The L_2 Rate of Convergence for Hazard Regression

Charles Kooperberg*
University of Washington

Charles J. Stone†
University of California at Berkeley

Young K. Truong‡
University of North Carolina at Chapel Hill

Technical Report No. 390
April 23, 1993
Department of Statistics
University of California
Berkeley, California 94720

Abstract

The logarithm of the conditional hazard function of a survival time given one or more covariates is approximated by a function having the form of a specified sum of functions of at most d of the variables. Subject to this form, the approximation is chosen to maximize the expected conditional log-likelihood. Maximum likelihood and sums of tensor products of polynomial splines are used to construct an estimate of this approximation based on a random sample. The components of this estimate possess a rate of convergence that depends only on d and a suitably defined smoothness parameter.

KEY WORDS: Conditional hazard function; Maximum likelihood; Tensor product splines.

*This research was supported in part by a grant from the Graduate School Fund of the University of Washington.

†This research was supported in part by National Science Foundation Grant DMS-9204247

‡This research was supported in part by a Research Council Grant from the University of North Carolina.

1 Introduction

Let T, C and $\mathbf{X}$ have a joint distribution, where T and C are nonnegative random variables and $\mathbf{X}$ is an M-dimensional random vector of covariates. In survival analysis, T and C are referred to as the survival time (or failure time) and censoring time, respectively. Set $Y = \min(T, C)$ and $\delta = \operatorname{ind}(T \leq C)$. Then the indicator random variable δ equals 1 if failure occurs on or before the censoring time (if $T \leq C$) and it equals 0 otherwise. The observable time Y is said to be uncensored or censored according as $\delta = 1$ or $\delta = 0$. For identifiability, T and C are assumed to be conditionally independent given $\mathbf{X}$.

Let $f(t|\mathbf{x})$ and $F(t|\mathbf{x})$ denote the conditional density function and conditional distribution function, respectively, of T given that $\mathbf{X} = \mathbf{x} \in \mathbb{R}^M$. The conditional survival, hazard and log-hazard functions are defined by

$$\bar{F}(t|\mathbf{x}) = 1 - F(t|\mathbf{x}), \quad h(t|\mathbf{x}) = f(t|\mathbf{x})/\bar{F}(t|\mathbf{x}) \quad \text{and} \quad \lambda(t|\mathbf{x}) = \log h(t|\mathbf{x}), \qquad t \geq 0.$$

Let $F_C(z|\mathbf{x})$ denote the conditional distribution function of C given that $\mathbf{X} = \mathbf{x}$, and set $\bar{F}_C(t|\mathbf{x}) = 1 - F_C(t|\mathbf{x})$.

A popular choice for the analysis of censored survival data with covariates is the proportional hazard model $\lambda(t|\mathbf{x}) = \lambda_0(t) + \mathbf{x}^T\boldsymbol{\beta}$, introduced by Cox (1972), where $\lambda_0(\cdot)$ is the baseline hazard function and $\boldsymbol{\beta} \in \mathbb{R}^M$ is a vector of parameters; see also Andersen et al (1993), Cox and Oakes (1984), Fleming and Harrington (1991), Kalbfleisch and Prentice (1980) and Miller (1981). In practice, it is more desirable to examine the covariate effects by using smooth, nonlinear functions. The *generalized additive model*

$$\lambda(t|\mathbf{x}) = \lambda_0(t) + \lambda_1(x_1) + \lambda_2(x_2) + \cdots + \lambda_M(x_M)$$

considered by Hastie and Tibshirani (1990), Sleeper and Harrington (1990) and Gray (1992) is a refinement of Cox's model. Here $\lambda_0(\cdot), \lambda_1(\cdot), \ldots, \lambda_M(\cdot)$ are smooth functions. In order to examine the interactions between covariates and time-varying coefficients, the generalized additive models can be further refined as follows. To motivate this approach, suppose $\mathbf{x} = (x_1, x_2)$ and write

$$\lambda(t|\mathbf{x}) = \lambda_0(t) + \lambda_1(x_1) + \lambda_2(x_2) + \lambda_{01}(t, x_1) + \lambda_{02}(t, x_2) + \lambda_{12}(x_1, x_2),$$

where $\lambda_0(\cdot), \lambda_1(\cdot), \lambda_2(\cdot), \ldots, \lambda_{12}(\cdot)$ are smooth functions. Here $\lambda_0(\cdot), \lambda_1(\cdot)$ and $\lambda_2(\cdot)$ are referred to as main effects, $\lambda_{12}(\cdot)$ is the interaction and $\lambda_{01}(\cdot)$ and $\lambda_{02}(\cdot)$ are components involving time-varying coefficients.

Given a random sample, consider an estimate

$$\hat{\lambda}(t|\mathbf{x}) = \hat{\lambda}_0(t) + \hat{\lambda}_1(x_1) + \hat{\lambda}_2(x_2) + \hat{\lambda}_{01}(t, x_1) + \hat{\lambda}_{02}(t, x_2) + \hat{\lambda}_{12}(x_1, x_2) \tag{1.1}$$

having the same form, where each component is empirically orthogonal to the corresponding lower order components. Such orthogonality will be defined precisely later in Section 2. We can think of $\hat{\lambda}(\cdot|\cdot)$ as an estimate of the log-hazard function $\lambda(\cdot|\cdot)$. Alternately, we can think of it as an estimate of the corresponding best theoretical approximation

$$\lambda^*(t|\mathbf{x}) = \lambda_0^*(t) + \lambda_1^*(x_1) + \lambda_2^*(x_2) + \lambda_{01}^*(t, x_1) + \lambda_{02}^*(t, x_2) + \lambda_{12}^*(x_1, x_2) \tag{1.2}$$

to the log-hazard function, where best means having the maximum expected likelihood subject to the indicated form and each component is theoretically orthogonal to the corresponding lower order components. According to Stone (1992), the right sides of (1.1) and (1.2) are referred to as the ANOVA decompositions of $\hat{\lambda}$ and λ^*, respectively. If the components of the ANOVA decomposition λ^* are estimated accurately by the corresponding ANOVA components of $\hat{\lambda}$, then examination of the components of the ANOVA decomposition of $\hat{\lambda}$ should shed light on the relationship of the survival time T and the covariates $\mathbf{X}$ through the function λ^* and, to a lesser extent, through the function λ.

In this paper, we consider the approximation λ^* to $\lambda = \log h$ having the form of a specified sum of functions of at most d of the variables $t, x_1, \ldots, x_M$ and, subject to this form, chosen to maximize the expected conditional log-likelihood. Given a random sample of size n from the distribution of $(Y, \delta, \mathbf{X})$, maximum likelihood and sums of tensor products of polynomial splines are used to construct estimates of λ^*. Its components are shown to possess the L_2 rate of convergence $n^{-p/(2p+d)}$, where p is a suitably defined smoothness parameter corresponding to λ^*. The problem of estimating the conditional density and survival functions are treated similarly by observing that

$$\bar{F}(t|\mathbf{x}) = \exp\left(-\int_0^t \exp(\lambda(u|\mathbf{x}))du\right), \qquad t \geq 0,$$

and

$$f(t|\mathbf{x}) = \exp(\lambda(t|\mathbf{x}))\exp\left(-\int_0^t \exp(\lambda(u|\mathbf{x}))du\right), \qquad t \geq 0.$$

The rest of the paper is organized as follows. Section 2.1 provides a preliminary discussion of the ANOVA decomposition. The formula for the expected log-likelihood function is derived in Section 2.2. The existence of the ANOVA decomposition of a specified form maximizing the expected log-likelihood is considered in Section 2.3. Maximum likelihood estimation based on a random sample is given in Section 2.4. Section 2.5 contains a discussion of works related to the current paper. Proofs are given in Section 3.

2 Statement of Results

2.1 Preliminaries

Given a subset s of $\{0, 1, \ldots, M\}$, let H_s denote the space of functions on $[0, \infty) \times \mathbb{R}^M$ that depend only on the variables

$$t, \text{ if } 0 \in s, \quad \text{and} \quad x_j, \text{ if } j \in s \cap \{1, \ldots, M\}.$$

Let $\mathcal{S}$ be a nonempty collection of subsets of $\{0, 1, \ldots, M\}$. It is assumed that $\mathcal{S}$ is *hierarchical*; that is, that if s is a member of $\mathcal{S}$ and r is a subset of s, then r is a member of $\mathcal{S}$. Let H denote the collection of functions of the form $a = \sum_{s \in \mathcal{S}} a_s$ with $a_s \in H_s$ for $s \in \mathcal{S}$.

2.2 Expected Log-Likelihood Function

The conditional likelihood based on $(Y, \delta, \mathbf{X})$ equals $[f(Y|\mathbf{X})]^{\delta}[\bar{F}(Y|\mathbf{X})]^{1-\delta}$. Observe that

$$\begin{aligned}
& E[\delta \log f(Y|\mathbf{X}) + (1-\delta)\log \bar{F}(Y|\mathbf{X})|\mathbf{X} = \mathbf{x}] \\
&= \iint_{t \le z} \log f(t|\mathbf{x})\, dF(t|\mathbf{x}) dF_C(z|\mathbf{x}) + \iint_{t > z} \log \bar{F}(z|\mathbf{x})\, dF(t|\mathbf{x}) dF_C(z|\mathbf{x}) \\
&= \int \bar{F}_C(t|\mathbf{x}) \log f(t|\mathbf{x})\, dF(t|\mathbf{x}) + \int \bar{F}(z|\mathbf{x}) \log \bar{F}(z|\mathbf{x})\, dF_C(z|\mathbf{x}) \\
&= \int \bar{F}_C(t|\mathbf{x})[\log \bar{F}(t|\mathbf{x}) + \log h(t|\mathbf{x})]\, dF(t|\mathbf{x}) + \int \bar{F}(t|\mathbf{x}) \log \bar{F}(t|\mathbf{x})\, dF_C(t|\mathbf{x}) \\
&= \int \bar{F}_C(t|\mathbf{x}) \log h(t|\mathbf{x})\, dF(t|\mathbf{x}) + \int \log \bar{F}(t|\mathbf{x})[\bar{F}_C(t|\mathbf{x}) dF(t|\mathbf{x}) + \bar{F}(t|\mathbf{x}) dF_C(t|\mathbf{x})] \\
&= \int \bar{F}_C(t|\mathbf{x}) \log h(t|\mathbf{x})\, dF(t|\mathbf{x}) - \int \left(\int_0^t h(u|\mathbf{x})\, du\right) [\bar{F}_C(t|\mathbf{x}) dF(t|\mathbf{x}) + \bar{F}(t|\mathbf{x}) dF_C(t|\mathbf{x})] \\
&= \int \bar{F}_C(t|\mathbf{x}) \log h(t|\mathbf{x})\, dF(t|\mathbf{x}) - \int \left(\int_u^\infty \bar{F}_C(t|\mathbf{x}) dF(t|\mathbf{x}) + \bar{F}(t|\mathbf{x}) dF_C(t|\mathbf{x})\right) h(u|\mathbf{x})\, du \\
&= \int \bar{F}_C(t|\mathbf{x}) \log h(t|\mathbf{x})\, dF(t|\mathbf{x}) - \int \bar{F}_C(u|\mathbf{x}) \bar{F}(u|\mathbf{x}) h(u|\mathbf{x})\, du.
\end{aligned}$$

Thus the expected conditional log-likelihood is given by

$$\begin{aligned}
& E[\delta \log f(Y|\mathbf{X}) + (1-\delta)\log \bar{F}(Y|\mathbf{X})] = \\
& \qquad \iint \bar{F}_C(t|\mathbf{x})[\log h(t|\mathbf{x}) f(t|\mathbf{x}) - \bar{F}(t|\mathbf{x}) h(t|\mathbf{x})]\, dt f_X(\mathbf{x})\, d\mathbf{x},
\end{aligned}$$

where $f_{\mathbf{X}}(\cdot)$ is the density function of the random vector $\mathbf{X}$. The expected conditional log-likelihood function $\Lambda(\cdot)$ is defined by

$$\Lambda(a) = \iint \bar{F}_C(t|\mathbf{x}) \left(a(t|\mathbf{x}) f(t|\mathbf{x}) - \bar{F}(t|\mathbf{x}) e^{a(t|\mathbf{x})}\right) dt f_X(\mathbf{x})\, d\mathbf{x}, \quad a \in H.$$

Note that $\Lambda(\cdot)$ is maximized at $\lambda = \log(f/\bar{F})$.

2.3 Existence

The first goal is to prove that $\Lambda(\cdot)$ has a maximum in H. Suppose the random vector $\mathbf{X}$ takes values in a compact interval $\mathcal{X} \subset \mathbb{R}^M$. Let $\mathcal{T}$ denote a compact interval of the form $[0, \tau]$ for some positive τ. Without loss of generality, we assume that $\mathcal{T} = [0, 1]$ and $\mathcal{X} = [0, 1]^M$.

Condition 1 *The joint density function of T and $\mathbf{X}$ is bounded away from zero and infinity on $\mathcal{T} \times \mathcal{X}$. Moreover, the survival function $\bar{F}(t|\mathbf{x})$ is bounded away from zero on $\mathcal{T} \times \mathcal{X}$.*

This condition implies that $\bar{F}(1|\mathbf{x}) = P(T > 1|\mathbf{X} = \mathbf{x}) > 0$ on $\mathcal{T} \times \mathcal{X}$ and that $|\lambda(t|\mathbf{x})|$ is bounded away from infinity on $\mathcal{T} \times \mathcal{X}$.

Condition 2 *$P(C \in \mathcal{T}|\mathbf{x}) = 1$ for $\mathbf{x} \in \mathcal{X}$ and $P(C = 1|\mathbf{x})$ is bounded away from zero on $\mathcal{X}$.*

This condition implies that $\bar{F}_C(t|\mathbf{x})$ is bounded away from zero on $[0, 1) \times \mathcal{X}$. According to this condition, censoring automatically occurs at time 1 if failure or censoring does not occur before this time.

Theorem 1 *Suppose Conditions 1 and 2 hold. Then there exists an essentially uniquely determined function $\lambda^* \in H$ such that $\Lambda(\lambda^*) = \max_{a \in H} \Lambda(a)$. If $\lambda \in H$, then $\lambda^* = \lambda$ almost everywhere.*

The uniqueness of ANOVA decomposition will be considered next. We first define inner products and orthogonality for functions on $\mathcal{T} \times \mathcal{X}$. Set

$$\langle a_1, a_2 \rangle = \int_{\mathcal{X}} \left(\int_{\mathcal{T}} a_1(y|\mathbf{x}) a_2(y|\mathbf{x}) f(y|\mathbf{x}) \bar{F}_C(y|\mathbf{x}) dy \right) f_X(\mathbf{x}) d\mathbf{x}$$

and $\|a\|^2 = \langle a, a \rangle$ for square integrable functions a_1, a_2, a on $\mathcal{T} \times \mathcal{X}$. For $s \in \mathcal{S}$, let H_s^2 denote the space of square integrable functions in H_s and set

$$H_s^0 = \{a \in H_s^2 : a \perp H_r^2 \quad \text{for} \quad r \subset s \quad \text{with } r \neq s\}.$$

(Here $a \perp H_r^2$ means that $\langle a, a_r \rangle = 0$ for $a_r \in H_r^2$.) Let H^2 denote the space of all functions of the form $\sum_{s \in \mathcal{S}} a_s$, $a_s \in H_s^2$ for $s \in \mathcal{S}$. Under Conditions 1 and 2, it can be shown that every function $a \in H^2$ can be written in an essentially unique manner as $\sum_{s \in \mathcal{S}} a_s$ where $a_s \in H_s^0$ for $s \in \mathcal{S}$; see Lemma 3.1 of Stone (1992). We refer to $\sum_{s \in \mathcal{S}} a_s$ as the ANOVA decompositions of a, and we refer to H_s^0, $s \in \mathcal{S}$, as the components of H^2.

Let $\#(s)$ denote the number of members of s, set $d = \max_{s \in \mathcal{S}} \#(s)$, and assume that $d \geq 1$. The component H_s^0 is referred to as the constant component if $\#(s) = 0$, as a main effect component if $\#(s) = 1$, and as an interactive component if $\#(s) \geq 2$.

Suppose λ^* in Theorem 1 is in H^2. Then it can be written in an essentially unique manner in the form $\lambda^* = \sum_{s \in \mathcal{S}} \lambda_s^*$ where $\lambda_s^* \in H_s^0$ for $s \in \mathcal{S}$. The rate of convergence in estimating λ^* depends on a smoothness condition on λ_s^*, $s \in \mathcal{S}$, which will now be described.

Let $0 < \beta \leq 1$. A function a on $\mathcal{T} \times \mathcal{X}$ is said to satisfy a Hölder condition with exponent β if there is a positive number γ such that $|a(\mathbf{z}) - a(\mathbf{z}_0)| \leq \gamma |\mathbf{z} - \mathbf{z}_0|^\beta$ for $\mathbf{z}, \mathbf{z}_0 \in \mathcal{T} \times \mathcal{X}$;

here $|\mathbf{z}|^2 = \sum_0^M z_j^2$ is the square of the Euclidean norm of $\mathbf{z} = (z_0, z_1, \ldots, z_M)$. Given an $(M+1)$-tuple $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_M)$ of nonnegative integers, set $[\alpha] = \alpha_0 + \alpha_1 + \cdots + \alpha_M$ and let D^α denote the differentiable operator defined by

$$D^\alpha = \frac{\partial^{[\alpha]}}{\partial z_0^{\alpha_0} \cdots \partial z_M^{\alpha_M}}.$$

Let m be a nonnegative integer and set $p = m + \beta$. A function a on $\mathcal{T} \times \mathcal{X}$ is said to be p-smooth if a is m times continuously differentiable on $\mathcal{T} \times \mathcal{X}$ and $D^\alpha a$ satisfies a Hölder condition with exponent β for all α with $[\alpha] = m$. In the following condition, it is assumed that $p > d/2$.

Condition 3 *There are p-smooth functions $\lambda_s^* \in H_s^0$, $s \in \mathcal{S}$, such that $\lambda^* = \sum_{s \in \mathcal{S}} \lambda_s^* \in H$ and $\Lambda(\lambda^*) = \max_{a \in H} \Lambda(a)$.*

2.4 Maximum Likelihood Estimation

Let $K = K_n$ be a positive integer, and let I_k, $1 \leq k \leq K$, denote the subintervals of $[0,1]$ defined by $I_k = [(k-1)/K, k/K)$ for $1 \leq k < K$ and $I_K = [1 - 1/K, 1]$. Let m and q be fixed integers such that $m \geq 0$ and $m > q \geq -1$. Let $S = S_n$ denote the space of functions g on $[0,1]$ such that

(i) the restriction of g to I_k is a polynomial of degree m (or less) for $1 \leq k \leq K$;

and, if $q \geq 0$, then

(ii) g is q-times continuously differentiable on $[0,1]$.

A function satisfying (i) is called a piecewise polynomial, and it is called a spline if it satisfies both (i) and (ii). Let B_j, $1 \leq j \leq J$, denote the usual basis of S consisting of B-splines [see de Boor (1978)]. Then $J = (m+1)K - (q+1)(K-1)$, so $K + m \leq J \leq (m+1)K$. Also, $B_j \geq 0$ on $[0,1]$, $B_j = 0$ on the complement of an interval of length $(m+1)/K$ for $1 \leq j \leq J$, and $\sum_j B_j = 1$ on $[0,1]$. Moreover, for $1 \leq j \leq J$, there are at most $2m+1$ values of $j' \in \{1, \ldots, J\}$ such that $B_j B_{j'}$ is not identically zero on $[0,1]$.

Let $G_\emptyset$ denote the space of constant functions on $\mathcal{T} \times \mathcal{X}$. Given a subset s of $\{0, 1, \ldots, M\}$, let G_s denote the space spanned by the functions g on $\mathcal{T} \times \mathcal{X}$ of the form

$$g(\mathbf{z}) = \prod_{j \in s} g_j(z_j), \qquad \text{where } \mathbf{z} = (z_0, z_1, \ldots, z_M) \text{ and } g_j \in S \text{ for } j \in s.$$

Then G_s has dimension $J^{\#(s)}$. Moreover, $G_r \subset G_s$ for $r \subset s$.

Consider a random sample $(T_1, C_1, \mathbf{X}_1), \ldots, (T_n, C_n, \mathbf{X}_n)$ from the distribution of $(T, C, \mathbf{X})$, and set $Y_i = \min(T_i, C_i)$ and $\delta_i = \operatorname{ind}(T_i \leq C_i)$ for $1 \leq i \leq n$. Let $\langle \cdot, \cdot \rangle_n$ denote the sample inner product defined by

$$\langle g_1, g_2 \rangle_n = \frac{1}{n} \sum_{i:\delta_i=1} g_1(Y_i|\mathbf{X}_i) g_2(Y_i|\mathbf{X}_i).$$

Let G_s^0 denote the space of functions in G_s, $s \in \mathcal{S}$ that are orthogonal (relative to $\langle \cdot, \cdot \rangle_n$) to each function in G_r for every proper subset r of s, and set

$$G = \left\{ \sum_{s \in \mathcal{S}} g_s : g_s \in G_s^0 \quad \text{for } s \in \mathcal{S} \right\}.$$

The space G is said to be nonidentifiable if there is a nonzero function g in the space such that $g(Y_i|\mathbf{X}_i) = 0$ for every $i \in \{1, \ldots, n\}$ such that $\delta_i = 1$; otherwise this space is said to be identifiable. Suppose G is identifiable, and let g be a member of this space. Then g can be written uniquely in the form $\sum_{s \in \mathcal{S}} g_s$, where $g_s \in G_s^0$ for $s \in \mathcal{S}$; see Lemma 3.2 of Stone (1992).

Condition 4 $J^{2d} = o(n^{1-\epsilon})$ *for some* $\epsilon > 0$.

It follows from Conditions 1, 2 and 4 [see Lemma 3.8 of Stone (1992)] that

$$P\Big(G \text{ is nonidentifiable}\Big) = o(1). \tag{2.1}$$

The likelihood corresponding to $(Y_1, \delta_1, \mathbf{X}_1), \ldots, (Y_n, \delta_n, \mathbf{X}_n)$ is given by

$$\prod_i [f(Y_i|\mathbf{X}_i)]^{\delta_i} [\bar{F}(Y_i|\mathbf{X}_i)]^{1-\delta_i},$$

and the log-likelihood is given by

$$\sum_i [\delta_i \log f(Y_i|\mathbf{X}_i) + (1-\delta_i) \log \bar{F}(Y_i|\mathbf{X}_i)].$$

For $s \in \mathcal{S}$, let $\mathcal{J}_s$ denote the collection of ordered $\#(s)$-tuples j_l, $l \in s$, with $j_l \in \{1, \ldots, J\}$ for $l \in s$. Then $\#(\mathcal{J}_s) = J^{\#(s)}$. For $\boldsymbol{j} \in \mathcal{J}_s$, let $B_{s\boldsymbol{j}}$ denote the function on $\mathcal{T} \times \mathcal{X}$ given by

$$B_{s\boldsymbol{j}}(y|\mathbf{x}) = \prod_{l \in s} B_{j_l}(x_l), \quad \mathbf{x} = (x_1, \ldots, x_M) \text{ and } x_0 = y.$$

Then the function $B_{s\boldsymbol{j}}$, $\boldsymbol{j} \in \mathcal{J}_s$, which are nonnegative and have sum one, form a basis of G_s.

Set $I = \sum_s \#(\mathcal{J}_s)$. Given an I-dimensional (column) vector $\boldsymbol{\theta}$ having entries $\theta_{s\boldsymbol{j}}$, $s \in \mathcal{S}$ and $\boldsymbol{j} \in \mathcal{J}_s$, set

$$g_s(\cdot|\cdot; \boldsymbol{\theta}) = \sum_{\boldsymbol{j} \in \mathcal{J}_s} \theta_{s\boldsymbol{j}} B_{s\boldsymbol{j}}(\cdot|\cdot), \quad s \in \mathcal{S},$$

and

$$g(\cdot|\cdot\,;\boldsymbol{\theta}) = \sum_{s\in\mathcal{S}} g_s(\cdot|\cdot\,;\boldsymbol{\theta}).$$

Then the log-likelihood function can be written as

$$\ell(\boldsymbol{\theta}) = \sum_i \delta_i g(Y_i|\mathbf{X}_i;\boldsymbol{\theta}) - \sum_i \int_0^{Y_i} \exp(g(u|\mathbf{X}_i;\boldsymbol{\theta}))\,du.$$

Define $\hat{\boldsymbol{\theta}}$ so that $\ell(\hat{\boldsymbol{\theta}}) = \max \ell(\boldsymbol{\theta})$, and consider $\hat{\lambda} = g(\cdot|\cdot\,;\hat{\boldsymbol{\theta}})$ as the maximum likelihood estimate of λ^* in G.

Theorem 2 *Suppose Conditions 1, 2 and 4 hold. Then, except on an event whose probability tends to zero with n, G is identifiable, the maximum likelihood estimate $\hat{\lambda}$ in G exists, and it can be written uniquely in the form $\sum_{s\in\mathcal{S}} \hat{\lambda}_s$ with $\hat{\lambda}_s \in G_s^0$ for $s \in \mathcal{S}$.*

Theorem 3 *Suppose Conditions 1–4 hold. Then*

$$\|\hat{\lambda}_s - \lambda_s^*\| = O_P\left(J^{-p} + \sqrt{J^d/n}\right), \quad s \in \mathcal{S},$$

so

$$\|\hat{\lambda} - \lambda^*\| = O_P\left(J^{-p} + \sqrt{J^d/n}\right).$$

Given positive numbers a_n and b_n for $n \geq 1$, let $a_n \sim b_n$ mean that a_n/b_n is bounded away from zero and infinity.

Corollary 1 *Suppose Conditions 1–3 hold and that $J \sim n^{1/(2p+d)}$. Then*

$$\|\hat{\lambda}_s - \lambda_s^*\| = O_P\left(n^{-p/(2p+d)}\right), \quad s \in \mathcal{S},$$

so

$$\|\hat{\lambda} - \lambda^*\| = O_P\left(n^{-p/(2p+d)}\right).$$

The L_2 rate of convergence in Corollary 1 depends on d, not on the dimension M of the random vector $\mathbf{X}$. This provides a nontrivial justification of the *heuristic dimensionality reduction principle* discussed by Stone (1985). When $d = M$, the rate is optimal according to Stone (1982) and Hasminskii and Ibragimov (1990).

2.5 Related Work

An excellent discussion of the literature on the estimation of hazard and survival functions from the counting process viewpoint is contained in Anderson et al (1993); see also Fleming and Harrington (1991). Here we discuss various approaches related to the theory developed in the present paper.

In the absence of censored observations, the theory for additive and generalized additive regression ($d = 1$) was considered by Stone (1985, 1986, 1989, 1990, 1991). The general ANOVA decomposition and its corresponding rates of convergence in nonparametric estimation were given in Stone (1992).

Except for the estimation of the baseline hazard function, numerical procedures for estimating main effects based on the generalized additive models were discussed by Hastie and Tibshirani (1990), Sleeper and Harrington (1990) and Gray (1992). The problem of density estimation (without covariates) under right, left and interval censorings was considered by Kooperberg and Stone (1992). From a methodological point of view, hazard regression based on adaptive model selection as in MARS (Friedman, 1991) was considered by Kooperberg, Stone and Truong (1993). The current paper lends theoretical support to such an adaptive methodology. Extensions of the present theory to handle time-dependent covariates as considered by Zucker and Karr (1990) and Hastie and Tibshirani (1993) should be practically useful.

3 Proofs

3.1 Proof of Theorem 1

Write

$$\Lambda(a) = \iint (ah - e^a)\bar{F}_C \bar{F} f_X, \qquad a \in H.$$

According to Condition 1, $h(\cdot|\cdot)$ is bounded away from zero and infinity on $\mathcal{T} \times \mathcal{X}$. Thus, by elementary algebra, there are positive constants A and ε such that

$$ah - e^a \leq A - \varepsilon |a|, \qquad a \in H.$$

It follows from Conditions 1 and 2 (with ε appropriately redefined) that

$$\Lambda(a) \leq A - \varepsilon \iint |a| f f_X, \qquad a \in H.$$

Thus, if $\iint |a| f f_X = \infty$, then $\Lambda(a) = -\infty$. Moreover, the function $\Lambda(\cdot)$ is bounded above by A. Hence, the numbers $\Lambda(a)$, $a \in H$, have a finite least upper bound L. Choose $a_k \in H$ such that $\Lambda(a_k) > -\infty$ and $\Lambda(a_k) \to L$ as $k \to \infty$. Observe that the numbers $\iint |a_k| f f_X$, $k \geq 1$, are bounded.

Let λ_1 and λ_2 be functions in H such that $\Lambda(\lambda_1) > -\infty$ and $\Lambda(\lambda_2) > -\infty$. For $u \in [0,1]$, set $\lambda^{(u)} = (1-u)\lambda_1 + u\lambda_2$ and $\Psi(u) = \Lambda(\lambda^{(u)})$. Then by the concavity of $u \mapsto \lambda^{(u)}h - e^{\lambda^{(u)}}$, $\Psi(\cdot)$ is a concave function. (Note that if λ_1 and λ_2 are bounded, then

$$\Psi''(u) = -\int\int (\lambda_2 - \lambda_1)^2 e^{\lambda^{(u)}} \bar{F}_C \bar{F} f_X.)$$

It follows from the argument of Theorem 4.1 in Stone (1992) that there is an integrable function λ^* such that $a_k \to \lambda^*$ in measure. By Lemma 4.1 of Stone (1992), we can assume that $\lambda^* \in H$. It follows from Fatou's Lemma that $\Lambda(a_k) \to \Lambda(\lambda^*) = L = \max_{a \in H} \Lambda(a)$. Furthermore, if $a \in H$ and $\Lambda(a) = \Lambda(\lambda^*)$, then $a = \lambda^*$ almost everywhere. Hence the first statement of the theorem is valid. The second statement follows from the fact that the function $a \mapsto ah - e^a$ has a unique maximum at $\lambda = \log h$.

3.2 Proof of Theorem 2

Throughout this subsection, it is assumed that Conditions 1–4 hold. Also, set $\|g\|_\infty = \sup_{t \in \mathcal{T}, \mathbf{x} \in \mathcal{X}} |g(t|\mathbf{x})|$.

Lemma 1 *Let U be a positive constant. Then there are positive constants M_1 and M_2 such that*

$$-M_1\|a - \lambda^*\|^2 \le \Lambda(a) - \Lambda(\lambda^*) \le -M_2\|a - \lambda^*\|^2$$

for all $a \in H$ such that $\|a\|_\infty \le U$.

Proof. Given $a \in H$ with $\|a\|_\infty \le U$ and given $u \in [0,1]$, set

$$\lambda^{(u)} = (1-u)\lambda^* + ua.$$

Then

$$\frac{d}{du}\Lambda(\lambda^{(u)})|_{u=0} = 0$$

and by integration by parts

$$\begin{aligned}\Lambda(a) - \Lambda(\lambda^*) &= \int_0^1 (1-u)\frac{d^2}{du^2}\Lambda(\lambda^{(u)})du \\ &= -\int_0^1 (1-u)\int\int (a - \lambda^*)^2 e^{\lambda^{(u)}} \bar{F}_C \bar{F} f_X du.\end{aligned}$$

The desired result now follows from Conditions 1 and 2. □

The next result is Lemma 4.3 of Stone (1992).

Lemma 2 *There is a positive constant M_3 such that $\|g\|_\infty \le M_3 J^{d/2}\|g\|$ for $g \in G$.*

Under Conditions 1 and 2, by an argument similar to that used to prove Theorem 1, there is a unique function $\lambda_n^* \in G$ such that $\Lambda(\lambda_n^*) = \max_{g\in G} \Lambda(g)$.

Lemma 3 $\|\lambda_n^* - \lambda^*\|^2 = O(J^{-2p})$ *and* $\|\lambda_n^* - \lambda^*\|_\infty = O(J^{d/2-p})$.

Proof. By Condition 3 and Theorem 12.8 of Schumaker (1981), there are functions $g_n \in G$ for $n \geq 1$ and a positive constant M_4 such that

$$\|g_n - \lambda^*\|_\infty \leq M_4 J^{-p}.$$

Consequently,

$$\|g_n - \lambda^*\|^2 \leq M_4^2 J^{-2p}.$$

By Lemma 1, there is a positive constant M_5 such that

$$\Lambda(g_n) - \Lambda(\lambda^*) \geq -M_5 J^{-2p}. \tag{3.1}$$

Let b be a positive constant. Choose $g \in G$ with

$$\|g - \lambda^*\|^2 = bJ^{-2p}.$$

Then

$$\|g_n - g\|^2 \leq 2\Big(\|g_n - \lambda^*\|^2 + \|\lambda^* - g\|^2\Big) \leq 2(b + M_4^2)J^{-2p}.$$

By Lemma 2, for J sufficiently large,

$$\|g\|_\infty \leq \|g - g_n\|_\infty + \|g_n - \lambda^*\|_\infty + \|\lambda^*\|_\infty \leq 1 + \|\lambda^*\|_\infty,$$

since $p > d/2$. Thus by Lemma 1, there is a positive constant M_6 such that, for J sufficiently large,

$$\Lambda(g) - \Lambda(\lambda^*) \leq -M_6 bJ^{-2p} \qquad \text{for all } g \in G \text{ with } \|g - \lambda^*\| = bJ^{-2p}. \tag{3.2}$$

Let b be chosen so that $b > M_4^2$ and $M_6 b > M_5$. By (3.1) and (3.2), for J sufficiently large,

$$\Lambda(g) < \Lambda(g_n) \qquad \text{for all } g \in G \text{ with } \|g - \lambda^*\| = bJ^{-2p}.$$

Therefore, $\Lambda(\cdot)$ has a local maximum on $\|g - \lambda^*\| < bJ^{-2p}$, and by its concavity,

$$\|\lambda_n^* - \lambda^*\| < bJ^{-2p}$$

for J sufficiently large. It follows from Lemma 2 and $\|\lambda_n^* - g_n\|^2 = O(J^{-2p})$ that

$$\|\lambda_n^* - g_n\|_\infty = O(J^{d/2-p}).$$

Consequently,

$$\|\lambda_n^* - \lambda^*\|_\infty = O(J^{d/2-p}). \quad \Box$$

Suppose Condition 4 holds. Let $\tau_n, n \geq 1$, be positive numbers such that $J^d \tau_n^2 = O(1)$ and $J^d \log n = o(n\tau_n^2)$. Let $\boldsymbol{\theta}^*$ be given by

$$\lambda_n^*(\cdot|\cdot) = g(\cdot|\cdot\,;\boldsymbol{\theta}^*) = \sum_{s\in\mathcal{S}} g_s(\cdot|\cdot\,;\boldsymbol{\theta}^*).$$

Lemma 4 *Given $b > 0$ and $\varepsilon > 0$, there is a $c > 0$ such that, for n sufficiently large,*

$$P\left(\left|\frac{\ell(g)-\ell(\lambda_n^*)}{n} - [\Lambda(g)-\Lambda(\lambda_n^*)]\right| \geq \varepsilon\tau_n^2\right) \leq 2\exp(-cn\tau_n^2)$$

for all $g \in G$ with $\|g-\lambda_n^*\| \leq b\tau_n$.

Proof. Write

$$\frac{\ell(g)-\ell(\lambda_n^*)}{n} - [\Lambda(g)-\Lambda(\lambda_n^*)] = n^{-1}\sum_i [W_i - E(W_i)],$$

where

$$\begin{aligned} W_i &= \delta_i g(Y_i|\mathbf{X}_i;\boldsymbol{\theta}) - \int_0^{Y_i} \exp(g(u|\mathbf{X}_i;\boldsymbol{\theta}))\,du \\ &\quad -\delta_i g(Y_i|\mathbf{X}_i;\boldsymbol{\theta}^*) + \int_0^{Y_i} \exp(g(u|\mathbf{X}_i;\boldsymbol{\theta}^*))\,du. \end{aligned}$$

By Lemma 2,

$$\|g(\cdot|\cdot\,;\boldsymbol{\theta}) - g(\cdot|\cdot\,;\boldsymbol{\theta}^*)\|_\infty = O(J^{d/2}\tau_n),$$

for $g(\cdot|\cdot\,;\boldsymbol{\theta})$ satisfying $\|g-\lambda_n^*\| \leq b\tau_n$. Thus there is a positive constant M_7 such that

$$\int_0^{Y_i} |\exp(g(u|\mathbf{X}_i;\boldsymbol{\theta})) - \exp(g(u|\mathbf{X}_i;\boldsymbol{\theta}^*))|\,du \leq M_7 \int_0^{Y_i} |g-\lambda_n^*|,$$

for $g(\cdot|\cdot\,;\boldsymbol{\theta})$ satisfying $\|g-\lambda_n^*\| \leq b\tau_n$. Hence, according to Condition 1, $|W_i| = O(J^{d/2}\tau_n)$. Moreover, by Condition 1,

$$E\left(\int_0^{Y_i} |g-\lambda_n^*|\right)^2 \leq E\left(\int_0^1 |g-\lambda_n^*|^2\right) \leq (b\tau_n)^2$$

and

$$E\{[\delta_i g(Y_i|\mathbf{X}_i;\boldsymbol{\theta}) - \delta_i g(Y_i|\mathbf{X}_i;\boldsymbol{\theta}^*)]^2\} \leq E\{[g(T_i|\mathbf{X}_i;\boldsymbol{\theta}) - g(T_i|\mathbf{X}_i;\boldsymbol{\theta}^*)]^2\} = O(\tau_n^2)$$

for $g(\cdot|\cdot\,;\boldsymbol{\theta})$ satisfying $\|g-\lambda_n^*\| \leq b\tau_n$. Hence $\mathrm{var}(W_i) = O(\tau_n^2)$. The desired result now follows from Bernstein's inequality [see (2.13) of Hoeffding (1963)]. □

Define the diameter of a set $\mathcal{E}$ of functions on $\mathcal{T} \times \mathcal{X}$ by

$$\sup\{\|g_1 - g_2\|_\infty : g_1, g_2 \in \mathcal{E}\}.$$

The next result is essentially the same as that of lemma 4.8 of Stone (1992).

Lemma 5 *Given $b > 0$ and $c > 0$, there is a $M_8 > 0$ such that, for n sufficiently large,*

$$\{g : g \in G \quad \text{and} \quad \|g - \lambda_n^*\| \le b\tau_n\}$$

can be covered by $O(\exp(M_8 J^d \log n))$ subsets each having diameter at most $c\tau_n^2$.

Lemma 6 *Let $b > 0$. Then, except on an event whose probability tends to zero with n, $\ell(g) < \ell(\lambda_n^*)$ for all $g \in G$ such that $\|g - \lambda_n^*\| = b\tau_n$.*

Proof. Choose $g \in G$ such that $\|g - \lambda_n^*\| = b\tau_n$. By Lemma 2,

$$\|g - \lambda_n^*\|_\infty = O(J^{d/2}\|g - \lambda_n^*\|) = O(J^{d/2}\tau_n) = O(1).$$

Thus by Lemma 3, $\|g\|_\infty = O(1)$. Hence,

$$\left|\frac{\ell(g_2) - \ell(g_1)}{n}\right| = O(\|g_2 - g_1\|_\infty) \quad \text{and} \quad |\Lambda(g_1) - \Lambda(g_2)| = O(\|g_2 - g_1\|_\infty)$$

for $g_1, g_2 \in G$ such that $\|g_i - \lambda_n^*\| \le b\tau_n$ for $i = 1, 2$. The desired result follows from Lemma 1, with λ^* replaced by λ_n^* and H by G, and Lemmas 4 and 5. □

Lemma 7 *The maximum likelihood estimate $\hat{\lambda} \in G$ of $\lambda = \log h$ exists, and is unique except on an event whose probability tends to zero with n. Moreover, $\|\hat{\lambda} - \lambda_n^*\|_\infty = o_P(1)$.*

Proof. The set $G_0 = \{g \in G : \|g - \lambda_n^*\| \le b\tau_n\}$ is a compact set, with boundary $\{g \in G : \|g - \lambda_n^*\| = b\tau_n\}$. By Lemma 6, the function $\ell(\cdot)$ has local maximum in the interior of G_0. It follows from the strictly concavity of the function $\ell(\cdot)$ that $\|\hat{\lambda} - \lambda_n^*\| = o_P(\tau_n)$ and hence from Lemma 2 that

$$\|\hat{\lambda} - \lambda_n^*\|_\infty = o_P(J^{d/2}\tau_n) = o_P(1). \quad \square$$

The proof of Theorem 2 now follows from Lemmas 3.2 and 3.8 of Stone (1992) and Lemma 7.

3.3 Proof of Theorem 3

The proof of the next result is similar to that of Lemma 5.3 of Stone (1992).

Lemma 8 *Suppose Conditions 1–4 hold. Then*

$$\|\lambda_{ns}^* - \lambda_s^*\|^2 = O_P\left(J^{-2p} + J^d/n\right), \quad s \in \mathcal{S}.$$

Recall that the log-likelihood function is given by

$$\ell(\boldsymbol{\theta}) = \sum_i \delta_i g(Y_i|\mathbf{X}_i;\boldsymbol{\theta}) - \sum_i \int_0^{Y_i} \exp(g(u|\mathbf{X}_i;\boldsymbol{\theta}))\,du$$

and that $I = \sum_s \#(\mathcal{J}_s)$. Let

$$\mathbf{S}(\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}}\ell(\boldsymbol{\theta})$$

denote the score at $\boldsymbol{\theta}$; that is, the I-dimensional vector with entries

$$\frac{\partial\ell(\boldsymbol{\theta})}{\partial\theta_{s\boldsymbol{j}}} = \sum_i \delta_i B_{s\boldsymbol{j}}(Y_i|\mathbf{X}_i) - \sum_i \int_0^{Y_i} B_{s\boldsymbol{j}}(u|\mathbf{X}_i)\exp(g(u|\mathbf{X}_i;\boldsymbol{\theta}))\,du.$$

Let

$$\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}$$

denote the Hessian of $\ell(\boldsymbol{\theta})$; that is, the $I \times I$ matrix having entries

$$\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\theta_{s_1\boldsymbol{j}_1}\partial\theta_{s_2\boldsymbol{j}_2}} = -\sum_i \int_0^{Y_i} B_{s_1\boldsymbol{j}_1}(u|\mathbf{X}_i)B_{s_2\boldsymbol{j}_2}(u|\mathbf{X}_i)\exp(g(u|\mathbf{X}_i;\boldsymbol{\theta}))\,du. \tag{3.3}$$

The maximum likelihood equation $\mathbf{S}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ can be written as

$$\int_0^1 \frac{d}{du}\mathbf{S}(\boldsymbol{\theta}^* + u(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*))du = -\mathbf{S}(\boldsymbol{\theta}^*).$$

This can further be written as $\mathbf{D}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = -\mathbf{S}(\theta^*)$, where $\mathbf{D}$ is the $I \times I$ matrix given by

$$\mathbf{D} = \int_0^1 \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\ell(\boldsymbol{\theta}^* + u(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*))du.$$

It follows from the maximum likelihood equation that

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T\mathbf{D}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = -(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T\mathbf{S}(\boldsymbol{\theta}^*). \tag{3.4}$$

We claim that

$$|\mathbf{S}(\boldsymbol{\theta}^*)|^2 = O_P(n) \tag{3.5}$$

and that there is positive constant M_9 such that

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T\mathbf{D}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \le -M_9 n J^{-d}|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*|^2 \tag{3.6}$$

except on an event whose probability tends to zero with n. Since

$$|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T\mathbf{S}(\boldsymbol{\theta}^*)| \le |\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*||\mathbf{S}(\boldsymbol{\theta}^*)|,$$

it follows from (3.4)–(3.6) that $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*|^2 = O_P(J^{2d}/n)$ and hence that

$$\|\hat{\lambda}_s - \lambda^*_{ns}\|^2 = O_P(J^d/n), \quad s \in \mathcal{S}, \tag{3.7}$$

and

$$\|\hat{\lambda} - \lambda^*_n\|^2 = O_P(J^d/n). \tag{3.8}$$

Theorem 3 follows from (3.7), (3.8) and Lemmas 3 and 8.

Proof of (3.5). By the definition of $\boldsymbol{\theta}^*$, we have

$$E\left(\frac{\partial \ell(\boldsymbol{\theta}^*)}{\partial \theta_{s\mathbf{j}}}\right) = 0.$$

Hence

$$E\left(\frac{\partial \ell(\boldsymbol{\theta}^*)}{\partial \theta_{s\mathbf{j}}}\right)^2 = \operatorname{var}\left(\sum_i \delta_i B_{s\mathbf{j}}(Y_i|\mathbf{X}_i) - \sum_i \int_0^{Y_i} B_{s\mathbf{j}}(u|\mathbf{X}_i) \exp(g(u|\mathbf{X}_i;\boldsymbol{\theta}^*))\, du\right).$$

It follows from

$$\operatorname{var}\left(\delta_i B_{s\mathbf{j}}(Y_i|\mathbf{X}_i) - \int_0^{Y_i} B_{s\mathbf{j}}(u|\mathbf{X}_i) \exp(g(u|\mathbf{X}_i;\boldsymbol{\theta}^*))\, du\right) = O(1)$$

that

$$E|\mathbf{S}(\boldsymbol{\theta}^*)|^2 = O(n).$$

So (3.5) holds.

Proof of (3.6). It follows from (3.3) that

$$\boldsymbol{\beta}^T \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \boldsymbol{\beta} = -\sum_i \int_0^{Y_i} g^2(u|\mathbf{X}_i;\boldsymbol{\beta}) \exp(g(u|\mathbf{X}_i;\boldsymbol{\theta}))\, du. \tag{3.9}$$

By Lemmas 3 and 7, there is a positive constant U such that

$$\lim_{n\to\infty} P\left(\|\lambda^*_n\|_\infty \le U \quad \text{and} \quad \|\hat{\lambda}\|_\infty \le U\right) = 1. \tag{3.10}$$

It follows from (3.9) and (3.10) that there is a positive constant ε such that, except on an event whose probability tends to zero with n,

$$\boldsymbol{\beta}^T \mathbf{D} \boldsymbol{\beta} \le -\varepsilon \sum_i \int_0^{Y_i} g^2(u|\mathbf{X}_i;\boldsymbol{\beta})\, du. \tag{3.11}$$

Since

$$\int_0^{Y_i} g_s^2(u|\mathbf{X}_i;\boldsymbol{\beta})\, du \le \int_0^{Y_i} \sum_j \beta_{sj}^2 B_{sj}(u|\mathbf{X}_i)\, du = O\left(J^{-1} \sum_j \beta_{sj}^2\right),$$

it follows from Condition 4 and Chebyshev's inequality that

$$\sum_i \int_0^{Y_i} g^2(u|\mathbf{X}_i;\boldsymbol{\beta})\,du - E\sum_i \int_0^{Y_i} g^2(u|\mathbf{X}_i;\boldsymbol{\beta})\,du = o_P\left(nJ^{-d}|\boldsymbol{\beta}|^2\right). \tag{3.12}$$

By Conditions 1 and 2, there is a $M_{10} > 0$ such that

$$\begin{aligned} & E\left(\int_0^{Y_i} g^2(u|\mathbf{X}_i;\boldsymbol{\beta})\,du\right) \\ &= \int\int\int_0^y g^2(u|\mathbf{x};\boldsymbol{\beta})du[\bar{F}_C(y|\mathbf{x})dF(y|\mathbf{x}) + \bar{F}(y|\mathbf{x})dF_C(y|\mathbf{x})]f_{\mathbf{X}}(\mathbf{x})\,d\mathbf{x} \\ &= \int\int g^2(u|\mathbf{x};\boldsymbol{\beta})\bar{F}_C(u|\mathbf{x})\bar{F}(u|\mathbf{x})f_{\mathbf{X}}(\mathbf{x})\,dud\mathbf{x} \\ &\geq M_{10}\int_{\mathcal{X}}\int_{\mathcal{T}} g^2(u|\mathbf{x};\boldsymbol{\beta})\,dud\mathbf{x}. \end{aligned} \tag{3.13}$$

According to Conditions 1 and 4 and Lemma 3.6 in Stone (1992), there is a $M_{11} > 0$ such that, except on an event whose probability tends to zero with n,

$$\int_{\mathcal{X}}\int_{\mathcal{T}} g^2(u|\mathbf{x};\boldsymbol{\beta})\,dud\mathbf{x} \geq M_{11}\sum_{s\in\mathcal{S}}\int_{\mathcal{X}}\int_{\mathcal{T}} g_s^2(u|\mathbf{x};\boldsymbol{\beta})\,dud\mathbf{x}, \qquad \boldsymbol{\beta}\in\mathbb{R}^I. \tag{3.14}$$

It follows from the basic properties of B-splines that, for some $\varepsilon > 0$,

$$\int_{\mathcal{X}}\int_{\mathcal{T}} g_s^2(u|\mathbf{x};\boldsymbol{\beta})\,dud\mathbf{x} \geq \varepsilon J^{-\#(s)}\sum_j \beta_{sj}^2, \quad s\in\mathcal{S} \text{ and } \boldsymbol{\beta}\in\mathbb{R}^I,$$

and hence

$$\sum_{s\in\mathcal{S}}\int_{\mathcal{X}}\int_{\mathcal{T}} g_s^2(u|\mathbf{x};\boldsymbol{\beta})\,dud\mathbf{x} \geq \varepsilon J^{-d}|\boldsymbol{\beta}|^2, \qquad \boldsymbol{\beta}\in\mathbb{R}^I. \tag{3.15}$$

Equation (3.6) follows from (3.11)–(3.15) applied to $\boldsymbol{\beta} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$. This completes the proof of Theorem 3.

References

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag, New York.

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

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* Chapman and Hall, London.

de Boor, C. (1978). *A Practical Guide to Splines.* Springer-Verlag, New York.

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

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19**, 1–141.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *J. Amer. Statist. Assoc.* **87**, 942–951.

Hasminskii, R. and Ibragimov, I. (1990). Kolmogorov's contributions to mathematical statistics. *Ann. Statist.* **18**, 1011–1016.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models (with discussion). *J. Roy. Statist. Soc. Ser. B* **55**. To appear.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58**, 13–30.

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

Kooperberg, C. and Stone, C. J. (1992). Logspline density estimation for censored data. *J. Comput. Graphical Statist.* **1**, 301–328.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1993). Hazard regression. Technical Report No. 389, Department of Statistics, University of California, Berkeley.

Miller, R. G. (1981). *Survival Analysis.* Wiley, New York.

Schumaker, L. L. (1981). *Spline Functions: Basic Theory.* Wiley, New York.

Sleeper, L. A. and Harrington, D. P. (1990). Regression splines in the Cox model with application to covariate effects in liver disease. *J. Amer. Statist. Assoc.* **85**, 941–949.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1040–1053.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13**, 689–705.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14**, 590–606.

Stone, C. J. (1989). Uniform error bounds involving logspline models. *In Probability, Statistics and Mathematics: Papers in Honor of Samuel Karlin.* (T. W. Anderson, K. B. Athrya, and D. L. Iglehart, eds.) 335–355. Academic Press, Boston.

Stone, C. J. (1990). Large-sample inference for log-spline models. *Ann. Statist.* **18**, 717–741.

Stone, C. J. (1991). Asymptotics for doubly logspline response models. *Ann. Statist.* **19**, 1832–1854.

Stone, C. J. (1992). The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.* To appear.

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
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

SCHOOL OF PUBLIC HEALTH
DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27599-7400