 Vol. 22: 143–157, 1995

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

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

1. Introduction

Let T, C and $\mathbf{X}$ have a joint distribution, where T and C are non-negative 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 $\delta = \text{ind}(T \leqslant C)$. Then the indicator random variable δ equals 1 if failure occurs on or before the censoring time (if $T \leqslant C$) 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 $f(t \mid \mathbf{x})$ and $F(t \mid \mathbf{x})$ 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

$$\bar{F}(t \mid \mathbf{x}) = 1 - F(t \mid \mathbf{x}), \quad \lambda(t \mid \mathbf{x}) = f(t \mid \mathbf{x})/\bar{F}(t \mid \mathbf{x}) \quad \text{and} \quad \alpha(t \mid \mathbf{x}) = \log \lambda(t \mid \mathbf{x}), \quad t \geqslant 0.$$

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

A popular choice for the analysis of censored survival data with covariates is the proportional hazard model $\alpha(t \mid \mathbf{x}) = \alpha_0(t) + \mathbf{x}^T\beta$ introduced by Cox (1972), where $\alpha_0(\cdot)$ is the baseline hazard function and $\beta \in \mathbb{R}^M$ is a vector of parameters; see also Kalbfleisch & Prentice (1980), Miller (1981), Cox & Oakes (1984), Fleming & Harrington (1991) and Andersen *et al.* (1993). In practice it is more desirable to examine the covariate effects by using smooth, non-linear functions. The generalized additive model

$$\alpha(t \mid \mathbf{x}) = \alpha_0(t) + \alpha_1(x_1) + \alpha_2(x_2) + \cdots + \alpha_M(x_M)$$

considered by Hastie and Tibshirani (1990), Sleeper & Harrington (1990) and Gray (1992) is a refinement of Cox's model. Here $\alpha_0(\cdot), a_1(\cdot), \ldots, \alpha_M(\cdot)$ are smooth functions. In order to examine the interactions between covariates and time-varying coefficients, the generalized

additive model can further be refined. To motivate our approach, suppose $\mathbf{x} = (x_1, x_2)$ and write

$$\alpha(t \mid \mathbf{x}) = \alpha_0(t) + \alpha_1(x_1) + \alpha_2(x_2) + \alpha_{01}(t, x_1) + \alpha_{02}(t, x_2) + a_{12}(x_1, x_2),$$

where $\alpha_0(\cdot), a_1(\cdot), \alpha_2(\cdot), \ldots, a_{12}(\cdot)$ are smooth functions. Here $\alpha_0(\cdot)$, $\alpha_1(\cdot)$ and $\alpha_2(\cdot)$ are referred to as main effects, $\alpha_{12}(\cdot)$ is the interaction component and $\alpha_{01}(\cdot)$ and $\alpha_{02}(\cdot)$ are components involving time-varying coefficients.

Given a random sample, consider an estimate

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

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 $\hat{\alpha}(\cdot \mid \cdot)$ as an estimate of the log-hazard function $\alpha(\cdot \mid \cdot)$. Alternatively, we can think of it as an estimate of the corresponding best theoretical approximation

$$\alpha^*(t \mid \mathbf{x}) = \alpha_0^*(t) + \alpha_1^*(x_1) + \alpha_2^*(x_2) + \alpha_{01}^*(t, x_1) + \alpha_{02}^*(t, x_2) + \alpha_{12}^*(x_1, x_2) \tag{1.2}$$

to the log-hazard function, where best means having the maximum expected log-likelihood subject to the indicated form and each component is theoretically orthogonal to the corresponding lower order components. According to Stone (1994), the right sides of (1.1) and (1.2) are referred to as the ANOVA decompositions of $\hat{\alpha}$ and α^*, respectively. If the components of the ANOVA decomposition of α^* are estimated accurately by the corresponding ANOVA components of $\hat{\alpha}$, then examination of the components of the ANOVA decomposition of $\hat{\alpha}$ should shed light on the relationship of the survival time T to the covariates $\mathbf{X}$ through the function α^* and, to a lesser extent, through the function α.

More generally, in this paper we consider the approximation α^* to $\alpha = \log \lambda$ 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 α^*. 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 α^*. The problem of estimating the conditional density and survival functions are treated similarly by observing that

$$\bar{F}(t \mid \mathbf{x}) = \exp\left(-\int_0^t \lambda(u \mid \mathbf{x})\, du\right) = \exp\left(-\int_0^t \exp(\alpha(u \mid \mathbf{x}))\, du\right), \quad t \geqslant 0,$$

and

$$f(t \mid \mathbf{x}) = \exp(\alpha(t \mid \mathbf{x})) \exp\left(-\int_0^t \exp(\alpha(u \mid \mathbf{x}))\, du\right), \quad t \geqslant 0.$$

The rest of the paper is organized as follows. Section 2.1 provides a preliminary discussion of the ANOVA decomposition by introducing the relevant notation. 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 described and the corresponding existence and rate of convergence results are stated in section 2.4. In section 2.5 we briefly discuss the closely related adaptive methodology in Kooperberg *et al.* (1995). Section 2.6 contains a discussion of other work related to the current paper. Proofs of the results in sections 2.3 and 2.4 are given in section 3.

2. Statement of results

2.1. Preliminaries

To prepare for the discussion of the ANOVA decomposition and its existence in the next two sections, we begin with some notation. Given a non-empty subset s of $\{0, 1, \ldots, M\}$, let H_s denote the space of functions on $[0, \infty) \times \mathbb{R}^M$ that depend on the variable t if $0 \in s$ and on the variable x_j for $j \in s \cap \{1, \ldots, M\}$ and on no other variables. Let $H_\emptyset$ denote the space of constant functions on $[0, \infty) \times \mathbb{R}^M$. Let $\mathcal{S}$ be a non-empty collection of sub-sets 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 sub-set of s, then r is a member of $\mathcal{S}$. Let H denote the collection of functions of the form $a = \Sigma_{s \in \mathcal{S}} a_s$ with $a_s \in H_s$ for $s \in \mathcal{S}$. For example, the functions given by (1.1) and (1.2) can be described as a member of H by setting $\mathcal{S} = \{\emptyset, \{0\}, \{1\}, \{2\}, \{0, 1\}, \{0, 2\}, \{1, 2\}\}$ and

$$\hat{\alpha}(t \mid \mathbf{x}) = \sum_{s \in \mathcal{S}} \hat{\alpha}_s \quad \text{and} \quad \alpha^*(t \mid \mathbf{x}) = \sum_{s \in \mathcal{S}} \alpha_s^*,$$

where $\hat{\alpha}_\emptyset$ and $\alpha_\emptyset^*$ are constants, $\hat{\alpha}_{\{0\}}$ and $\alpha_{\{0\}}^*$ are functions of the variable t, $\hat{\alpha}_{\{1\}}$ and $\alpha_{\{1\}}^*$ are functions of the variable x_1, and so on.

2.2. Expected log-likelihood function

The conditional log-likelihood based on $(Y, \delta, \mathbf{X})$ is given by

$$\begin{aligned}\log \{[f(Y \mid \mathbf{X})]^\delta [\bar{F}(Y \mid \mathbf{X})]^{1-\delta}\} &= \delta \log \lambda(Y \mid \mathbf{X}) + \log \bar{F}(Y \mid \mathbf{X}) \\ &= \delta \log \lambda(Y \mid \mathbf{X}) - \int_0^Y \lambda(u \mid \mathbf{X})\, du.\end{aligned}$$

Using integration by parts, we get that

$$\begin{aligned}E\left(\int_0^Y \lambda(u \mid \mathbf{X})\, du \,\middle|\, \mathbf{X} = \mathbf{x}\right) &= \int \left(\int_0^t \lambda(u \mid \mathbf{x})\, du\right)(\bar{F}_C(t \mid \mathbf{x})\, dF(t \mid \mathbf{x}) + \bar{F}(t \mid \mathbf{x})\, dF_C(t \mid \mathbf{x})) \\ &= \int \lambda(t \mid \mathbf{x}) \bar{F}_C(t \mid \mathbf{x}) \bar{F}(t \mid \mathbf{x})\, dt.\end{aligned}$$

Thus the expected conditional log-likelihood is given by

$$\begin{aligned}&E\left(\delta \log \lambda(Y \mid \mathbf{X}) - \int_0^Y \lambda(u \mid \mathbf{X})\, du\right) \\ &\quad = \int\int \bar{F}_C(t \mid \mathbf{x})(\log \lambda(t \mid \mathbf{x}) f(t \mid \mathbf{x}) - \bar{F}(t \mid \mathbf{x}) \lambda(t \mid \mathbf{x}))\, dt\, f_X(\mathbf{x})\, d\mathbf{x},\end{aligned}$$

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

$$\Lambda(a) = \int\int \bar{F}_C(t \mid \mathbf{x})(a(t \mid \mathbf{x}) f(t \mid \mathbf{x}) - \bar{F}(t \mid \mathbf{x}) \exp(a(t \mid \mathbf{x})))\, dt\, f_X(\mathbf{x})\, d\mathbf{x}, \quad a \in H.$$

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

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 $\mathscr{T} \times \mathscr{X}$. Moreover, the survival function $\bar{F}(t \mid \mathbf{x})$ is bounded away from zero on $\mathscr{T} \times \mathscr{X}$.

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

Condition 2

$P(C \in \mathscr{T} \mid \mathbf{x}) = 1$ for $\mathbf{x} \in \mathscr{X}$ and $P(C = 1 \mid \mathbf{x})$ is bounded away from zero on $\mathscr{X}$.

This condition implies that $\bar{F}_C(t \mid \mathbf{x})$ is bounded away from zero on $[0, 1) \times \mathscr{X}$. 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 $\alpha^* \in H$ such that $\Lambda(\alpha^*) = \max_{a \in H} \Lambda(a)$. If $\alpha \in H$, then $\alpha^* = \alpha$ almost everywhere.

In the statement of theorem 1, "essentially uniquely determined" means that any two such functions are equal almost everywhere. Note that uniqueness of the components of α^* is not required in this theorem.

To establish the rates of convergence, it is necessary to have a unique ANOVA decomposition of α^*, which will be considered next. We first define inner products and orthogonality for functions on $\mathscr{T} \times \mathscr{X}$. Set

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

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

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

here $a \perp H_r^2$ means that $\langle a, a_r \rangle = 0$ for $a_r \in H_r^2$. Observe that $H_\emptyset^0 = H_\emptyset^2$ is the space of square-integrable functions on $\mathscr{T} \times \mathscr{X}$ that equal some constant almost everywhere. Let H^2 denote the space of all functions of the form $\Sigma_{s \in \mathscr{S}} a_s$, where $a_s \in H_s^2$ for $s \in \mathscr{S}$. 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 $\Sigma_{s \in \mathscr{S}} a_s$, where $a_s \in H_s^0$ for $s \in \mathscr{S}$; see lem. 3.1 of Stone (1994). We refer to $\Sigma_{s \in \mathscr{S}} a_s$ as the ANOVA decomposition of a, and we refer to H_s^0, $s \in \mathscr{S}$, as the components of H^2. Observe that $a_\emptyset = E(a \mid T \leqslant C)$. Given any non-empty set $s \in \mathscr{S}$, it follows from the orthogonality of $H_\emptyset^2$ and H_s^0 that $E(a_s \mid T \leqslant C) = 0$.

Let $\#(s)$ denote the number of members of s, set $d = \max_{s \in \mathscr{S}} \#(s)$, and assume that $d \geqslant 1$. 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 $\#(s) \geqslant 2$.

Suppose the function α^* in theorem 1 is a member of H^2. Then it can be written in an essentially unique manner in the form $\alpha^* = \Sigma_{s \in \mathscr{S}} \alpha_s^*$ where $\alpha_s^* \in H_s^0$ for $s \in \mathscr{S}$. The rate of convergence in estimating α^* depends on a smoothness condition on α_s^*, $s \in \mathscr{S}$, which will now be described.

Let $0 < \beta \leqslant 1$. A function a on $\mathscr{T} \times \mathscr{X}$ is said to satisfy a Hölder condition with exponent β if there is a positive number γ such that $|a(\mathbf{z}) - a(\mathbf{z}_0)| \leqslant \gamma |\mathbf{z} - \mathbf{z}_0|^\beta$ for $\mathbf{z}, \mathbf{z}_0 \in \mathscr{T} \times \mathscr{X}$; here $|\mathbf{z}|^2 = \Sigma_0^M z_j^2$ is the square of the Euclidean norm of $\mathbf{z} = (z_0, z_1, \ldots, z_M)$. Given an $(M+1)$-

tuple $\mathbf{i} = (i_0, i_1, \ldots, i_M)$ of non-negative integers, set $[\mathbf{i}] = i_0 + i_1 + \cdots + i_M$ and let $D^{\mathbf{i}}$ denote the differential operator defined by

$$D^{\mathbf{i}} = \frac{\partial^{[\mathbf{i}]}}{\partial z_0^{i_0} \cdots \partial z_M^{i_M}}.$$

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

Condition 3

There are p-smooth functions $\alpha_s^* \in H_s^0$, $s \in \mathscr{S}$, such that $\alpha^* = \Sigma_{s \in \mathscr{S}} \alpha_s^* \in H$ and $\Lambda(\alpha^*) = \max_{a \in H} \Lambda(a)$.

2.4. *Maximum likelihood estimation*

Let $K = K_n$ be a positive integer, and let I_k, $1 \leqslant k \leqslant K$, denote the sub-intervals of $[0, 1]$ defined by $I_k = [(k-1)/K, k/K)$ for $1 \leqslant k < K$ and $I_K = [1 - 1/K, 1]$. Let m and q be fixed integers such that $m \geqslant 0$ and $m > q \geqslant -1$. Let 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 $1 \leqslant k \leqslant K$;

and, if $q \geqslant 0$, 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, $1 \leqslant j \leqslant J$, denote the usual basis of S_n consisting of B-splines (see de Boor, 1978). Then $J = (m+1)K - (q+1)(K-1)$, so $K + m \leqslant J \leqslant (m+1)K$. Also, $B_j \geqslant 0$ on $[0, 1]$, $B_j = 0$ on the complement of an interval of length $(m+1)/K$ for $1 \leqslant j \leqslant J$, and $\Sigma_j B_j = 1$ on $[0, 1]$. Moreover, for $1 \leqslant j \leqslant J$, 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 $\mathscr{T} \times \mathscr{X}$. Given a sub-set s of $\{0, 1, \ldots, M\}$, let G_s denote the space spanned by the functions g on $\mathscr{T} \times \mathscr{X}$ of the form

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

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 $\delta_i = \text{ind}(T_i \leqslant C_i)$ for $1 \leqslant i \leqslant n$. Let $\langle \cdot, \cdot \rangle_n$ denote the sample inner product defined by

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

Given $s \in \mathscr{S}$ let G_s^0 denote the space of functions in G_s that are orthogonal (relative to $\langle \cdot, \cdot \rangle_n$) to G_r for every proper sub-set r of s. Also, set

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

The space G is said to be *non-identifiable* if there is a non-zero function g in the space such that $g(Y_i \mid \mathbf{X}_i) = 0$ 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 $\Sigma_{s \in \mathscr{S}} g_s$, where $g_s \in G_s^0$ for $s \in \mathscr{S}$; see lem. 3.2 of Stone (1994).

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

It follows from conditions 1, 2 and 4 (see lem. 3.8 of Stone, 1994) that

$$P(G \text{ is non-identifiable}) = o(1). \tag{2.1}$$

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

$$\prod_i \{[\lambda(Y_i \mid \mathbf{X}_i)]^{\delta_i} \bar{F}(Y_i \mid \mathbf{X}_i)\},$$

and the log-likelihood is given by

$$\sum_i \left(\delta_i \log \lambda(Y_i \mid \mathbf{X}_i) - \int_0^{Y_i} \lambda(u \mid \mathbf{X}_i)\, du \right).$$

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

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

Then the functions $B_{s\mathbf{j}}$, $\mathbf{j} \in \mathscr{J}_s$, which are non-negative and have sum one, form a basis of G_s.

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

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

and

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

Then the log-likelihood function can be written as

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

Thus, by the identifiability of G, the log-likelihood function is strictly concave except on an event whose probability tends to zero with n. Define $\hat{\boldsymbol{\theta}}$ so that $l(g(\cdot \mid \cdot\,; \hat{\boldsymbol{\theta}})) = \max_{g \in G} l(g)$, and consider $\hat{\alpha} = g(\cdot \mid \cdot\,; \hat{\boldsymbol{\theta}})$ as the maximum likelihood estimate of α^* in G. It follows from the strict concavity of the log-likelihood function that the Newton–Raphson method can be used to compute these maximum-likelihood estimates.

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

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

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

so

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

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*

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

so

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

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 another non-trivial justification of the heuristic dimensionality reduction principle discussed by Stone (1986). When $d = M$, the rate is optimal according to Stone (1982) and Hasminskii & Ibragimov (1990).

Conditions 1 and 2 and the compactness of $\mathcal{X}$ are required for the validity of theorem 3 and corollary 1 and, more generally, for the mathematical tractability of the estimates studied in this paper. The closely related adaptive methodology in Kooperberg *et al.* (1995), which is summarized in section 2.5, does not depend on these conditions for its applicability. Also, in condition 3 it is assumed that the various components in the ANOVA decomposition of α^* have the same smoothness parameter. If these components have different smoothness parameters, then the rate of convergence of the corresponding estimates of these components and of their sum α^* is governed by the smallest such smoothness parameter.

2.5. Adaptive methodology for hazard regression

The pragmatic importance of the theoretical results in section 2.4 is based on the principle that, in the context of the present paper and in similar contexts involving (at least) regression, generalized regression, density estimation, conditional density estimation and spectral density estimation, the successful development of mathematical theory for non-adaptive procedures under mildly restrictive conditions implies that practically useful methodology involving closely related adaptive procedures can be developed.

With this motivation, the authors of the present paper have used splines and their selected tensor products to develop an adaptive methodology for estimating the conditional log-hazard function. This methodology, referred to as HARE and described in Kooperberg *et al.* (1995), is similar in spirit to MARS (Friedman, 1991).

The HARE approach can be described by introducing the notion of an allowable space. Recall the description of the space G in section 2.4. A family $\mathcal{G}$ of such spaces is said to be *allowable* if it satisfies the following properties:

1. each $G \in \mathcal{G}$ is a linear space having dimension $p \geq p_{\min}$;
2. there is only one $G \in \mathcal{G}$ with dimension $p_{\min}$;

3. if $G \in \mathcal{G}$ has dimension $p > p_{\min}$, there is at least one sub-space $G_0 \in \mathcal{G}$ of G with dimension $p-1$;
4. if $G_0 \in \mathcal{G}$ has a dimension p, there is at least one space $G \in \mathcal{G}$ with dimension $p+1$ and containing G_0 as a sub-space.

We refer to $G_{\min} \in \mathcal{G}$ with minimal dimension $p_{\min}$ as the minimal allowable space.

In order to avoid numerical integration and other complications in the context of stepwise knot addition, we use linear (rather than quadratic or cubic) splines. The allowable spaces are constructed as follows. Let K_0 be a non-negative integer; if $K_0 \geq 1$, let t_k, $1 \leq k \leq K_0$ be distinct positive numbers, and consider the basis functions $B_{0k}(t) = (t_k - t)_+$, $1 \leq k \leq K_0$, where $t_+ = \max(t, 0)$. Next, for $1 \leq m \leq M$, let K_m be an integer with $K_m \geq -1$; if $K_m \geq 0$, consider the basis function $B_{m0}(x_m) = x_m$; if $K_m \geq 1$, let x_{mk}, $1 \leq k \leq K_m$ be distinct real numbers and consider the additional basis functions $B_{mk}(x_m) = (x_m - x_{mk})_+$, $1 \leq k \leq K_m$.

Let G be the linear space having basis functions 1, $B_{0k}(t)$ for $1 \leq k \leq K_0$, $B_{mk}(x_m)$ for $1 \leq m \leq M$ and $0 \leq k \leq K_m$, and perhaps certain tensor products of two such basis functions. It is required that if $B_{mj}(x_m)B_{0k}(t)$ be among the basis functions for some $j \geq 1$, then $B_{m0}(x_m)B_{0k}(t) = x_m B_{0k}(t)$ be among the basis functions. Similarly, it is required that if $B_{lj}(x_l)B_{mk}(x_m)$ be among the basis functions for some $j \geq 1$, then $B_{l0}(x_l)B_{mk}(x_m) = x_l B_{mk}(x_m)$ and hence $x_l x_m$ be among the basis functions. It is easy to check that the collection $\mathcal{G}$ of such spaces is allowable. In particular, the minimal allowable space $G_{\min}$ is the space of constant functions given by $\alpha(t \mid \boldsymbol{\theta}) = \alpha(t \mid \mathbf{X}; \boldsymbol{\theta}) = \theta_1$, $t \geq 0$ (that is, $p = 1$ and $B_1 = 1$). The corresponding conditional distribution of T given $\mathbf{X}$ is exponential with mean $\exp(-\theta_1)$, which does not depend on $\mathbf{X}$.

Initially, the minimal allowable space is used to model $\alpha(t \mid \mathbf{X})$. Then we proceed with stepwise addition. Here we successively replace the $(p-1)$-dimensional allowable space G_0 by a p-dimensional allowable space G containing G_0 as a sub-space, choosing among the various candidates for a new basis function by maximizing the absolute value of the corresponding Rao statistic, see (6e.3.6) of Rao (1973).

When the number of basis functions reaches a specified number (the default being chosen in light of the theoretical results presented in the current paper), we stop the stepwise addition stage and proceed to stepwise deletion. Here we successively replace the p-dimensional allowable space G by a $(p-1)$-dimensional allowable sub-space G_0 until we arrive at the minimal allowable space, at each step choosing the candidate space G_0 so that the Wald statistic for a basis function that is in G but not in G_0 is smallest in magnitude.

In this process, we obtain a sequence of models indexed by v with the vth model having p_v parameters and fitted log-likelihood $\hat{l}_v$. The final model is selected from this sequence so as to minimize (say) the Bayesian Information Criterion $\mathrm{BIC}_v = -2\hat{l}_v + (\log n)p_v$.

2.6. Other 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 & 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 the context of non-parametric regression, generalized regression, density estimation and conditional density estimation were given in Stone (1994).

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 & Tibshirani (1990), Sleeper & Harrington (1990) and Gray (1992) using smoothing splines. Kooperberg & Stone (1992) consider logspline density estimation (without covariates) under right, left and interval censoring. Time-dependent coefficient models have been discussed by Zucker & Karr (1990) and Hastie & Tibshirani (1993) using penalized partial likelihood method. Extensions of the present theory to handle time-dependent covariates as considered by O'Sullivan (1993) and bivariate censored variables should be practically useful.

3. Proofs

3.1. Proof of theorem 1

Write

$$\Lambda(a) = \int\int (a\lambda - \exp(a))\bar{F}_C \bar{F} f_X, \quad a \in H.$$

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

$$a\lambda - \exp(a) \leqslant A - \varepsilon|a|, \quad a \in H.$$

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

$$\Lambda(a) \leqslant A - \varepsilon \int\int |a| f f_X, \quad a \in H.$$

Thus, if $\int\int |a| f f_X = \infty$, 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 $\int\int |a_k| f f_X$, $k \geqslant 1$, are bounded.

Let α_1 and α_2 be functions in H such that $\Lambda(\alpha_1) > -\infty$ and $\Lambda(\alpha_2) > -\infty$. For $u \in [0, 1]$, set $\alpha^{(u)} = (1-u)\alpha_1 + u\alpha_2$ and $\Psi(u) = \Lambda(\alpha^{(u)})$. Then, by the concavity of $\alpha^{(u)}\lambda - \exp(\alpha^{(u)})$ as a function of u, $\Psi(\cdot)$ is a concave function. (Note that if α_1 and α_2 are bounded, then

$$\Psi''(u) = -\int\int (\alpha_2 - \alpha_1)^2 \exp(\alpha^{(u)})\bar{F}_C \bar{F} f_X.)$$

It follows from the argument of th. 4.1 in Stone (1994) that there is an integrable function α^* such that $a_k \to \alpha^*$ in measure as $k \to \infty$. By lem. 4.1 of Stone (1994), we can assume that $\alpha^* \in H$. It follows from Fatou's Lemma that $\Lambda(a_k) \to \Lambda(\alpha^*) = L = \max_{a \in H} \Lambda(a)$ as $k \to \infty$. Furthermore, if $a \in H$ and $\Lambda(a) = \Lambda(\alpha^*)$, then it follows from the concavity described above that $a = \alpha^*$ almost everywhere. Hence the first statement of the theorem is valid. The second statement follows from the fact that $a\lambda - \exp(a)$, as a function of a, has a unique maximum at $\alpha = \log \lambda$.

3.2. Proof of theorem 2

Set $\|g\|_\infty = \sup_{t \in \mathscr{T}, \mathbf{x} \in \mathscr{X}} |g(t \mid \mathbf{x})|$. Throughout this sub-section, it is assumed that conditions 1–4 hold.

Lemma 1

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

$$-M_1\|a-\alpha^*\|^2 \leqslant \Lambda(a)-\Lambda(\alpha^*) \leqslant -M_2\|a-\alpha^*\|^2$$

for all $a \in H$ *with* $\|a\|_\infty \leqslant U$.

Proof. Given $a \in H$ with $\|a\|_\infty \leqslant U$ and given $u \in [0, 1]$, set $\alpha^{(u)} = (1-u)\alpha^* + ua$. Then

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

and, by integration by parts,

$$\Lambda(a)-\Lambda(\alpha^*) = \int_0^1 (1-u)\frac{d^2}{du^2}\Lambda(\alpha^{(u)})\,du$$
$$= -\int_0^1 (1-u)\int\int (a-\alpha^*)^2 \exp(\alpha^{(u)})\bar{F}_C\bar{F}f_X\,du.$$

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

The next result is lem. 4.3 of Stone (1994).

Lemma 2
There is a positive constant M_3 *such that* $\|g\|_\infty \leqslant M_3 J^{d/2}\|g\|$ *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 $\alpha_n^* \in G$ such that $\Lambda(\alpha_n^*) = \max_{g\in G}\Lambda(g)$.

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

Proof. By condition 3 and th. 12.8 of Schumaker (1981), there are functions $g_n \in G$ for $n \geqslant 1$ and a positive constant M_4 such that $\|g_n-\alpha^*\|_\infty \leqslant M_4 J^{-p}$. Consequently, $\|g_n-\alpha^*\|^2 \leqslant M_4^2 J^{-2p}$. By lemma 1, there is a positive constant M_5 such that

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

Let b be a positive constant. Choose $g \in G$ with $\|g-\alpha^*\|^2 = bJ^{-2p}$. Then

$$\|g_n-g\|^2 \leqslant 2(\|g_n-\alpha^*\|^2 + \|\alpha^*-g\|^2) \leqslant 2(b+M_4^2)J^{-2p}.$$

Since $p > d/2$, it follows from lemma 2 that, for J sufficiently large,

$$\|g\|_\infty \leqslant \|g-g_n\|_\infty + \|g_n-\alpha^*\|_\infty + \|\alpha^*\|_\infty \leqslant 1 + \|\alpha^*\|_\infty.$$

Thus by lemma 1, there is a positive constant M_6 such that, for J sufficiently large,

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

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,

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

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

$$\|\alpha_n^*-\alpha^*\|^2 < bJ^{-2p}$$

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

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

Consequently, $\|\alpha_n^* - \alpha^*\|_\infty = O(J^{d/2-p})$. □

Suppose condition 4 holds. Let τ_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

$$\alpha_n^*(\cdot \mid \cdot) = g(\cdot \mid \cdot\,; \boldsymbol{\theta}^*) = \sum_{s \in \mathscr{S}} g_s(\cdot \mid \cdot\,; \boldsymbol{\theta}^*).$$

Lemma 4

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

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

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

Proof. Write

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

where

$$W_i = \delta_i g(Y_i \mid X_i; \boldsymbol{\theta}) - \int_0^{Y_i} \exp(g(u \mid \mathbf{X}_i; \boldsymbol{\theta}))\, du$$

$$- \delta_i g(Y_i \mid \mathbf{X}_i; \boldsymbol{\theta}^*) + \int_0^{Y_i} \exp(g(u \mid \mathbf{X}_i; \boldsymbol{\theta}^*))\, du.$$

By lemma 2, $\|g(\cdot \mid \cdot\,; \boldsymbol{\theta}) - g(\cdot \mid \cdot\,; \boldsymbol{\theta}^*)\|_\infty = O(J^{d/2}\tau_n)$ for $g(\cdot \mid \cdot\,; \boldsymbol{\theta})$ satisfying $\|g - \alpha_n^*\| \leq b\tau_n$. Thus there is a positive constant M_7 such that

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

for $g(\cdot \mid \cdot\,; \boldsymbol{\theta})$ satisfying $\|g - \alpha_n^*\| \leq b\tau_n$. It now follows from condition 1 that $|W_i| = O(J^{d/2}\tau_n)$. It also follows from this condition that

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

and

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

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

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

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

The next result is essentially that of lem. 4.8 of Stone (1994).

Lemma 5

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

$$\{g : g \in G \text{ and } \|g - \alpha_n^*\| \leq b\tau_n\}$$

can be covered by $O(\exp(M_8 J^d \log n))$ *sub-sets 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, $l(g) < l(\alpha_n^*)$ *for all* $g \in G$ *such that* $\|g - \alpha_n^*\| = b\tau_n$.

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

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

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

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

for $g_1 = g(\cdot \mid \cdot\,; \theta_1)$ and $g_2 = g(\cdot \mid \cdot\,; \theta_2) \in G$ such that $\|g_i - \alpha_n^*\| \leqslant b\tau_n$ for $i = 1, 2$. The desired result follows from lemma 1, with α^* replaced by α_n^* and H by G, and lemmas 4 and 5. □

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

Proof. The set $G_0 = \{g \in G: \|g - \alpha_n^*\| \leqslant b\tau_n\}$ is compact with boundary $\{g \in G: \|g - \alpha_n^*\| = b\tau_n\}$. By lemma 6, the log-likelihood function has a local maximum in the interior of G_0. It follows from condition 2 and the formula for the Hessian of the log-likelihood function (see (3.3) below) that the log-likelihood function is strictly concave except on an event whose probability tends to zero as $n \to \infty$. Consequently, $\|\hat{\alpha} - \alpha_n^*\| = o_P(\tau_n)$, so we conclude from lemma 2 that $\|\hat{\alpha} - \alpha_n^*\|_\infty = o_P(J^{d/2}\tau_n) = o_P(1)$. □

The proof of theorem 2 now follows from lem. 3.2 and 3.8 of Stone (1994) and lemma 7.

3.3. Proof of theorem 3

The proof of the next result is similar to that of lem. 5.3 of Stone (1994).

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

$$\|\alpha_{ns}^* - \alpha_s^*\|^2 = O_P(J^{-2p} + J^d/n), \quad s \in \mathscr{S}.$$

Recall that the log-likelihood function is given by

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

and that $I = \Sigma_s \#(\mathscr{J}_s)$. Let

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

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

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

Let

$$\frac{\partial^2 l(g)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^{\mathrm{T}}}$$

denote the Hessian of $l(g)$; that is, the $I \times I$ matrix having entries

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

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

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

It follows from the maximum likelihood equation that

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

We claim that

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

and that there is positive constant M_9 such that

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

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

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

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

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

and

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

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

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

Hence

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

Since

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

we conclude that $E|\mathbf{S}(\boldsymbol{\theta}^*)|^2 = O(n)$ and hence that (3.5) is valid.

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

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

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

$$\lim_{n \to \infty} P(\|\alpha_n^*\|_\infty \leqslant U \text{ and } \|\hat{a}\|_\infty \leqslant U) = 1. \tag{3.10}$$

It follows from (3.9) and (3.10) that there is a positive constant M_{10} such that, except on an event whose probability tends to zero with n,

$$\boldsymbol{\beta}^{\mathrm{T}} \mathbf{D} \boldsymbol{\beta} \leqslant -M_{10} \sum_i \int_0^{Y_i} g^2(u \mid \mathbf{X}_i; \boldsymbol{\beta})\, du. \tag{3.11}$$

By conditions 1 and 2, there is a positive constant $M_{11} > 0$ such that

$$P(Y \geqslant 1 \mid \mathbf{X} = \mathbf{x}) > M_{11} \quad \text{for } \mathbf{x} \in \mathscr{X}. \tag{3.12}$$

Let $I_n = \{i: 1 \leqslant i \leqslant n,\ Y_i \geqslant 1\}$. Write $g(\cdot \mid \cdot\,; \boldsymbol{\beta}) = \Sigma_{s \in \mathscr{S}} g_s(\cdot \mid \cdot\,; \boldsymbol{\beta})$ with $\boldsymbol{\beta} \in \mathbb{R}^I$ chosen such that $g_s(\cdot \mid \cdot\,; \boldsymbol{\beta}) \in G_s^0$ for $s \in \mathscr{S}$. Let $\Theta \subset \mathbb{R}^I$ be the set of all such $\boldsymbol{\beta}$. According to lem. 3.7 of Stone (1994), there is a positive constant M_{12} such that, except on an event whose probability tends to zero with n,

$$\sum_{i \in I_n} g^2(u \mid \mathbf{X}_i; \boldsymbol{\beta}) \geqslant n M_{12} E g^2(u \mid \mathbf{X}_1; \boldsymbol{\beta}), \quad \boldsymbol{\beta} \in \Theta \text{ and } 0 \leqslant u \leqslant 1. \tag{3.13}$$

Hence there is a positive constant M_{13} such that

$$\begin{aligned} \sum_i \int_0^{Y_i} g^2(u \mid \mathbf{X}_i; \boldsymbol{\beta})\, du &\geqslant \int_0^1 \sum_{i \in I_n} g^2(u \mid \mathbf{X}_i; \boldsymbol{\beta})\, du \\ &\geqslant n M_{12} \int_0^1 E g^2(u \mid \mathbf{X}_1; \boldsymbol{\beta})\, du \\ &\geqslant M_{13} n \int_{\mathscr{X}} \int_{\mathscr{T}} g^2(u \mid \mathbf{x}; \boldsymbol{\beta})\, du\, d\mathbf{x}, \quad \boldsymbol{\beta} \in \Theta. \end{aligned} \tag{3.14}$$

According to conditions 1 and 4 and lem. 3.6 in Stone (1994), there is a $M_{14} > 0$ such that, except on an event whose probability tends to zero with n,

$$\int_{\mathscr{X}} \int_{\mathscr{T}} g^2(u \mid \mathbf{x}; \boldsymbol{\beta})\, du\, d\mathbf{x} \geqslant M_{14} \sum_{s \in \mathscr{S}} \int_{\mathscr{X}} \int_{\mathscr{T}} g_s^2(u \mid \mathbf{x}; \boldsymbol{\beta})\, du\, d\mathbf{x}, \quad \boldsymbol{\beta} \in \Theta. \tag{3.15}$$

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

$$\int_{\mathscr{X}} \int_{\mathscr{T}} g_s^2(u \mid \mathbf{x}; \boldsymbol{\beta})\, du\, d\mathbf{x} \geqslant C J^{-\#(s)} \sum_j \beta_{sj}^2, \quad s \in \mathscr{S} \text{ and } \boldsymbol{\beta} \in \Theta,$$

and hence that

$$\sum_{s \in \mathscr{S}} \int_{\mathscr{X}} \int_{\mathscr{T}} g_s^2(u \mid \mathbf{x}; \boldsymbol{\beta})\, du\, d\mathbf{x} \geqslant C J^{-d} |\boldsymbol{\beta}|^2, \quad \boldsymbol{\beta} \in \Theta. \tag{3.16}$$

Equation (3.6) follows from (3.11)–(3.16) applied to $\boldsymbol{\beta} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$. This completes the proof of theorem 3. □

Acknowledgments

C.K. was supported in part by a grant from the Graduate School Fund of the University of Washington; C.J.S. was supported in part by National Science Foundation Grant DMS-

9204247; and Y.K.T. was supported in part by a Research Council Grant from the University of North Carolina.

References

Andersen, P. K., Borgan, Ø., Gill, R. D. & 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. & Oakes, D. (1984). *Analysis of survival data*. Chapman & Hall, London.

de Boor, C. (1978). *A practical guide to splines*. Springer-Verlag, New York.

Fleming, T. R. & 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. & Ibragimov, I. (1990). Kolmogorov's contributions to mathematical statistics. *Ann. Statist.* **18,** 1011–1016.

Hastie, T. & Tibshirani, R. (1990). *Generalized additive models*. Chapman & Hall, London.

Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models (with discussion). *J. Roy. Statist. Soc. Ser. B* **55,** 757–796.

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

Kalbfleisch, J. D. & Prentice, R. L. (1980). *The statistical analysis of failure time data*. Wiley, New York.

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

Kooperberg, C., Stone, C. J. & Truong, Y. K. (1995). Hazard regression. *J. Amer. Statist. Assoc.* **90,** 78–94.

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

O'Sullivan, F. (1993). Nonparametric estimation in the Cox Model. *Ann. Statist.* **21,** 124–145.

Rao, C. R. (1973). *Linear statistical inference and its applications*, second edition. Wiley, New York.

Schumaker, L. L. (1981). *Spline functions: basic theory*. Wiley, New York.

Sleeper, L. A. & 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* (eds. T. W. Anderson, K. B. Athreya & D. L. Iglehart), 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 flexible log-spline response models. *Ann. Statist.* **19,** 1832–1854.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Ann. Statist.* **22,** 118–184.

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

Received May 1993, in final form November 1994

Charles Kooperberg, Department of Statistics, University of Washington, Seattle, Washington 98195, USA.

Charles J. Stone, Department of Statistics, University of California, Berkeley, California 94720, USA.

Young K. Truong, School of Public Health, Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27599-7400, USA.