The Annals of Statistics
1999, Vol. 27, No. 5, 1536–1563

EFFICIENT ESTIMATION OF THE PARTLY LINEAR ADDITIVE COX MODEL

BY JIAN HUANG

University of Iowa

The partly linear additive Cox model is an extension of the (linear) Cox model and allows flexible modeling of covariate effects semiparametrically. We study asymptotic properties of the maximum partial likelihood estimator of this model with right-censored data using polynomial splines. We show that, with a range of choices of the smoothing parameter (the number of spline basis functions) required for estimation of the nonparametric components, the estimator of the finite-dimensional regression parameter is root-n consistent, asymptotically normal and achieves the semiparametric information bound. Rates of convergence for the estimators of the nonparametric components are obtained. They are comparable to the rates in nonparametric regression. Implementation of the estimation approach can be done easily and is illustrated by using a simulated example.

1. Introduction. The partly linear additive Cox model is an extension of the Cox (1972) model, in which the log-relative risk takes the partly linear additive form. So the conditional hazard of the failure time given the covariate value $z = (x, w) \in R^d \times R^J$ is modeled as

$$\lambda(t|x, z) = \lambda_0(t)\exp\bigl(x'\beta + \phi_1(w_1) + \cdots + \phi_J(w_J)\bigr), \tag{1.1}$$

where λ_0 is the unknown baseline hazard function, β is a d-dimensional regression parameter and $\phi_1, \ldots, \phi_J$ are unknown and smooth functions. In many situations, our main interest is in estimating the regression parameter β, which provides a concise and easily interpretable measure of the effect of the covariate X in the presence of the auxiliary covariate W. For instance, when X is a treatment covariate and W is a vector of covariates describing other characteristics of the patients, β can be interpreted as a measure of the treatment effect after adjusting for the effect of W. Although a categorical-type covariate X is our main motivation for this model, X can also be a continuous-type variable or a mixture of the two types. In the proportional hazards model framework with multidimensional covariates, this model allows flexible modeling of the covariate effect and at the same time maintains the features of being parsimonious and easy to interpret enjoyed by the Cox model.

Received January 1998; revised May 1999.

AMS 1991 *subject classifications*. Primary 62G05, 62G20; secondary 62G07, 62P99.

Key words and phrases. Additive regression, asymptotic normality, right-censored data, partial likelihood, polynomial splines, projection, rate of convergence, semiparametric information bound.

Model (1.1) is closely related to the partly linear Cox model,

$$\lambda(t|x,w) = \lambda_0(t)\exp(x'\beta_0 + b(\omega)), \tag{1.2}$$

where b: $R^J \to R$. In this model, no further assumption is made on the form of b. For high-dimensional covariate W, it may require unrealistic large samples to estimate this model because of "curse of dimensionality." Indeed, a range of models between and beyond (1.1) and (1.2) can be considered. For example, an ANOVA type decomposition for b can be considered, and model (1.1) can be viewed as a first order approximation. Interaction between the variables in X and W can also be considered. Excellent discussions on these issues can be found in Stone (1984, 1994). We focus on the partly linear additive model (1.1) because it directly generalizes the (linear) Cox model. See also the discussions in Hastie and Tibshirani (1986, 1990).

Many authors have considered nonparametric and semiparametric modeling of covariate effects on the censored failure time. For example, Sasieni (1992a, b) calculated an information bound for estimating β in model (1.2) and suggested using a spline-based partial likelihood to estimate this model. His calculation suggests that it is possible to estimate β at the usual root-n rate of convergence despite the presence of two nonparametric functions and despite that the function b cannot be estimated at the root-n rate. Grambsch, Therneau and Fleming (1990) and Fleming and Harington [(1991), Section 4.5, pages 163–168] proposed using smoothed martingale residuals to explore the functional form of the covariate effect in the Cox model. The martingale residual approach was further discussed by Grambsch, Therneau and Fleming (1995). Hastie and Tibshirani (1986, 1990) have considered a fully nonparametric additive Cox model in exploratory data analysis. Their estimation approach is to maximize a penalized partial likelihood. Kooperberg, Stone and Truong (1995) considered a general nonparametric hazard regression problem. Their approach is to maximize the likelihood function over an approximating parameter space consisting of sums of tensor products of polynomial splines as in Stone (1994). They established the rate of convergence of their estimator. Several generalizations of the Cox model have also been studied in the literature. For example, O'Sullivan (1993) considered the proportional hazards model with a fully nonparametric relative risk function. He obtained the rate of convergence of the penalized partial likelihood estimator uniformly with respect to the penalty parameter. Sasieni (1992a) calculated the information bound for the continuously stratified Cox model and suggested a simple form of kernel-smoothed partial likelihood estimator. Dabrowska (1997) proved asymptotic normality of the estimators of the regression parameter and the stratified cumulative hazard in this model based on a general kernel-smoothed partial likelihood. A survey of other regression models for censored survival data can be found in Andersen, Borgan, Gill and Keiding [(1993), Chapter VII].

In this paper, we study the asymptotic properties of the partial likelihood estimator of β and $(\phi_{01}, \ldots, \phi_{0J})$ of model (1.1) using polynomial splines. The

use of polynomial splines in estimating the fully nonparametric additive Cox model based on the partial likelihood was first suggested by Stone (1986b). It appears that systematic study of this estimation approach for the partly linear additive model (1.1) has not been done in the literature. Although previous results on the asymptotic normality of the maximum partial likelihood estimator in the (linear) Cox model [Tsiatis (1981), Andersen and Gill (1982)] and the information calculation (see Section 4) suggests that β should be estimable at the usual root-n rate of convergence in the present model, the proof is complicated by the presence of the nonparametric component ϕ_j's in the partial likelihood and the fact that their estimators converge at rates slower than root-n. We deal with these difficulties by using some results from empirical process theory and the projection idea in information calculation for semiparametric models. Under appropriate conditions, we show that, with a range of choices of the smoothing parameter (the number of B-spline basis functions) required for estimation of the nonparametric components, the maximum partial likelihood estimator of β is root-n consistent, asymptotically normal and achieves information bound, although the convergence rate of the estimator of the nonparametric part is slower than root-n. The result that a range of the smoothing parameter is allowed for the asymptotic normality of $\hat{\beta}_n$ suggests that the first-order asymptotic performance of $\hat{\beta}_n$ is relatively insensitive to the specification of the smoothing parameter. This differs from nonparametric curve estimation in which the optimal choice of the smoothing parameter is required to achieve the optimal rate of convergence. Rates of convergence for nonparametric component estimators are also obtained. These rates are comparable to those obtained in nonparametric regression.

The organization of this paper is as follows. Section 2 describes the estimator using polynomial splines. An example is included to illustrate the computation of the estimator in Splus. The main results are stated in Section 3. In Section 4, we calculate the information bound for β in the partly linear additive Cox model. Section 5 contains proofs of the main results. Section 6 briefly discusses some aspects of incorporating time-dependent covariates in model (1.1). Several technical details are put together in the Appendix.

2. Definition and computation of the estimator. Let T^u and T^c be the failure time and censoring time, respectively. The observable random variable is $(T, \Delta, Z) \in R^+ \times \{0, 1\} \times R^{d+J}$, where $T = \min\{T^u, T^c\}$, $\Delta = I\{T^u \leq T^c\}$, $Z = (X, W)$ with $X \in R^d$, $W \in R^J$. Throughout, we assume that T^u and T^c are conditionally independent given the covariate Z. Let (T_i, Δ_i, Z_i), $i = 1, \ldots, n$ be an independent random sample identically distributed as (T, Δ, Z).

We assume that W takes values in $[a, b]^J$ where a and b are finite numbers. Let $a = \xi_0 < \xi_1 < \cdots < \xi_K < \xi_{K+1} = b$ be a partition of $[a, b]$ into K subintervals $I_{Kt} = [\xi_t, \xi_{t+1})$, $t = 0, \ldots, K - 1$ and $I_{KK} = [\xi_K, \xi_{K+1}]$, where $K \equiv K_n = n^v$ with $0 < v < 0.5$ is a positive integer such that $\max_{1 \leq k \leq K+1} |\xi_k - \xi_{k-1}| = O(n^{-v})$. The precise range of v will be given in Theorem 3.3 in

Section 3. Let $\mathscr{S}_n$ be the space of polynomial splines of order $l \geq 1$ consisting of functions s satisfying (i) the restriction of s to I_{Kt} is a polynomial of order l for $1 \leq t \leq K$; (ii) for $l \geq 2$ and $0 \leq l' \leq l-2$, s is l' times continuously differentiable on $[a, b]$. This definition is phrased after Stone (1985), which is a descriptive version of Schumaker [(1981), page 108, Definition 4.1].

Let Φ_n be the collection of functions ϕ on $[a, b]^J$ with the additive form $\phi(w) = \phi_1(w_1) + \cdots + \phi_J(w_J)$, where each component ϕ_j belongs to $\mathscr{S}_n$. According to Schumaker [(1981), page 117, Corollary 4.10], there exists a local basis $\{B_t,\ 1 \leq t \leq q_n\}$ for $\mathscr{S}_n$, where $q_n \equiv K_n + l$. Thus for any $\phi_j \in \mathscr{S}_n$, we can write

$$\phi_j(z_j) = \sum_{t=1}^{q_n} b_{jt} B_t(z_j), \qquad 1 \leq j \leq J. \tag{2.1}$$

Let $\mathbf{b} = \{b_{jt}\colon 1 \leq j \leq J,\ 1 \leq t \leq q_n\}$ be the collection of all the coefficients in the representation (2.1). Under suitable smoothness assumptions, ϕ_{0j}'s can be well approximated by functions in $\mathscr{S}_n$. Therefore, we seek a member of Φ_n along with a value of β that maximizes the partial likelihood function. Specifically, let $\hat{\theta} \equiv (\hat{\beta}_n, \hat{\mathbf{b}}_n)$ with $\hat{\mathbf{b}}_n = \{\hat{b}_{jt}\colon 1 \leq j \leq J,\ 1 \leq t \leq q_n\}$ be the value that maximizes

$$l_n(\beta, \phi) = n^{-1} \sum_{i=1}^{n} \Delta_i \Bigg\{ X_i'\beta + \phi(W_i) - \log \sum_{k:\, T_k \geq T_i} \exp\big[X_k'\beta + \phi(W_k)\big] \Bigg\}, \tag{2.2}$$

with $\phi(W_i) = \sum_{j=1}^{J} \phi_j(W_{ji})$, $1 \leq i \leq n$ where ϕ_j is given in (2.3), with respect to $(\beta, \mathbf{b}) \in R^d \times R^{q_n}$. Because the regression function ϕ_j's can only be identified up to an additive constant, we will center the estimators of ϕ_j's as follows. Let

$$\hat{\phi}_{jn}^*(z_j) = \sum_{t=1}^{q_n} \hat{b}_{jt} B_t(z_j) \quad \text{and} \quad \overline{\phi}_{jn}^* = \frac{\sum_{i=1}^{n} \Delta_i \hat{\phi}_{jn}^*(W_{ji})}{\sum_{i=1}^{n} \Delta_i}.$$

The resulting estimator of ϕ_j is defined to be

$$\hat{\phi}_{jn}(z_j) = \hat{\phi}_{jn}^*(z_j) - \overline{\phi}_{jn}^*, \qquad 1 \leq j \leq J.$$

So $\hat{\phi}_{jn}$ is a centered version of $\hat{\phi}_{jn}^*$ and satisfies $\sum_{i=1}^{n} \Delta_i \hat{\phi}_{jn}(W_{ji}) = 0$, $1 \leq j \leq J$. Notice that $(\hat{\beta}_n, \hat{\phi}_{1n}^*, \ldots, \hat{\phi}_{Jn}^*)$ maximizes the partial likelihood if and only if $(\hat{\beta}_n, \hat{\phi}_{1n}, \ldots, \hat{\phi}_{Jn})$ maximizes the partial likelihood. The use of this particular form of centering instead of the usual centering by average is to simplify the asymptotic analysis; see the comments in Section 3.

To make statistical inferences about β, it is necessary to know or to approximate the sampling distribution of $\hat{\beta}_n$. As stated in the next section, the distribution of $\hat{\beta}_n$ can be approximated by a normal distribution in the large sample sense. Unfortunately, the variance matrix of this normal distri-

bution cannot be expressed in terms of quantities that can be easily estimated. We suggest using the (inverse of the) observed partial information matrix, taking into account that we are also estimating the "nuisance" parameter $\mathbf{b}$, to estimate the variance matrix of $\hat{\beta}_n$ as in the (linear) Cox model. This variance estimator is available from any program that fits the Cox regression model. We have not been able to prove the consistency of this variance estimator. Heuristics based on the finite-dimensional parametric model and some limited simulation suggest that this estimator should work well. An example is given at the end of this section.

We have used a prespecified partition $\{\xi_t,\ 1 \le t \le K_n\}$ and fixed basis functions. It is probably preferable to adaptively select the partition and the basis functions. Large sample theory of data-driven procedures for the present problem appears to be extremely difficult and is beyond the scope of this paper. On the other hand, if our main purpose is to estimate β, then any reasonable choice of $\{\xi_t,\ 1 \le t \le K_n\}$ may work well. This is because as long as it guarantees that the estimators of ϕ_j's converge at a certain rate, which may be much slower than $n^{1/2}$, then $\hat{\beta}_n$ has $n^{1/2}$ rate of convergence and is asymptotically normal. See Theorem 3.3 and Remark 3.1 in Section 3, where the range of K_n that ensures asymptotic normality of $\hat{\beta}_n$ is given.

Although other estimation approaches can also be used, such as the penalized partial likelihood method used by Hastie and Tibshirani (1990) and O'Sullivan (1993), the above method has the advantage that it can be implemented with the existing Cox regression program. For example, in Splus [Version 3.4, (1996) MathSoft Inc.], two functions `coxph` and `bs` can accomplish most of the computation, where `coxph` is for fitting the Cox model and `bs` creates a basis matrix for polynomial splines. For the `bs` function, it has arguments for knots placement and `degree` of the polynomials. The default value 3 of `degree` gives the cubic-spline basis. There are two approaches to the placement of knots. The first is to explicitly specify the knots such as $\texttt{bs}(x, \texttt{knots} = (2, 4, 6))$ which places three knots at three points 2, 4 and 6. A simpler way is to specify the degrees of freedom. For example, $\texttt{bs}(x, \texttt{df} = 6)$ places the knots at the twenty-fifth, fiftieth and seventy-fifth percentile of x. Detailed description of these two functions can be found in the help file of Splus.

We now give a simple simulated example. The model we used to generate the pseudo-random numbers is

$$\lambda(t|x, w_1, w_1) = \lambda_0 \exp(\beta x + \phi_1(w_1) + \phi_2(w_2)),$$

where $\beta = 1$ and where $\phi_1(w_1) = 1.5(w_1 - 1.2)^2$ and $\phi_2(w_2) = 2\log(200 + (w_2 - 1.2)^3)$. The joint distribution of (X, W_1, W_2) is multivariate normal with mean $(0, 1.2, 1.2)$, standard deviation $(0.6, 0.6, 0.6)$ and all the pairwise correlations equal to 0.6. The baseline λ_0 is taken to be a constant equal to $\exp(-12)$ where 12 is approximately the expectation of $X + \phi_1(W_1) + \phi_2(W_2)$. The distribution of the censoring time given (x, w_1, w_2) is exponential with mean equal to $\exp(2) = 7.39$. So the censuring distribution does not depend

on the parameters of the distribution of T. The expected censoring rate is 20%. The following three commands complete the main part of the computation:

```
p1 ← bs(w1, df = 6); p2 ← bs(w2, df = 6)
sim.dat ← list(time = T, status = censor.ind, x, p1, p2)
sim.fit ← coxph(Surv(time, status) ~ x + p1 + p2, sim.dat,
iter.max = 20)
```

Here T is the vector of the simulated event times, `censor.ind` is the censoring indicator. `Surv` is a Splus function that generates the appropriate response variable for `coxph`.

To examine the performance of the estimator of β in this example, 1000 datasets are generated. In each dataset, the sample size $n = 140$. Table 1 summarizes the results. We also computed the estimator $\hat{\beta}_c$ using the true form of ϕ_1 and ϕ_2. This estimator serves as a bench mark for evaluating $\hat{\beta}_n$.

In the table, *mean* is the average of the 1000 estimated β's; *bias* is the difference between the *mean* and the generating value $\beta = 1$; *sd* is the sample standard deviation of the 1000 estimated β's, which represents the true variability of the estimators; and *mean se* is the average of 1000 standard error estimates of the estimated β from `coxph`. It is seen that the performance of $\hat{\beta}_n$ in terms of *bias* and *sd* is slightly worse than, but comparable to, that of $\hat{\beta}_c$. This is expected because $\hat{\beta}_c$ is estimated under the generating model. Observe that the standard error estimate based on the observed (partial) information works well for this example.

3. Main results. In this section, we state the results on the information bound, the asymptotic distribution of $\hat{\beta}_n$ and rate of convergence of $\hat{\phi}_{jn}$'s. We first state the conditions for the asymptotic results. These conditions combine the usual conditions in the asymptotic studies of nonparametric regression estimators and the Cox regression model with right-censored data.

Let k be a nonnegative integer, and let $\alpha \in (0, 1]$ be such that $p = k + \alpha > 0.5$. Let $\mathscr{A}$ be the class of functions h on $[0, 1]$ whose kth derivative $h^{(k)}$ exists and satisfies a Lipschitz condition of order α,

$$\left|h^{(k)}(s) - h^{(k)}(t)\right| \leq C|s - t|^{\alpha} \quad \text{for } s, t \in [0, 1].$$

TABLE 1
Summary of the example

	mean	**bias**	**sd**	**mean se**
$\hat{\beta}_n$	1.04	0.04	0.20	0.19
$\hat{\beta}_c$	1.01	0.01	0.16	0.15

(B1) (i) The regression parameter β_0 belongs to an open subset (not necessarily bounded) of R^d, and each $\phi_j \in \mathscr{A}$ for $1 \le j \le J$; (ii) $E(\Delta X) = 0$ and $E[\Delta \phi_j(W_J)] = 0$, $1 \le j \le J$.

The requirement that β_0 not be on the boundary of the parameter space is standard for asymptotic normality. The smoothness assumption of ϕ_j's is also often used in nonparametric curve estimation. Usually, $p = 1$ (i.e., $k = 0$ and $\alpha = 1$) or $p = 2$ (i.e., $k = 1$ and $\alpha = 1$) should be satisfied in many situations. These two cases roughly correspond to assuming that ϕ_j's have bounded first-order derivative or bounded second order derivative. (B1)(ii) requires the covariate X and the regression function to be suitably centered. Because the regression functions can only be identified up to a constant, centering removes this ambiguity. Observe that the partial likelihood does not change when each X_i is centered by the sample version of $E(\Delta X)$; therefore, this centering does not impose any real restriction. Centering by $E(\Delta X)$ or $E[\Delta \phi_j(W_j)]$ instead of the simpler $E(X)$ or $E[\phi_j(W_j)]$ simplifies information calculation and asymptotic analysis; see Sections 4 and 5.

(B2) The failure time T^u and the censoring time T^c are conditionally independent given the covariate Z.

(B3) (i) Only the observations for which the event time T_i, $1 \le i \le n$ is in a finite interval, say $[0, \tau]$, are used in the partial likelihood. At this point τ, the baseline cumulative hazard function $\Lambda_0(\tau) \equiv \int_0^\tau \lambda_0(s)\, ds < \infty$. (ii) The covariate X takes values in a bounded subset of R^d, and the covariate W takes values in $[a, b]^J$.

(B4) There exists a small positive constant ε such that (i) $P(\Delta = 1|Z) > \varepsilon$ and (ii) $P(T^c > \tau|Z) > \varepsilon$ almost surely with respect to the probability measure of Z.

Condition (B2) is sufficient for the censoring mechanism to be noninformative, which is often assumed in analyzing right-censored data. (B3)(i) is a major technical assumption, which avoids the unboundedness of the partial likelihood and the partial score functions at the end point of the support of the observed event time. Condition (B3)(ii) places the boundedness condition on the covariates, which is unpleasant, but it is not too restrictive in many situations because one is often able to put some bound on the covariates. A similar assumption is often used in asymptotic analysis of nonparametric regression problems.

Condition (B4)(i) ensures that the probability of being uncensored is positive regardless of the covariate value. Condition (B4)(ii) prevents censoring from being too heavy. Conditions (B3) and (B4) were also used by Sasieni [(1992b), Appendix] as sufficient conditions to ensure that the sumspace of the tangent spaces for the hazard and the regression functions be closed, so that the projections and information bound are well defined.

(B5) Let $0 < c_1 < c_2 < \infty$ be two constants. The joint density $f(t, w, \Delta = 1)$ of $(T, W, \Delta = 1)$ satisfies $c_1 \le f(t, w, \Delta = 1) < c_2$ for all $(t, w) \in [0, \tau] \times [0, 1]^J$.

This condition and the centering condition in (B1)(ii) are needed for the model to be identifiable. Note that weaker conditions can be formulated if only identifiability is required. However, (B5) is also used in information bound calculation and in obtaining the rate of convergence for the estimator of each nonparametric component in the model.

(B6) Let $q \geq 1$ be a positive integer. For $1 \leq j \leq J$, the qth partial derivative of the joint density $f(t, x, w, \Delta = 1)$ of $(T, X, W, \Delta = 1)$ with respect to t or w_j exists and is bounded. [For discrete covariate X, $f(t, x, w, \Delta = 1)$ is defined to be $(\partial^2/\partial t\, \partial w)P(T \leq t,\ X = x,\ W \leq w,\ \Delta = 1)$.]

This condition is used in showing that the partial score functions of the nonparametric components in the least favorable direction are nearly zero, which is a key step in proving the root-n convergence rate and asymptotic normality of the finite-dimensional estimator.

Let $r(z) = \exp(x'\beta + \phi(w))$, and let

$$M(t) \equiv M(t|Z) = \Delta 1_{[T \leq t]} - \int_0^t 1_{[T \geq u]} r(Z)\, d\Lambda_0(u) \tag{3.1}$$

be the usual counting process martingale associated with the Cox model. Throughout, Let $\|\cdot\|$ denote the Euclidean norm, and let $\|\cdot\|_2$ denote the L_2-norm with respect to a probability measure which should be clear in the context. Also, let $\|\cdot\|_\infty$ denote the supremum norm.

THEOREM 3.1. *Under conditions* (B1) *to* (B5), *the efficient score for estimation of* β *in the partly linear additive Cox model* (1.1) *is*

$$l_\beta^*(T, \Delta, Z) = \int_0^\tau (X - a^*(t) - h^*(W))\, dM(t),$$

where $h^*(w) = h_1^*(w_1) + \cdots + h_J^*(w_J)$ *and* $(a^*, h_1^*, \ldots, h_J^*)$ *are the unique* L_2 *functions that minimize*

$$E\Delta \| X - a(T) - h_1(W_1) - \cdots - h_J(W_J) \|^2.$$

Here a^* *can expressed as* $a^*(t) = E[X - h^*(W)|T = t,\ \Delta = 1]$. *The information bound for estimation of* β *is*

$$I(\beta) = E\left[l_\beta^*(T, \Delta, Z) \right]^{\otimes 2} = E\left[\Delta (X - a^*(T) - h^*(W))^{\otimes 2} \right],$$

where $x^{\otimes 2} \equiv xx'$ *for any column vector* $x \in R^d$.

THEOREM 3.2. *Suppose that conditions* (B1) *to* (B5) *hold and* $0 < v < 0.5$. *Then*

$$E\Delta \left[X'\hat{\beta}_n + \hat{\phi}_n(W) - (X'\beta + \phi(W)) \right]^2 = O_p(n^{-2vp} + n^{-(1-v)}).$$

Furthermore, if $I(\beta)$ *is nonsingular, then*

$$\| \hat{\beta}_n - \beta \|^2 = O_p(n^{-2vp} + n^{-(1-v)})$$

and

$$\|\hat{\phi}_{jn} - \phi_j\|_2^2 = O_p(n^{-2vp} + n^{-(1-v)}), \qquad 1 \le j \le J.$$

If $v = 1/(1 + 2p)$, the rate of convergence of $\hat{\phi}_{jn}$ is $n^{p/(1+2p)}$ which is the same as the optimal rate in nonparametric regression. The following theorem states that the rate of convergence of $\hat{\beta}_n$ achieves $n^{1/2}$ under condition (B6) in addition to conditions (B1)–(B5).

THEOREM 3.3. *Suppose that conditions* (B1)–(B6) *hold and that* $I(\beta)$ *is nonsingular. If* v *satisfies the restrictions* $0.25/p < v < 0.5$ *and* $v(q + p) > 0.5$, *where* p *is the measure of smoothness of* ϕ_j *defined in* (B2), *and* q *is defined in* (B6), *then*

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) = n^{-1/2} I^{-1}(\beta) \sum_{i=1}^{n} l_\beta^*(T_i, \Delta_i, Z_i) + o_p(1) \to_d N(0, \Sigma),$$

where $\Sigma = I^{-1}(\beta)$.

REMARK 3.1. It is interesting to notice that the $n^{1/2}$ rate of convergence and asymptotic normality of $\hat{\beta}_n$ hold for a range of the number of knots $K_n = O(n^v)$, although the rate of convergence of $\hat{g}_{jn}$ is slower than $n^{1/2}$. Here v plays the role of a smoothness parameter. The range of v that ensures asymptotic normality of $\hat{\beta}_n$ depends on p and q, where p measures the smoothness of the nonparametric parameters and q can be regarded as a measure of the smoothness of the model. If $p = 1$ and $q = 1$, then asymptotic normality of $\hat{\beta}_n$ holds for $1/4 < v < 1/2$. If $p = 2$ and $q = 2$, then asymptotic normality of $\hat{\beta}_n$ holds for $1/8 < v < 1/2$.

For estimating ϕ_j's, the optimal choice of v is $v = 1/(1 + 2p)$. This choice of v satisfies the restriction on v stated in Theorem 3.3. With this choice, both $\hat{\beta}_n$ and $\hat{g}_n$ achieve the optimal rates of convergence, $n^{1/2}$ and $n^{p/(1+2p)}$, respectively.

REMARK 3.2. Because $\hat{\beta}_n$ achieves this information lower bound and is asymptotically linear, it is asymptotically efficient among all the regular estimators. See for example, Van der Vaart (1991) and Bickel, Klaassen, Ritov and Wellner [(1993), Chapter 3 (in particular, Section 3.4)] for a systematic discussion on the information bounds for finite-dimensional parameters in infinite-dimensional models.

4. Information bound calculation. In this section, we calculate the information bound for the estimation of β given in Theorem 3.1. General theory on the asymptotic information bound for parameters in infinite-dimensional models can be found in Van der Vaart (1991) and Bickel, Klaassen, Ritov and Wellner (1993). The calculation here is based on the approach of Sasieni (1992b), who carried out information calculation in the partly linear Cox model (1.2) in which projection onto a sumspace of two nonorthogonal L_2

spaces was calculated. We extend this method to the partly additive model (1.1) in which projection onto the sumspace of $J + 1$ nonorthogonal L_2 spaces needs to be calculated.

We start with the log-likelihood function and the score functions associated with the parameters. The log-likelihood for a sample of size one is, up to an additive term not dependent on (β, ϕ, Λ),

$$l(\beta, \phi, \Lambda) = \Delta \log \lambda(T) + \Delta[X'\beta + \phi(W)] - \Lambda(T)\exp[X'\beta + \phi(W)],$$

where $\phi(W) = \phi_1(W_1) + \cdots + \phi_J(W_J)$. Consider a parametric smooth submodel $\{\lambda_{(\eta)}\colon\ \eta \in R\}$ and $\{\phi_{j(\eta_j)}\colon\ \eta_j \in R,\ 1 \le j \le J\}$ in which $\lambda_{(0)} = \lambda$ and $\phi_{j(0)} = \phi_j$ and

$$\frac{\partial \log \lambda_{(\eta)}}{\partial \eta}(t)|_{\eta=0} = a$$

and

$$\frac{\partial \phi_{j(\eta_j)}}{\partial \eta_j}(w_j)|_{\eta_j=0} = h_j(w_j), \qquad 1 \le j \le J.$$

Recall $r(z) = \exp(x'\beta + \phi(w))$ and M is the martingale defined in (3.1). The score operators for the hazard Λ and regression functions ϕ_j and the score vector for β are the partial derivatives of the likelihood $l(\beta, \phi_{1(\eta_1)}, \ldots, \phi_{J(\eta_J)}, \Lambda_{(\eta)})$ with respect to $\eta, \eta_1, \ldots, \eta_J$ and β evaluated at $\eta = 0$, $\eta_1 = 0, \ldots, \eta_J = 0$,

$$\dot{l}_\Lambda a \equiv \Delta a(T) - r(Z)\int_0^\infty Y(t)a(t)\,d\Lambda(t) = \int_0^\infty a(t)\,dM(t), \tag{4.1}$$

$$\dot{l}_{\phi_j} h_j \equiv h_j(W_j)[\Delta - r(Z)\Lambda(T)] = \int_0^\infty h_j(W_j)\,dM(t), \qquad 1 \le j \le J, \tag{4.2}$$

$$\dot{l}_\beta \equiv X[\Delta - r(Z)\Lambda(T)] = \int_0^\infty X\,dM(t). \tag{4.3}$$

Define $L_2(P_T^{(u)}) \equiv \{a\colon\ E[\Delta a^2(T)] < \infty\}$, and $L_2^0(P_{W_j}^{(u)}) \equiv \{h_j\colon\ E[\Delta h_j(W_j)] = 0;\ E[\Delta h_j^2(W_j)] < \infty\}$, $1 \le j \le J$. Let

$$A_\Lambda = \left\{\dot{l}_\Lambda a\colon a \in L_2\left(P_T^{(u)}\right)\right\}$$

and

$$H_j = \left\{\dot{l}_{\phi_j} h_j\colon h_j \in L_2^0\left(P_{W_j}^{(u)}\right)\right\}, \qquad 1 \le j \le J.$$

Let $h^* = (h_1^*, \ldots, h_J^*)$ and $\dot{l}_\phi h^* = \dot{l}_{\phi_1} h_1^* + \cdots + \dot{l}_{\phi_J} h_J^*$. To calculate the information bound for β, we need to find the (least favorable) direction $(a^*, h_1^*, \ldots, h_J^*)$ such that $\dot{l}_\beta - \dot{l}_\Lambda a^* - \dot{l}_\phi h^*$ is orthogonal to the sumspace $\mathbf{A} = A_\Lambda + H_1 + \cdots + H_J$. That is, $(a^*, h_1^*, \ldots, h_J^*)$ must satisfy

$$E\left\{\left[\dot{l}_\beta - \dot{l}_\Lambda a^* - \dot{l}_\phi h^*\right]\dot{l}_\Lambda a\right\} = 0, \qquad a \in L_2\left(P_T^{(u)}\right),$$

$$E\left\{\left[\dot{l}_\beta - \dot{l}_\Lambda a^* - \dot{l}_\phi h^*\right]\dot{l}_{\phi_j} h_j\right\} = 0, \qquad h_j \in L_2^0\left(P_{W_j}^{(u)}\right),\ 1 \le j \le J.$$

By the martingale representations given in (4.1), (4.2) and (4.3), these two equations can be written as

$$E\left[\int (X - a^* - h_1^* - \cdots - h_J^*)\, dM \int a\, dM\right] = 0, \tag{4.4}$$

$$E\left[\int (X - a^* - h_1^* - \cdots - h_J^*) dM \int h_j\, dM\right] = 0, \qquad 1 \le j \le J. \tag{4.5}$$

By Lemma 1 of Sasienin (1992b), for any measurable ψ_k: $R^+ \times R^{d+J} \to R^q$ satisfying $E[\psi_j^2(T, Z)] < \infty$, $k = 1, 2$,

$$E\left[\int \psi_1(t, Z)\, dM(t) \int \psi_2(t, Z)\, dM(t)\right] = E\left[\Delta \psi_1(T, Z) \psi_2(T, Z)\right], \tag{4.6}$$

provided that the compensator of M is absolutely continuous. So (4.4) and (4.5) are equivalent to

$$E\left[\Delta(X - a^* - h_1^* - \cdots - h_J^*) a\right] = 0, \qquad a \in L_2\left(P_T^{(u)}\right),$$

$$E\left[\Delta(X - a^* - h_1^* - \cdots - h_J^*) h_j\right] = 0, \qquad h_j \in L_2^0\left(P_{W_j}^{(u)}\right), 1 \le j \le J.$$

Therefore, we can take $(a^*, h_1^*, \ldots, h_J^*)$ to be the solution to the following equations:

$$E\left[X - a^* - h_1^* - \cdots - h_J^* | T = t, \Delta = 1\right] = 0 \quad \text{a.s. } P_T^{(u)}, \tag{4.7}$$

$$E\left[X - a^* - h_1^* - \cdots - h_J^* | W_j = w_j, \Delta = 1\right] = 0 \quad \text{a.s. } P_{W_j}^{(u)}, \tag{4.8}$$

$$1 \le j \le J.$$

It follows that $a^* + h_1^* + \cdots + h_J^*$ is the projection of X onto the sumspace $\mathbf{L} \equiv L_2(P_T^{(u)}) + L_2^0(P_{W_1}^{(u)}) + \cdots + L_2^0(P_{W_J}^{(u)})$.

We now show that, under condition (B3), the sumspace $\mathbf{L}$ is closed, so that the projection is well defined. According to Proposition 2, part A, of Bickel, Klaassen, Ritov and Wellner [(1993), Appendix 4, pages 440 and 441] it suffices to show that for $a \in L_2(P_T^{(u)})$ and $h_j \in L_2^0(P_{W_j}^{(u)})$, $1 \le j \le J$,

$$\begin{aligned} &E\left[\Delta \|a + h_1 + \cdots + h_J\|^2\right] \\ &\quad \ge c\left\{E\left[\Delta \|a\|^2\right] + E\left[\Delta \|h_1\|^2\right] + \cdots + E\left[\Delta \|h_J\|^2\right]\right\} \end{aligned} \tag{4.9}$$

for a constant $c > 0$. Under conditions (B4) and (B5), (4.9) follows from Lemma 1 of Stone (1985). Moreover, because of (4.9), $(a^*, h_1^*, \ldots, h_J^*)$ is unique, and the population version of the back-fitting algorithm, which is the inner loop of the ACE algorithm of Breiman and Friedman (1985), converges to $(a^*, h_1^*, \ldots, h_J^*)$.

The above calculation directly projects $\dot{l}_\beta$ onto the sumspace $\mathbf{A}$. By (4.6), the problem is transformed to the calculation of the projection of X onto the sumspace $\mathbf{L}$. Because $\mathbf{L}$ has a more transparent structure than $\mathbf{A}$, the calculation becomes much easier. Also, with $\mathbf{A}$, it is easier to formulate appropriate conditions so that the projection is well defined and unique.

A different approach is first to eliminate the hazard function by projecting the scores $\dot{l}_\beta$ and $\dot{l}_{\phi_j}$ onto the tangent space for the hazard and then projecting the residual of the projection of $\dot{l}_\beta$ onto the sumspace generated by the residual scores of the projection of $\dot{l}_{\phi_j}$. This route was used by Sasieni (1992b). An advantage of this approach is that it is more naturally related to the partial likelihood.

We now outline this approach for the present model. Let S be an operator taking measurable functions of Z to functions of t defined by $Sa(t) = E[a(Z)r(Z)1_{[T \geq t]}]$ and let $S_0(t) = E[r(Z)1_{[T \geq t]}]$. Denote $S_k(t) = SZ_1^k$, $k = 0, 1$. A useful identity due to Sasieni [(1992b), Lemma 2] is

$$\frac{Sa(t)}{S_0(t)} = E\big[a(Z)|T = t,\ \delta = 1\big]. \tag{4.10}$$

By Proposition 1(iii) of Sasieni (1992b), regression scores orthogonal to the tangent space for the hazard are

$$K_j h_j = \dot{l}_{\phi_j} h_j - \dot{l}_\Lambda\left(\frac{Sh_j}{S_0}\right) \equiv \int D_j h_j(z, t)\, dM(t|z),$$

where D_j: $L_2^0(P_{W_j}^{(u)}) \to L_2(P^{(u)})$ is defined by

$$D_j h_j(w_j, t) = h_j(w_j) - \frac{Sh_j}{S_0}(t) = h_j(w_j) - E\big[h_j|T = t, \Delta = 1\big],$$

$$1 \leq j \leq J.$$

In other words, K_j's are the residual scores of the projection of $\dot{l}_{\phi_j}$ onto the tangent space for the hazard. By Theorem 1(ii) of Sasieni (1992b), the residual scores of the projection of $\dot{l}_\beta$ onto the tangent space for the hazard is

$$K_\beta = i_\beta - i_\Lambda\left(\frac{S_1}{S_0}\right) = \int D_X(z, t)\, dM(t|z),$$

where

$$D_X(z, t) = z - \frac{S_1}{S_0}(t) = z - E\big[Z|T = t, \Delta = 1\big].$$

In the remainder of this section, let $h \equiv (h_1, \ldots, h_J)$ where $h_j \in L_2^0(P_{W_j}^{(u)})$, denote $Kh = K_1 h_1 - \cdots - K_J h_J$. The least favorable direction is $h^* \equiv (h_1^*, \ldots, h_J^*)$ with $h_j^* \in L_2^0(P_{W_j}^{(u)})$ that minimizes

$$E\|K_\beta - Kh\|^2. \tag{4.11}$$

Equivalently, h^* is the direction such that $K_\beta - Kh^*$ is orthogonal to Kh for all $h = (h_1, \ldots, h_J)$ with $h_j \in L_2^0(P_{W_j}^{(u)})$. Therefore, h^* must satisfy

$$E\big[(K_\beta - Kh^*)K_j h_j\big] = 0 \quad \text{for every } h_j \in L_2^0\big(P_{W_j}^{(u)}\big),\ 1 \leq j \leq J. \tag{4.12}$$

To see that such an h^* exists, denote $Dh^* = D_1 h_1^* + \cdots + D_J h_J^*$. By Lemma 1 of Sasieni (1992b), as in the proof of Proposition 2 of the same article, we have

$$E\big[(K_\beta - Kh^*)K_j h_j\big] = E\big[D_j'(D_X - Dh^*)h_j\big].$$

Therefore,

$$D_j'(D_X - Dh^*) = 0 \quad \text{a.s.} - P_{W_j}^{(u)}, 1 \le j \le J.$$

By Lemma 3 of Sasieni (1992),

$$D_j'D_X(w_j) = E\big[X - E(X|T = t, \Delta = 1)|W_j = w_j, \Delta = 1\big],$$
$$D_j'Dh^*(w_j) = E\big[h^*(W) - E(h^*(W)|T = t, \Delta = 1)|W_j = w_j, \Delta = 1\big].$$

Therefore, h^* satisfies

$$E\big[X - h^*(W) - E(X - h^*(W)|T = t, \Delta = 1)|W_j = w_j, \Delta = 1\big] = 0 \quad \text{a.s.} - P_{W_j}^{(u)}. \tag{4.13}$$

for $1 \le j \le J$. Let

$$a^*(t) = E\big[X - h^*(W)|T = t, \Delta = 1\big]. \tag{4.14}$$

It is seen that (4.13) and (4.14) are equivalent to (4.7) and (4.8).

5. Rate of convergence and asymptotic normality. In this section, we prove Theorems 3.2 and 3.3. In the proof of Theorem 3.2, we first obtain a suboptimal convergence rate by taking advantage of the concavity of the partial likelihood. This enables us to work in a sufficiently small neighborhood of the parameters. We then use Theorem 3.4.1 of Van der Vaart and Wellner [(1996), pages 322–323] to obtain the rates of convergence. The proof of Theorem 3.3 is based on Theorem 6.1 of Huang (1996), which provides a set of sufficient conditions for the maximum likelihood estimator of the finite-dimensional parameter in a class of semiparametric models to satisfy a central limit theorem. Although we are dealing with a partial likelihood, the approach there can be adapted to the present situation.

Throughout this section, denote the regression function by $g(z) = x'\beta + \phi(w)$ with $\phi(w) = \phi_1(w_1) + \cdots + \phi_J(w_J)$. To avoid confusion, let (β_0, ϕ_0) be the true parameter value. Denote $g_0(z) = x'\beta_0 + \phi_0(w)$. By Lemma A.5, there exists $\phi_n \in \Phi_n$ such that $\|\phi_n - \phi_0\|_\infty = O_p(n^{-vp} + n^{-(1-v)})$. Let $g_n(z) = x'\beta_0 + \phi_n(w)$. Also denote the estimator of g_0 by $\hat{g}_n(z) = x'\hat{\beta}_n + \hat{\phi}_n(w)$.

Let P_n be the empirical measure of (T_i, Δ_i, Z_i), $1 \le i \le n$ and let P be the probability measure of (T, Δ, Z). Let $P_{\Delta n}$ be the (subprobability) empirical measure of $(T_i, \Delta_i = 1, Z_i)$, $1 \le i \le n$ and let P_Δ be one subprobability measure of $(T, \Delta = 1, Z)$. It is convenient to use linear functional notation. So, for example, $P_{\Delta n}f = \int f dP_{\Delta n} = \int \Delta f dP_n = n^{-1}\sum_{i=1}^n \Delta_i f(T_i, \Delta_i, Z_i)$ for any f such that this integral is well defined.

5.1. *Rate of convergence.* Throughout this subsection, we assume that conditions (B1) to (B5) hold and $0 < v < 0.5$. For $0 \le t \le \tau$, let $Y(t) = 1_{[T \ge t]}$

and $Y_j(t) = 1_{[T_j \geq t]}$, $1 \leq j \leq n$. Denote

$$S_{0n}(t, g) = n^{-1} \sum_{j=1}^{n} Y_j(t) \exp\big(g(Z_j)\big), \tag{5.1}$$
$$S_0(t, g) = EY(t) \exp(g(Z))$$

and

$$S_{1n}(t, g)[h] = n^{-1} \sum_{j=1}^{n} Y_j(t) h(Z_j) \exp(g(Z_i)), \tag{5.2}$$
$$S_1(t, g)[h] = EY(t) h(Z) \exp(g(Z)).$$

Let τ be given in condition (B3)(i). The logarithm of the partial likelihood is

$$M_n(g) = n^{-1} \sum_{i=1}^{n} 1_{[0 \leq T_i \leq \tau]} \Delta_i \big[g(Z_i) - \log S_{0n}(T_i, g) \big].$$

Denote $m_n(t, x, g) = [g(z) - \log S_{0n}(t, g)] 1_{0[\leq t \leq \tau]}$ and $m_0(t, x, g) = [g(z) - \log S_0(t, g)] 1_{[0 \leq t \leq \tau]}$. Then

$$M_n(g) = P_{\Delta n} m_n(\cdot, g).$$

Let

$$M_0(g) = P_{\Delta} m_0(\cdot, g).$$

For notational convenience, in the remainder of the proofs including those in this section and in the Appendix, we will drop the indicator function $1_{[0 \leq t \leq \tau]}$ in the summation and integration or in the integrand of the subprobability measure P_{Δ} or the empirical measure $P_{\Delta n}$.

LEMMA 5.1. *Let $q_n = K_n + l$ be the number of polynomial spline basis functions defined in Section* 2:

$$\|\hat{g}_n - g_n\|_2^2 = o_p\big(q_n^{-1}\big).$$

Subsequently, by Lemma 7 *of Stone* (1986*a*), $\|\hat{g}_n - g_n\|_{\infty} = o_p(1)$.

PROOF. Let $b \in R^d$ and $\psi_n \in \Phi_n$ be such that $\|x'b + \psi_n(z)\|_2^2 = O(q_n^{-1})$. Denote $h_n(z) = x'b + \psi_n(z)$. Let $H_n(\alpha) = M_n(g_n + \alpha h_n)$. The derivative of H_n is

$$H_n'(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \Delta_i \left[h_n(Z_i) - \frac{n^{-1} \sum_{j=1}^{n} Y_j(t) h_n(Z_j) \exp\big[(g_n + \alpha h_n)(Z_i)\big]}{n^{-1} \sum_{j=1}^{n} Y_j(t) \exp\big[(g_n + \alpha h_n)(Z_i)\big]} \right]$$
$$= P_{\Delta n} \left[h_n - \frac{S_{1n}(\cdot, g_n + \alpha h_n)[h_n]}{S_{0n}(\cdot, g_n + \alpha h_n)} \right].$$

By concavity of $M_n(g)$, $H_n'(\alpha)$ is a nonincreasing function. Therefore, to prove the lemma, it suffices to show that for any $\alpha = \alpha_0 > 0$, $H_n'(\alpha_0) < 0$ and $H_n'(-\alpha_0) > 0$ except on an event with probability tending to zero, because then $\hat{g}_n$ must be between $g_n - \alpha_0 h_n$ and $g_n + \alpha_0 h_n$, and so $\|\hat{g}_n - g_n\|_2 \leq$

$\alpha_0\|h_n\|_2$. Let $b_n = g_n + \alpha_0 h_n$, and let

$$A_n(t) = \frac{S_1(\cdot, b_n)[h_n]}{S_0(\cdot, b_n)} - \frac{S_{1n}(\cdot, b_n)[h_n]}{S_{0n}(\cdot, b_n)}.$$

By adding and subtracting terms, we have

$$\begin{aligned} H_n'(\alpha_0) &= P_{\Delta n} A_n + (P_{\Delta n} - P_\Delta)\left[h_n - \frac{S_1(\cdot, b_n)[h_n]}{S_0(\cdot, b_n)}\right] \\ &\quad + P_\Delta\left[h_n - \frac{S_1(\cdot, b_n)[h_n]}{S_0(\cdot, b_n)}\right] \\ &\equiv I_{1n} + I_{2n} + I_{3n}. \end{aligned}$$

The first term $|I_{1n}| \le \sup_{0 \le t \le \tau}|A_n(t)|$. Write

$$\begin{aligned} &S_0(t, b_n) S_{0n}(t, b_n) A_n(t) \\ &\quad = S_1(t, b_n)[h_n]\{S_{0n}(t, b_n) - S_0(t, b_n)\} \\ &\qquad - S_0(t, b_n)\{S_{1n}(t, b_n)[h_n] - S_1(t, b_n)[h_n]\} \\ &\quad \equiv J_{1n}(t) + J_{2n}(t). \end{aligned}$$

By Lemma 7 of Stone (1986a), $\|h_n\|_\infty \le c q_n^{1/2}\|h_n\|_2 = O_p(1)$. By Lemma A.1, using Corollary A.2 on the bracket number for $\mathscr{M}_2$ and using $\sup_{0 \le t \le \tau}|S_1(t, b_n)[h_n]| \le \|h_n\|_2 = O(q_n^{-1/2})$, we have

$$\sup_{0 \le t \le \tau} |J_{1n}(t)| = O_p(1)\|h_n\|_2 n^{-1/2}\left(q_n^{1/2} + q_n^{-1/2}\log^{0.5} q_n\right) = O_p(n^{-1/2})$$

and

$$\sup_{0 \le t \le \tau} |J_{2n}(t)| = O_p(1) n^{-1/2} q_n^{-1/2}\left(q_n^{1/2} + \log^{0.5} q_n\right) = O_p(n^{-1/2}).$$

Thus $I_{1n} = O_p(n^{-1/2})$, since $\inf_{0 \le t \le \tau} S_0(t, b_n) S_{0n}(t, b_n) > 1/c_1$ for some constant $c_1 > 0$. Likewise, the second term $I_{2n} = O_p(n^{-1/2})$. For the third term, because

$$P_\Delta\left[h_n - \frac{S_1(\cdot, g_0)[h_n]}{S_0(\cdot, g_0)}\right] = 0,$$

we have, by adding and subtracting terms,

$$\begin{aligned} I_{3n} &= -P_\Delta\left[\frac{S_1(\cdot, b_n)[h_n]}{S_0(\cdot, b_n)} - \frac{S_1(\cdot, g_n)[h_n]}{S_0(\cdot, g_n)}\right] \\ &\quad - P_\Delta\left[\frac{S_1(\cdot, g_n)[h_n]}{S_0(\cdot, g_n)} - \frac{S_1(\cdot, g_0)[h_n]}{S_0(\cdot, g_0)}\right]. \end{aligned}$$

By Lemma A.4, using $P_{\Delta n} g_n = 0$ and $P_{\Delta n} h_n = 0$, as in the proof of Lemma A.6, we have

$$I_{3n} \le -c_2 \alpha_0 q_n^{-1} + O_p\left(n^{-1} q_n\right) = -c_2 \alpha_0 n^{-v} + O_p(n^{-(1-v)}).$$

Therefore, because $0 < v < 0.5$, we have,

$$H'(\alpha_0) \le -c_2\alpha_0 n^{-v} + O(n^{-1/2}) + O(n^{-(1-v)}) < 0,$$

except on an event with probability converging to zero. Similarly, we can show that $H'(-\alpha_0) > 0$ with high probability. This completes the proof of the lemma. □

PROOF OF THEOREM 3.2. We first prove that

$$\begin{aligned} &E \sup_{\eta/2 \le \|g-g_n\|_2 \le \eta} \left| M_n(g) - M_n(g_n) - (M_0(g) - M_0(g_n)) \right| \\ &\quad = n^{-1/2}\eta\Big(q_n + \sqrt{\log(1/\eta)}\Big). \end{aligned} \tag{5.3}$$

Observe that

$$\begin{aligned} &M_n(g) - M_n(g_n) - (M_0(g) - M_0(g_n)) \\ &\quad V = (P_{\Delta n} - P_\Delta)[m_0(\cdot, g) - m_0(\cdot, g_n)] \\ &\qquad - P_{\Delta n}\left[\log\frac{S_{0n}(\cdot, g)}{S_{0n}(\cdot, g_n)} - \log\frac{S_0(\cdot, g)}{S_0(\cdot, g_n)}\right] \\ &\quad \equiv I_{1n}(g) - I_{2n}(g). \end{aligned} \tag{5.4}$$

For the first term I_{1n}, by Van der Vaart and Wellner [(1996), Lemma 3.4.1],

$$E \sup_{\|g-g_n\|_2 \le \eta} |I_{1n}(g)| = n^{-1/2} m_n^{1/2}\eta.$$

For the second term I_{2n}, we have, for a constant $c_1 > 0$,

$$\begin{aligned} \sup_{\|g-g_n\|_2 \le \eta} |I_{2n}| &\le 2 \sup_{0 \le t \le \tau, \|g-g_n\|_2 \le \eta} \left| \log\frac{S_{0n}(\cdot, g)}{S_{0n}(\cdot, g_n)} - \log\frac{S_0(\cdot, g)}{S_0(\cdot, g_n)} \right| \\ &\le c_1 \sup_{0 \le t \le \tau, \|g-g_n\|_2 \le \eta} \left| \frac{S_{0n}(\cdot, g)}{S_{0n}(\cdot, g_n)} - \frac{S_0(\cdot, g)}{S_0(\cdot, g_n)} \right| \\ &\le c_1 n^{-1/2}\eta\big(q_n^{1/2} + \log^{0.5}(\eta^{-1})\big) \end{aligned}$$

with probability arbitrarily close to one for n sufficiently large, where the last inequality follows from Lemma A.3(i). Therefore, by Van der Vaart and Wellner [(1996), Theorem 3.4.1, pages 322 and 323], choosing the distance d_n defined in that theorem to be $d_n^2(\hat{g}_n, g_n) = -(P_\Delta m_0(\cdot, \hat{g}_n) - P_\Delta m_0(\cdot, g_n))$, we have

$$-r_{1n}^2[P_\Delta m_0(\cdot, \hat{g}_n) - P_\Delta m_0(\cdot, g_n)] = O_p(1),$$

where r_{1n} satisfies

$$r_{1n}^2\big(r_{1n}^{-1}q_n^{1/2} + r_{1n}^{-1}\log^{1/2} r_{1n}\big) = O(n^{1/2}).$$

It follows that $r_{1n} = q_n^{-1/2}n^{1/2} = n^{(1-v)/2}$. Therefore, by Lemma A.6, for a constant $c_2 > 0$,

$$c_2\|\hat{g}_n - g_n\|_2^2 \le O_p(n^{-(1-v)} + n^{-2vp}).$$

Because $\|g_n - g_0\|_\infty^2 = O_p(n^{-2vp} + n^{-(1-v)})$, we have

$$\|\hat{g}_n - g_0\|_2^2 = O_p(n^{-(1-v)} + n^{-2vp}).$$

By conditions (B4) and (B5), it follows that

$$E\Delta\left\| X'\hat{\beta}_n + \hat{\phi}_n(W) - \left(X'\beta_0 + \phi_0(W)\right)\right\|_2^2 = O_p(n^{-(1-v)} + n^{-2vp}).$$

Therefore, for the projections a^* and $h^*(w) = h_1^*(w_1) + \cdots + h_J^*(w_J)$ defined in Section 4,

$$\begin{aligned}
&E\Delta\Big\|\left(X - a^*(T) - h^*(W)\right)'\left(\hat{\beta}_n - \beta_0\right) \\
&\quad + \left(a^*(T) + h^*(W)\right)'\left(\hat{\beta}_n - \beta_0\right) + \left(\hat{\phi}_n(W) - \phi_0(W)\right)\Big\|_2^2 \\
&\quad = E\Delta\Big\|\left(X - a^*(T) - h^*(W)\right)'\left(\hat{\beta}_n - \beta_0\right)\Big\|_2^2 \\
&\qquad + E\Delta\Big\|\left(a^*(T) + h^*(W)\right)'\left(\hat{\beta}_n - \beta_0\right) + \left(\hat{\phi}_n(W) - \phi_0(W)\right)\Big\|_2^2 \\
&\quad = O_p(n^{-(1-v)} + n^{-2vp}),
\end{aligned}$$

where the first equality follows from orthogonality given in (4.7) and (4.8). Because $E[\Delta(X - a^*(T) - h^*(W))]^{\otimes 2}$ is assumed to be nonsingular, it follows that $\|\hat{\beta}_n - \beta_0\|_2^2 = O_p(n^{-(1-v)} + n^{-2vp})$. This in turn implies

$$E\Delta\left\|\hat{\phi}_n(W) - \phi_0(W)\right\|_2^2 = O_p(n^{-(1-v)} + n^{-2vp}).$$

Thus by Lemma 1 of Stone (1985), (B4) and (B5),

$$E\left\|\hat{\phi}_{jn}(W) - \phi_j(W)\right\|_2^2 = O_p(n^{-(1-v)} + n^{-2vp}), \qquad 1 \le j \le J.$$

The result follows. □

5.2. *Asymptotic normality and efficiency.* Throughout this section, we assume that conditions (B1)–(B6) hold. The proof of Theorem 3.3 is built on the following three lemmas.

Let $u = (t, x, w)$. For a real-valued function h of $w \in R^J$, define

$$s_n(u, g)[h] = h(w) - \frac{S_{1n}(t, g)[h]}{S_{0n}(t, g)}$$

and

$$s(u, g)[h] = h(w) - \frac{S_1(t, g)[h]}{S_0(t, g)},$$

where S_{kn} and S_k, $k = 0, 1$ are defined in (5.1) and (5.2), but now we take h to be a function of $w \in R^J$. To simplify (and slightly abuse) the notation, for a vector $x \in R^d$ and the identity map $I(x) = s$, denote

$$s_n(u, g)[x] = x - \frac{S_{1n}(t, g)[I]}{S_{0n}(t, g)} \quad \text{and} \quad s(u, g)[x] = x - \frac{S_1(t, g)[I]}{S_0(t, g)}.$$

We also write $s_n(u, \beta, \phi) = s_n(u, g)$ and so on.

As in likelihood estimation, we shall call the derivatives of the partial likelihood with respect to the parameters (partial) score functions. The score function based on the partial likelihood for β is

$$\dot{l}_{n\beta}(\beta, \phi) = P_{\Delta n} s_n(\cdot, \beta, \phi)[x].$$

The score function based on the partial likelihood for ϕ in a direction $h_n \in \Phi_n$ is

$$\dot{l}_{n\phi}(\beta, \phi)[h_n] = P_{\Delta n} s_n(\cdot, \beta, \phi)[h_n].$$

By the definition of $(\hat{\beta}_n, \hat{\phi}_n)$ (i.e., it maximizes the partial likelihood),

$$\dot{l}_{n\beta}\big(\hat{\beta}_n, \hat{\phi}_n\big) \equiv P_{\Delta n} s_n\big(\cdot, \hat{\beta}_n, \hat{\phi}_n\big)[x] = 0, \tag{5.5}$$

and for any $h_n \in \Phi_n$,

$$\dot{l}_{n\phi}\big(\hat{\beta}_n, \hat{\phi}_n\big)[h_n] \equiv P_{\Delta n} s_n\big(\cdot, \hat{\beta}_n, \hat{\phi}_n\big)[h_n] = 0. \tag{5.6}$$

The first key step (Lemma 5.2) in proving Theorem 3.3 is to show that the partial score function $\dot{l}_{n\phi}$ evaluated at $(\hat{\beta}_n, \hat{\phi}_n)$ along the least favorable direction is nearly zero.

LEMMA 5.2. *Let h^* be defined by $h^*(w) = h_1^*(w_1) + \cdots + h_J^*(w_J)$ (note that this is different from the notation we used in Section 4),*

$$\dot{l}_{n\phi}\big(\hat{\beta}_n, \hat{\phi}_n\big)[h^*] \equiv P_{\Delta n} s_n\big(\cdot, \hat{\beta}_n, \hat{\phi}_n\big)[h^*] = o_p(n^{-1/2}). \tag{5.7}$$

PROOF. By condition (B6), and equations (4.7) and (4.8), it can be shown that h^* is qth differentiable and its qth derivative is bounded. Thus according to Corollary 6.21 of Schumaker [(1981), page 227] there exists an $h_n^* \in \Phi_n$ such that

$$\|h_n^* - h^*\|_\infty = O(q_n^{-q}).$$

By (5.6),

$$\begin{aligned}
\dot{l}_{n\phi}\big(\hat{\beta}_n, \hat{\phi}_n\big)[h^*] &= \dot{l}_{n\phi}\big(\hat{\beta}_n, \hat{\phi}_n\big)[h^*] - \dot{l}_{n\phi}\big(\hat{\beta}_n, \hat{\phi}_n\big)[h_n^*] \\
&= P_{\Delta n}\left[h^* - h_n^* - \left(\frac{S_{1n}(\cdot, \hat{g}_n)[h^*]}{S_{0n}(\cdot, \hat{g}_n)} - \frac{S_{1n}(\cdot, \hat{g}_n)[h_n^*]}{S_{0n}(\cdot, \hat{g}_n)}\right)\right] \\
&= P_{\Delta n}\left[h^* - h_n^* - \frac{S_{1n}(\cdot, \hat{g}_n)[h^* - h_n^*]}{S_{0n}(\cdot, \hat{g}_n)}\right] \\
&\equiv I_{1n} + I_{2n} + I_{3n},
\end{aligned}$$

where

$$I_{1n} = (P_{\Delta n} - P_\Delta)\left[h^* - h_n^* - \frac{S_1(\cdot, \hat{g}_n)[h^* - h_n^*]}{S_0(\cdot, \hat{g}_n)}\right],$$

$$I_{2n} = P_{\Delta n}\left[\frac{S_1(\cdot, \hat{g}_n)[h^* - h_n^*]}{S_0(\cdot, \hat{g}_n)} - \frac{S_{1n}(\cdot, \hat{g}_n)[h^* - h_n^*]}{S_0(\cdot, \hat{g}_n)}\right]$$

and

$$I_{3n} = P_\Delta\left[h^* - h_n^* - \frac{S_1(\cdot, \hat{g}_n)[h^* - h_n^*]}{S_0(\cdot, \hat{g}_n)}\right].$$

By the maximal inequality in Lemma A.1 and some entropy calculation similar to those in Corollary A.1, it follows that $I_{1n} = o_p(n^{-1/2})$. By Lemma A.3(ii), $I_{2n} = o_p(n^{-1/2})$. Now consider the third term I_{3n}. By (4.10),

$$P_\Delta\left[h^* - h_n^* - \frac{S_1(\cdot, g_0)[h^* - h_n^*]}{S_0(\cdot, g_0)}\right]$$
$$= E\big[\Delta(h^* - h_n^*) - E(\Delta(h^* - h_n^*)|T = t, \Delta = 1)\big] = 0,$$

so we have

$$I_{3n} = P_\Delta\left[\frac{S_1(\cdot, g_0)[h^* - h_n^*]}{S_0(\cdot, g_0)} - \frac{S_1(\cdot, \hat{g}_n)[h^* - h_n^*]}{S_0(\cdot, \hat{g}_n)}\right].$$

By Lemma A.4, there exists a constant $c > 0$ such that

$$|I_{3n}| \le c\|h^* - h_n^*\|_\infty \|\hat{g}_n - g_0\|_2.$$

Therefore, $I_{3n} = n^{-qv}O_p(n^{-vp} + n^{-(1-v)/2}) = o_p(n^{-1/2})$ by the restriction on v stated in Theorem 3.3. This proves the lemma. □

LEMMA 5.3.

$$P_{\Delta n}\{s_n(\cdot, \hat{g}_n)[x] - s_n(\cdot, g_0)[x]\}$$
$$- P_\Delta\{s(\cdot, \hat{g}_n)[x] - s(\cdot, g_0)[x]\} = o_p(n^{-1/2}) \tag{5.8}$$

and

$$P_{\Delta n}\{s_n(\cdot, \hat{g}_n)[h^*] - s_n(\cdot, g_0)[h^*]\}$$
$$- P_\Delta\{s(\cdot, \hat{g}_n)[h^*] - s(\cdot, g_0)[h^*]\} = o_p(n^{-1/2}). \tag{5.9}$$

PROOF. We only prove (5.9), because the proof of (5.8) is similar. The right side of (5.9) is bounded by the sum of two terms,

$$I_{1n} \equiv \big|(P_{\Delta n} - P_\Delta)\{s(\cdot, \hat{g}_n)[h^*] - s(\cdot, g_0)[h^*]\}\big|$$

and

$$I_{2n} \equiv \big|P_{\Delta n}\{s_n(\cdot, \hat{g}_n)[h^*] - s_n(\cdot, g_0)[h^*] - (s(\cdot, \hat{g}_n)[h^*] - s(\cdot, g_0)[h^*])\}\big|.$$

For I_{1n}, by Lemma A.4 in the Appendix, $P[s(\cdot, \hat{g}_n)[h^*] - s(\cdot, g_0)[h^*]]^2 \le O(\|\hat{g}_n - g_0\|_2^2)$, and the ε-bracketing number of the class of functions $S_1(\eta) = \{s(\cdot, g)[h^*] - s(\cdot, g_0)[h^*]: \|g - g_0\|_2 \le \eta \in G\}$ is $q_n \log(\eta/\varepsilon) + \log(1/\varepsilon)$. The corresponding entropy integral $J_\square(\eta, S_1(\eta), L_2(P))$ is $\eta(q_n^{1/2} + \log^{0.5}(1/\eta))$. Therefore, by Lemma A.1 and Theorem 3.2, for $r_n \equiv n^{(1-v)/2} + n^{vp}$,

$$EI_{1n} \le O(1)n^{-1/2}r_n^{-1}\big[q_n + \log^{0.5}(r_n)\big] = o(n^{-1/2}).$$

For I_{2n}, the integrand of the empirical measure is

$$II_{2n}(t) \equiv \frac{S_{1n}(t,\hat{g}_n)[h^*]}{S_{0n}(t,\hat{g}_n)} - \frac{S_{1n}(t,g_0)[h^*]}{S_{0n}(t,g_0)} - \left[\frac{S_1(t,\hat{g}_n)[h^*]}{S_0(t,\hat{g}_n)} - \frac{S_1(t,g_0)[h^*]}{S_0(t,g_0)}\right].$$

It is shown in Lemma A.7 in the Appendix that

$$\sup_{0\le t\le t} |II_{2n}(t)| = o_p(n^{-1/2}).$$

This completes the proof. □

LEMMA 5.4.

$$\begin{aligned} &P_\Delta\{s(\cdot,\hat{g}_n)[x-h^*] - s(\cdot,g_0)[x-h^*]\} \\ &\quad = -P_\Delta\{s(\cdot,\hat{g}_n)[x-h^*]\}^{\otimes 2}(\hat{\beta}_n-\beta_0) + O\big(\|\hat{\beta}_n-\beta_0\|^2 + \|\hat{\phi}_n-\phi_0\|_2^2\big) \\ &\quad = -P_\Delta\{s(\cdot,\hat{g}_n)[x-h^*]\}^{\otimes 2}(\hat{\beta}_n-\beta_0) + o_p(n^{-1/2}). \end{aligned}$$

PROOF. By Lemma A.4 in the Appendix, we have

$$\begin{aligned} &P_\Delta\{s(\cdot,\hat{g}_n)[x-h^*] - s(\cdot,g_0)[x-h^*]\} \\ &\quad = -P_\Delta s(\cdot,g_0)[x-h^*]s(\cdot,g_0)[x](\hat{\beta}_n-\beta_0) \\ &\qquad - P_\Delta s(\cdot,g_0)[x-h^*]s(\cdot,g_0)\big[\hat{\phi}_n-\phi_0\big] \\ &\qquad + O\big(\|\hat{\beta}_n-\beta_0\|^2 + \|\hat{\phi}_n-\phi_0\|_2^2\big). \end{aligned}$$

However, by (4.12) in Section 4,

$$P_\Delta s(\cdot,g_0)[x-h^*]s(\cdot,g_0)\big[\hat{\phi}_n-\phi_0\big] = 0$$

and

$$P_\Delta s(\cdot,g_0)[x-h^*]s(\cdot,g_0)[x] = P_\Delta\{s(\cdot,g_0)[x-h^*]\}^{\otimes 2}.$$

Because $\|\hat{\beta}_n-\beta_0\|^2 = o_p(n^{-1/2})$ and $\|\hat{\phi}_n-\phi_0\|_2^2 = o_p(n^{-1/2})$ by Theorem 3.2, the lemma follows. □

PROOF OF THEOREM 3.3. By Lemmas 5.2, 5.3 and 5.4 and using the same proof of Huang [(1996), Theorem 6.1] we have

$$n^{1/2}P_\Delta\{s(\cdot,g_0)[x-h^*]\}^{\otimes 2}(\hat{\beta}_n-\beta_0) = n^{1/2}P_{\Delta n}s_n(\cdot,g_0)[x-h^*] + o_p(1).$$

So asymptotic normality of $\hat{\beta}_n$ follows directly by the martingale central limit theorem. However, the following argument shows that $\hat{\beta}_n$ is asymptotically linear in the efficient influence function. Let

$$M_i(t) = \Delta_i 1_{[T_i\le t]} - \int_0^t Y_i(u)\exp(g_0(Z_i))\,d\Lambda_0(u), \qquad 1\le i\le n.$$

We can write

$$n^{1/2}P_{\Delta n}s_n(\cdot, g_0)[x - h^*]$$
$$= n^{-1/2}\sum_{i=1}^{n}\int_0^\tau \left[X_i - h^*(W_i) - \frac{S_{1n}(t, g_0)[x - h^*]}{S_{0n}(t, g_0)}\right] dM_i(t).$$

Thus

$$n^{1/2}P_{\Delta n}s_n(\cdot, g_0)[x - h^*]$$
$$- n^{-1/2}\sum_{i=1}^{n}\int_0^\tau \left[X_i - h^*(W_i) - \frac{S_1(t, g_0)[x - h^*]}{S_0(t, g_0)}\right] dM_i(t)$$
$$= n^{-1/2}\sum_{i=1}^{n}\int_0^\tau \left[\frac{S_1(t, g_0)[x - h^*]}{S_0(t, g_0)} - \frac{S_{1n}(t, g_0)[x - h^*]}{S_{0n}(t, g_0)}\right] dM_i(t).$$

Because

$$n^{-1}\sum_{i=1}^{n}\int_0^\tau \left[\frac{S_1(t, g_0)[x - h^*]}{S_0(t, g_0)} - \frac{S_{1n}(t, g_0)[x - h^*]}{S_{0n}(t, g_0)}\right]^2$$
$$\times Y_i(u)\exp(g_0(Z_i))d\Lambda_0(u) \to_p 0,$$

by Lenglart's inequality as stated in Theorem 3.4.1 and Corollary 3.4.1 of Fleming and Harrington (1991) or Andersen, Borgan, Gill and Keiding [(1993), page 86] we have

$$n^{1/2}P_{\Delta n}s_n(\cdot, g_0)[x - h^*]$$
$$= n^{-1/2}\sum_{i=1}^{n}\int_0^\tau \left[X_i - h^*(W_i) - \frac{S_1(t, g_0)[x - h^*]}{S_0(t, g_0)}\right] dM_i(t) + o_p(1).$$

However, by (4.10),

$$\frac{S_1(t, g_0)[x - h^*]}{S_0(t, g_0)} = E[X - h^*(W)|T = t, \Delta = 1] = a^*(t).$$

By the definition of the efficient score function l_β^*, we have

$$\sqrt{n}\,P_{\Delta n}s(\cdot, g_0)[x - h^*] = n^{-1/2}\sum_{i=1}^{n} l_\beta^*(T_i, \Delta_i, Z_i) + o_p(1) \to_d N(0, I(\beta_0)).$$

Therefore, the result follows. This completes the proof. □

6. Concluding remarks. In this paper, we studied asymptotic properties of the maximum partial likelihood estimator of the partly linear additive Cox model using polynomial splines for the nonparametric regression components. We have elected to consider only the model with time-independent covariates in the random right-censorship setting, because this type of data arises often in practice and because the technical details involved already appear to not be straightforward. It seems that the results can be extended to

the multiplicative intensity model with time-dependent covariates (and a partly linear regression function) as described by Andersen and Gill (1982). If methods similar to the present one are to be used, two aspects not discussed in this paper need to be addressed. First, the information bound calculation (or the similar type of projection calculation) must be done for partial score operators and partial score functions involving time-dependent covariates. This calculation will be helpful in separating the root-n consistent estimator from estimators with lower rates of convergence. Second, there should be a maximal inequality similar to Lemma A.1 for martingale integrals indexed by classes of functions with appropriate bracketing entropy numbers. This type of inequality is useful in establishing rates of convergence and controlling remainder terms in the asymptotic normality proof.

APPENDIX

Technical lemmas. In this Appendix we collect several lemmas that are used in the previous sections.

For any probability measure Q, define $L_2(Q) = \{f: \int f^2\, dQ < \infty\}$. Let $\|\cdot\|_2$ be the usual L_2-norm, that is, $\|f\|_2 = (\int f^2\, dQ)^{1/2}$. For any subclass $\mathscr{F}$ of $L_2(Q)$, define the bracketing number $\mathscr{N}_{[]}(\varepsilon, \mathscr{F}, L_2(Q)) = \min\{m$: there exist $f_1^L, f_1^U, \ldots, f_m^L, f_m^U$ such that for each $f \in \mathscr{F}$, $f_i^L \le f \le f_i^U$ for some i, and $\|f_i^U - f_i^L\|_2 \le \varepsilon\}$. Denote $J_{[]}(\eta, \mathscr{F}, L_2(Q)) = \int_0^\eta \sqrt{1 + \log N_{[]}(\varepsilon, \mathscr{F}, L_2(Q))}\, d\varepsilon$.

The following lemma used in the previous sections is Lemma 3.4.2 of van der Vaart and Wellner (1996). Let $X_1, \ldots, X_n$ be i.i.d. random variables with distribution Q, and Q_n be the empirical measure of these random variables. Denote $G_n = \sqrt{n}(Q_n - Q)$, and $\|G_n\|_{\mathscr{F}} = \sup_{f \in \mathscr{F}} |G_n f|$ for any measurable class of functions $\mathscr{F}$.

LEMMA A.1. *Let M_0 be a finite positive constant. Let $\mathscr{F}$ be a uniformly bounded class of measurable functions such that $Qf^2 < \eta^2$ and $\|f\|_\infty \le M_0$. Then*

$$E_Q^* \|G_n\|_{\mathscr{F}} \le C_0 J_{[]}(\eta, \mathscr{F}, L_2(Q)) \left(1 + \frac{J_{[]}(\eta, \mathscr{F}, L_2(Q))}{\eta^2 \sqrt{n}} M_0\right),$$

where C_0 is a finite constant not dependent on n.

LEMMA A.2. *For any $\eta > 0$, let*

$$\Theta_n = \left\{x'\beta + \phi(w): \|\beta - \beta_0\| \le \eta,\ \phi_j \in \mathscr{S}_n, \|\phi_j - \phi_{0j}\|_2 \le \eta,\ 1 \le j \le J\right\}.$$

Then, for any $\varepsilon \le \eta$,

$$\log N_{[]}\big((\varepsilon, \Theta_n, L_2(P))\big) \le c\big(q_n \log(\eta/\varepsilon)\big).$$

(*Recall that $q_n = K_n + l + 1$ is the number of spline basis functions.*)

PROOF. By the calculation of Shen and Wong [(1994), page 597] $\log N_{[]}(\varepsilon, \mathscr{S}_n, L_2(P)) \le c_1(q_n \log(\eta/\varepsilon))$. Therefore, the logarithm of the bracketing number of the class

$$\Phi_n = \{\phi(z): \phi(z) = \phi_1(z_1) + \cdots + \phi_J(z_J): \phi_j \in \mathscr{S}_n, 1 \le j \le J\}$$

is also $c_2(q_n \log(\eta/\varepsilon))$. Since the neighborhood $B(\eta) \equiv \{\beta: \|\beta - \beta_0\| \le \eta\}$ in R^d can be covered by $c_3(\eta/\varepsilon)^d$ balls with radius ε, the logarithm of the bracketing number of Θ_n is bounded by $c_2 q_n \log(\eta/\varepsilon) + c_3 \log(\eta/\varepsilon)] \le cq_n \log(\eta/\varepsilon)$ for $c = \max\{c_2, c_3\}$. □

As a consequence of Lemma A.2, we have:

COROLLARY A.1. *Let* $m_0(t, \eta, x, z; \beta, \phi) = x'\beta + \phi(z) - \log S_0(t; \beta, \phi)$, $m_1(t, x, z; s, \beta, \phi) = 1_{[\tau \ge t \ge s]} \exp(x'\beta + \phi(z))$, *and* $m_2(t, x, z; s, \beta, b, \phi) = 1_{[\tau \ge t \ge s]} h(z) \exp(x'\beta + \phi(z))$. *Define the classes of functions*

$$\mathscr{M}_0(\eta) = \{m_0: \|\beta - \beta_0\| \le \eta, \|\phi - \phi_0\|_2 \le \eta\},$$

$$\mathscr{M}_1(\eta) = \{m_1: 0 \le s \le \tau, \|\beta - \beta_0\| \le \eta, \|\phi - \phi_0\|_2 \le \eta\}$$

and

$$\mathscr{M}_2(\eta) = \{m_2: 0 \le s \le \tau, \|\beta - \beta_0\| \le \eta, \|h\|_2 \le \eta, \|\phi - \phi\|_2 \le \eta\}.$$

Then for any $\varepsilon < \eta$,

$$\log N_{[]}(\varepsilon, \mathscr{M}_0(\eta), L_2(P)) \le c_0 q_n \log(\eta/\varepsilon)$$

and

$$\log N_{[]}(\varepsilon, \mathscr{M}_j(\eta), L_2(P)) \le c_j[q_n \log(\eta/\varepsilon) + \log(\tau/\varepsilon)], \qquad j = 1, 2,$$

Consequently,

$$J_{[]}(\eta, \mathscr{M}_0, L_2(P)) \le c_0 q_n^{1/2} \eta$$

and

$$J_{[]}(\eta, \mathscr{M}_j, L_2(P)) \le c_j[q_n^{1/2}\eta + \eta \log^{1/2}(1/\eta)], \qquad j = 1, 2.$$

PROOF. Because exp is monotone, by Lemma A.2, the entropy of the class consisting of functions $\exp(x'\beta + \phi(z))$ for $x'\beta + \phi(z) \in \Theta_n$ is bounded by $cq_n \log(\eta/\varepsilon)$. The ε-bracketing entropy of the indicator functions $1_{[\tau \ge t \ge s]}$, $s \in [0, \tau]$, is bounded by $\log(\tau/\varepsilon)$. The class $\mathscr{M}_1(\eta)$ is obtained by multiplying $\exp(x'\beta + \phi(z))$ by $1_{[\tau \ge t \ge s]}$; therefore, its bracketing entropy is bounded by the sum of $cq_n \log(\eta/\varepsilon) + c \log(\tau/\varepsilon)$. □

LEMMA A.3. (i) *Let* c *be a finite constant and* η *be a small positive constant. Define the class of functions*

$$\mathscr{G} = \{g: g(z) = x'\beta + \phi(w), \phi(w) \in \Phi_n, \|g - g_n\| \le \eta, \|g\|_\infty \le c\}.$$

Then

$$\sup_{t \in [0,1],\, g \in \mathscr{G}} \left| \frac{S_{0n}(t; g)}{S_{0n}(t; g_0)} - \frac{S_0(t; g)}{S_0(t; g_0)} \right| = \eta n^{-1/2} O_p(q_n^{1/2} + \log^{0.5}(\eta^{-1})). \tag{A.1}$$

(ii) *Suppose* $h_n \in \Phi_n$ *is a sequence of uniformly bounded functions and* $\|h_n\|_2 = O(q_n^{-1})$. *Then*

$$\sup_{t\in[0,1]}\left|\frac{S_{1n}(t;g_n)[h_n]}{S_{0n}(t;g_n)} - \frac{S_1(t;g_n)[h_n]}{S_0(t;g_n)}\right| = o_p(n^{-1/2}). \tag{A.2}$$

PROOF. (i) Because

$$\frac{S_{0n}(t;g)}{S_{0n}(t;g_n)} - \frac{S_0(t;g)}{S_0(t;g_n)} = \frac{S_{0n}(t;g)S_0(t;g_n) - S_{0n}(t;g_n)S_0(t;g)}{S_{0n}(t;g_n)S_0(t;g_n)}$$

and because the denominator on the right side is bounded away from zero with probability tending to one, we need only to consider the numerator. Write

$$\begin{aligned} &S_{0n}(t;g)S_0(t;g_n) - S_{0n}(t;g_n)S_0(t;g) \\ &\quad = S_0(t;g_n)\big[S_{0n}(t;g) - S_{0n}(t;g_n) - S_0(t;g) + S_0(t;g_n)\big] \\ &\qquad - \big[S_{0n}(t;g_n) - S_0(t;g_n)\big]\big[S_0(t;g) - S_0(t;g_n)\big]. \end{aligned}$$

The first term on the right side is

$$(P_n - P)\big\{y(t)\big[\exp(g(z)) - \exp(g_n(z))\big]\big\} = n^{-1/2}\eta O_p\big(q_n^{1/2} + \log^{0.5}(\eta^{-1})\big).$$

Because $S_{0n}(t;g_n) - S_0(t;g_n) = O_p(n^{-1/2}q_n^{1/2})$, and

$$\big|S_0(t;g) - S_0(t;g_n)\big| \le E\{Y(t)|\exp(g) - \exp(g_n)\} \le C\Big[E(g_n - g)^2\Big]^{1/2},$$

we have $[S_{0n}(t;g_n) - S_0(t;g_n)][S_0(t;g) - S_0(t;g_n)] = O_p(n^{-1/2}q_n^{1/2}\eta)$. Therefore,

$$S_{0n}(t;g_n)S_0(t;g_0) - S_{0n}(t;g_0)S_0(t;g_n) = n^{-1/2}\eta O_p\big(q_n^{1/2} + \log^{0.5}(\eta^{-1})\big).$$

(ii) Write

$$\begin{aligned} &\frac{S_{1n}(t;g_n)[h_n]}{S_{0n}(t;g_n)} - \frac{S_1(t;g_n)[h_n]}{S_0(t;g_n)} \\ &\quad = \frac{\begin{array}{r} S_0(t;g_n)\big[S_{1n}(t;g_n)[h_n] - S_1(t;g_n)[h_n]\big] - S_1(t;g_n)[h_n] \\ \times\big[S_{0n}(t;g_n) - S_0(t;g_n)\big] \end{array}}{S_0(t;g_n)S_{0n}(t;g_n)}. \end{aligned}$$

Because $g_n \to_p g_0$, the denominator of the right side of this equation is bounded away from zero. The first term in the numerator is equal to

$$(P_{\Delta n} - P_\Delta)\big\{y(t)h_n(z)\exp(g_n(z))\big\} = o_p(n^{-1/2}).$$

The second term in the numerator is equal to

$$EY(t)\big[h_n(z)\exp(g_n(z))\big](P_{\Delta n} - P_\Delta)\big[y(t)\exp(g_n(z))\big] = o_p(n^{-1/2}).$$

This completes the proof. □

LEMMA A.4. *For a number* $0 \le s \le 1$, *let*

$$H(t,s) = \frac{S_1(t; g_0 + sd)[h]}{S_0(t; g_0 + sd)}.$$

Denote $W_s(t) = Y(t)\exp(g_0 + sd)/[S_0(g_0 + sd)]$. *We have*

$$\begin{aligned}\frac{\partial}{\partial s}H(t;s) &= E[W_s(t)h(Z)d(Z)] - E[W_s(t)h(Z)]E[W_s(t)d(Z)] \\ &= E\{W_s(t)[h(Z) - E(W_s(t)h(Z))][d(Z) - E(W_s(t)d(Z))]\}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2}{\partial s^2}H(t;s) &= E[W_s(t)h(Z)d^2(Z)] - 2E[W_s(t)d(Z)]E[W_s(t)h(Z)d(Z)] \\ &\quad - E[W_s(t)h(Z)]E[W_s(t)d^2(Z)] \\ &\quad + 2E[W_s(t)h(Z)]E[W_s(t)d(Z)]^2.\end{aligned}$$

The lemma follows by direct calculation of the derivatives. Details are omitted.

LEMMA A.5. *Let* $1 \le j \le J$ *be the integer associated with the* j*th covariate* W_j. *Suppose that* $\xi \in \mathscr{G}$ *and* $E[\Delta\xi(W_j)] = 0$. *There exists a* $\xi_n \in \mathscr{S}_n$ *with* $P_{\Delta n}\xi_n = 0$ *and*

$$\|\xi_n - \xi\|_\infty = O_p(n^{-vp} + n^{-(1-v)/2}).$$

PROOF. According to Corollary 6.21 of Schumaker [(1981), page 227] there exists a $\xi_n^* \in \mathscr{S}_n$ such that $\|\xi_n^* - \xi\|_= O(n^{-vp})$. Let $n_\Delta = n^{-1}\sum_{i=1}^n \Delta_i$. Let $\xi_n = \xi_n^* - n_\Delta^{-1}n^{-1}\sum_{i=1}^n \Delta_i \xi_n^*(W_{ji}) = \xi_n^* - n_\Delta^{-1}P_{\Delta n}\xi_n^*$; then $P_{\Delta n}\xi_n = 0$. Because $|\xi_n - \xi| \le |\xi_n^* - \xi| + |P_{\Delta n}\xi_n^*|$, we only need to consider

$$P_{\Delta n}\xi_n^* = (P_{\Delta n} - P_\Delta)\xi_n^* + P_\Delta(\xi_n^* - \xi).$$

Since $(P_{\Delta n} - P_\Delta)\xi_n^* = O_p(n^{-1/2}n^{v/2})$, and $|P_\Delta(\xi_n^* - \xi)| \le E(\Delta)\|\xi_n^* - \xi\|_\infty = O(n^{-vp})$, the lemma follows from the triangle inequality. □

LEMMA A.6. *Denote* $m_0(t, \delta, x, z; g) = g(z) - \log S_0(t; g)$. *Let* η *be a positive constant. For any* g *with* $\|g - g_n\|_\infty \le \eta$ *and* $E[\Delta g(Z)] = 0$, *there exist constants* $0 < c_1, c_2 < \infty$ *such that*

$$\begin{aligned}&-c_1\|g - g_n\|_2^2 + O(n^{-2vp} + n^{-(1-v)}) \\ &\quad \le P_\Delta m_0(\cdot; g) - P_\Delta m_0(\cdot; g_n) \\ &\quad \le -c_2\|g - g_n\|_2^2 + O_p(n^{-2vp} + n^{-(1-v)}).\end{aligned}$$

PROOF. Let g_0 be the true value and let $h = g - g_0$. First consider

$$L_1(s) \equiv P_\Delta m_0(\cdot; g_0 + sh) - P_\Delta m_0(\cdot; g_0).$$

The first and the second derivatives of $L_1(s)$ are

$$L_1'(s) = P_\Delta\big[h - H_1(t,s)\big],$$
$$L''(s) = -P_\Delta\Big\{E\big[W_s(t)h^2(Z)\big] - \big[EW_s(t)h(Z)\big]^2\Big\}.$$

In particular,

$$L'(0) = P_\Delta\big[h - E(h|T=t,\delta=1)\big] = 0$$

and

$$\begin{aligned} L''(0) &= -P_\Delta\Big\{E\big[W_0(t)h^2(Z)\big] - \big[EW_0(t)h(Z)\big]^2\Big\} \\ &= -P_\Delta\Big\{EW_0(t)\big[h(Z) - EW_0(t)h(Z)\big]^2\Big\}. \end{aligned}$$

By the definition of $W_0(t)$,

$$E\big[W_0(t)|Z=z\big] = P(T\ge t|Z=z)\exp\big(g_0(z)\big)/S_0(t,g_0).$$

So there exist constants $c_1 > c_2 > 0$ such that

$$c_2 \le E\big[W_0(t)|Z=z\big] \le c_1.$$

It follows that

$$\begin{aligned} c_1P_\Delta E\big[h(Z) - EW_0(t)h(Z)\big]^2 &\le P_\Delta\Big\{EW_0(t)\big[h(Z) - EW_0(t)h(Z)\big]^2\Big\} \\ &\le c_2P_\Delta E\big[h(Z) - EW_0(t)h(Z)\big]^2. \end{aligned}$$

Now by (4.10), $EW_0(t)h(Z) = E[h(Z)|T=t,\Delta=1]$, and $E[\Delta h(Z)] = 0$, we have

$$\begin{aligned} &P_\Delta E\big[h(Z) - EW_0(t)h(Z)\big]^2 \\ &\quad = P_\Delta\,\delta E\,h^2 - 2Eh(Z)E\big[\Delta h(Z)\big] + P_\Delta\big[EW_0(t)h(Z)\big]^2 \\ &\quad = P_\Delta Eh^2 + P_\Delta\big[EW_0(t)h(Z)\big]^2. \end{aligned}$$

Furthermore, by Lemma A.4,

$$\big|L^{(3)}(s)\big| = O(1)\big[P_\Delta|h|^3 + P_\Delta|h|P_\Delta|h|^2\big] \le O(1)\eta P_\Delta|h|^2.$$

It follows that

$$-c_1\|g-g_0\|_2^2 \le P_\Delta m_0(\cdot,g) - P_\Delta m_0(\cdot,g_0) \le -c_2\|g-g_0\|_2^2.$$

The same argument as above gives that

$$\big|P_\Delta m_0(\cdot;g_n) - P_\Delta m_0(\cdot;g_0)\big| = O_p(1)\|g_n-g_0\|_2^2 = O_p\big(n^{-2vp} + n^{-(1-v)}\big),$$

where the second equality follows from Lemma A.5. Finally, since

$$\begin{aligned} &P_\Delta m_0(\cdot;g) - P_\Delta m_0(\cdot;g_n) \\ &\quad = P_\Delta m_0(\cdot;g) - P_\Delta m_0(\cdot;g_0) + P_\Delta m_0(\cdot;g_0) - P_\Delta m_0(\cdot;g_n), \end{aligned}$$

and by the triangle inequality,

$$\|g - g_n\|_2^2 - \|g_n - g_0\|_2^2 \le \|g - g_0\|_2^2 \le \|g - g_n\|_2^2 + \|g_n - g_0\|_2^2,$$

the lemma follows in view of Lemma A.5. □

LEMMA A.7. *Let*

$$I_{2n}(t) \equiv \frac{S_{1n}(t, \hat{g}_n)[h^*]}{S_{0n}(t; \hat{g}_n)} - \frac{S_{1n}(t, g_0)[h^*]}{S_{0n}(t; g_0)} - \left[\frac{S_1(t, \hat{g}_n)[h^*]}{S_0(t; \hat{g}_n)} - \frac{S_1(t, g_0)[h^*]}{S_0(t; g_0)}\right].$$

We have

$$\sup_{0 \le t \le 1} |I_{2n}(t)| = o_p(n^{-1/2}).$$

PROOF. Write $S_0(t, g_0) = S_0(g_0), S_1(t, g_0)[h^*] = S_1(g_0)$ and so on. Let

$$\begin{aligned}
A_{1n}(t) &= S_{1n}(g_0) - S_{1n}(\hat{g}_n) - [S_1(g_0) - S_1(\hat{g}_n)],\\
A_{2n}(t) &= [S_1(\hat{g}_n) - S_1(g_0)][S_0(\hat{g}_n) - S_{0n}(\hat{g}_n)],\\
A_{3n}(t) &= S_{0n}(g_0) - S_{0n}(\hat{g}_n) - [S_0(g_0) - S_0(\hat{g}_n)],\\
A_{4n}(t) &= [S_0(g_0) - S_0(\hat{g}_n)][S_{1n}(g_0) - S_1(g_0)],\\
A_{5n}(t) &= [S_0(g_0) - S_0(\hat{g}_n)][S_0(\hat{g}_n) - S_{0n}(\hat{g}_n)],\\
A_{6n}(t) &= [S_0(g_0) - S_0(\hat{g}_n)][S_0(g_0) - S_{0n}(g_0)],\\
A_{7n}(t) &= S_0(g_0)S_0(\hat{g}_n)S_{0n}(g_0)S_{0n}(\hat{g}_n).
\end{aligned}$$

Some algebra shows that

$$\begin{aligned}
A_{7n}(t)I_{2n}(t) = {} & S_0(g_0)S_0(\hat{g}_n)S_{0n}(g_0)A_{1n}(t) + S_0(g_0)S_{0n}(g_0)A_{2n}(t)\\
& + S_{1n}(g_0)S_0(g_0)S_0(\hat{g}_n)A_{3n}(t) + S_0(g_0)S_0(\hat{g}_n)A_{4n}(t)\\
& + S_0(t, g_0)S_1(g_0)A_{5n}(t) - S_1(g_0)S_{0n}(\hat{g}_n)A_{6n}(t).
\end{aligned}$$

Because there exists a constant $c > 0$ such that $\inf_{0 \le t \le 1} A_{7n}(t) \ge c$, and

$$\sup_{0 \le t \le 1} |A_{jn}(t)| = o_p(n^{-1/2}), \qquad 1 \le j \le 6,$$

the lemma follows from the triangle inequality. □

REFERENCES

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

ANDERSEN, P. K., BORGAN, O., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

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

BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–598.
COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220.
DABROWSKA, D. M. (1997). Smoothed Cox regression. *Ann. Statist.* **25** 1510–1540.
FLEMING, T. R. and HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.
GRAMBSCH, P. M., THERNEAU, T. M. and FLEMING, T. R. (1990). Martingale-based residuals for survival models. *Biometrika* **77** 147–160.
GRAMBSCH, P. M., THERNEAU, T. M. and FLEMING T. R. (1995). Diagnostic plots to reveal functional form for covariates in multiplicative intensity models. *Biometrics* **51** 1469–1482.
HASTIE, T. and TIBSHIRANI, R. (1986). Generalized additive models. *Statist. Sci.* **1** 297–318.
HASTIE, T. and TIBSHIRANI, R. (1990). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics* **46** 1005–1016.
HUANG, J. (1996). Efficient estimation for the Cox model with interval censoring. *Ann. Statist.* **24** 540–568.
KOOPERBERG, C., STONE, C. and TRUONG, Y. K. (1995). The L2 rate of convergence for hazard regression. *Scand. J. Statist.* **22** 143–158.
O'SULLIVAN, F. (1993). Nonparametric estimation in the Cox model. *Ann. Statist.* **21** 124–145.
SASIENI, P. (1992a). Information bounds for the conditional hazard ratio in a nested family of regression models. *J. Roy. Statist. Soc. Ser. B* **54** 617–635.
SASIENI, P. (1992b). Non-orthogonal projections and their application to calculating the information in a partly linear Cox model. *Scand. J. Statist.* **19** 215–233.
SCHUMAKER, L. (1981). *Spline Functions: Basic Theory*. Wiley, New York.
SHEN, X. and WONG, W. H. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22** 580–615.
STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705.
STONE, C. J. (1986a). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590–606.
STONE, C. J. (1986b). Comment on "Generalized Additive Models" by T. Hastie and R. Tibshirani. *Statist. Sci.* **1** 312–314.
STONE, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Ann. Statist.* **22** 118–184.
TSIATIS, A. A. (1981). A large sample study of Cox's regression model. *Ann. Statist.* **9** 93–108.
VAN DER VAART, A. W. (1991). On differentiable functionals. *Ann. Statist.* **19** 178–204.
VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

DEPARTMENT OF STATISTICS
AND ACTUARIAL SCIENCE
UNIVERSITY OF IOWA
IOWA CITY, IOWA 52242
E-MAIL: jian@stat.uiowa.edu