The Annals of Statistics
2016, Vol. 44, No. 4, 1708–1738
DOI: 10.1214/16-AOS1441

OPTIMAL ESTIMATION FOR THE FUNCTIONAL COX MODEL

BY SIMENG QU*, JANE-LING WANG[1,†] AND XIAO WANG[2,*]

*Purdue University** *and University of California, Davis*†

Functional covariates are common in many medical, biodemographic and neuroimaging studies. The aim of this paper is to study functional Cox models with right-censored data in the presence of both functional and scalar covariates. We study the asymptotic properties of the maximum partial likelihood estimator and establish the asymptotic normality and efficiency of the estimator of the finite-dimensional estimator. Under the framework of reproducing kernel Hilbert space, the estimator of the coefficient function for a functional covariate achieves the minimax optimal rate of convergence under a weighted L_2-risk. This optimal rate is determined jointly by the censoring scheme, the reproducing kernel and the covariance kernel of the functional covariates. Implementation of the estimation approach and the selection of the smoothing parameter are discussed in detail. The finite sample performance is illustrated by simulated examples and a real application.

1. Introduction. The proportional hazard model, known as the Cox model, was introduced by Cox (1972), where the hazard function of the survival time T for a subject with covariate $Z(t) \in \mathbb{R}^p$ is represented by

$$h(t|Z) = h_0(t)e^{\theta_0' Z(t)}, \tag{1.1}$$

where h_0 is an unspecified baseline hazard function and $\theta_0 \in \mathbb{R}^p$ is an unknown parameter. Some or all of the p components in Z may be time-independent, meaning that they are constant over time t, or may depend on t. The aim of this paper is to develop a different type of model, the functional Cox model, by incorporating functional predictors along with scalar predictors. Chen et al. (2011) first proposed such a model when studying the survival of diffuse large-B-cell lymphoma (DLBCL) patients, which is thought to be influenced by genetic differences. The functional predictor, denoted by $X(\cdot) : \mathcal{S} \to \mathbb{R}$ on a compact domain $\mathcal{S}$, is a smooth stochastic process related to the high-dimensional microarray gene expression of DLBCL patients. The entire trajectory of X has an effect on the hazard function, which makes it different from the Cox model (1.1) with time-varying covariates, where only the current value of X at time t affects the hazard function at time t.

Received January 2015; revised January 2016.
[1]Supported by NSF Grant DMS-09-06813.
[2]Supported by NSF Grants CMMI-1030246 and DMS-10-42967.
MSC2010 subject classifications. 62C20, 62G05, 62N01, 62N02.
Key words and phrases. Cox models, functional data, minimax rate of convergence, partial likelihood, right-censored data.

Specifically, the functional Cox model with a vector covariate Z and functional covariate $X(t)$ represents the hazard function by

$$h(t|X) = h_0(t)\exp\Big\{\theta_0' Z + \int_{\mathcal{S}} X(s)\beta_0(s)\,ds\Big\}, \tag{1.2}$$

where β_0 is an unknown coefficient function. Without loss of generality, we take $\mathcal{S}$ to be $[0, 1]$.

Under the right censorship model and letting T^u and T^c be, respectively, the failure time and censoring time, we observe i.i.d. copies of $(T, \Delta, X(s), s \in \mathcal{S})$, $(T_1, \Delta_1, X_1), \ldots, (T_n, \Delta_n, X_n)$, where $T = \min\{T^u, T^c\}$ is the observed time event and $\Delta = I\{T^u \le T^c\}$ is the censoring indicator. Our goal is to estimate $\alpha_0 = (\theta_0, \beta_0(\cdot))$ to reveal how the functional covariates $X(\cdot)$ and other scalar covariates Z relate to survival.

Let $\hat{\alpha} = (\hat{\theta}, \hat{\beta}(\cdot))$ be an estimate from the data. It is critical to define the risk function to measure the accuracy of the estimate. Let $W = (Z, X)$ and

$$\eta_\alpha(W) = \theta' Z + \int_0^1 \beta(s)X(s)\,ds.$$

Define an L_2-distance such that

$$d^2(\hat{\alpha}, \alpha_0) = \mathbb{E}\big\{\Delta\big(\eta_{\hat{\alpha}}(W) - \eta_{\alpha_0}(W)\big)^2\big\}. \tag{1.3}$$

Based on this L_2-distance, we show that the accuracy of $\hat{\theta}$ is measured by the usual L_2-norm $\|\hat{\theta} - \theta\|_2$ and the accuracy of $\hat{\beta}$ is measured by a weighted L_2-norm $\|\hat{\beta} - \beta_0\|_{C_\Delta}$, where

$$C_\Delta(s,t) = \mathrm{Cov}\big(\Delta X(s), \Delta X(t)\big) \quad \text{and} \quad \|\beta\|_{C_\Delta}^2 = \int\int \beta(s)C_\Delta(s,t)\beta(t)\,ds\,dt.$$

We now explain why we do not consider the convergence of $\hat{\beta}$ with respect to the usual L_2-norm in the present paper. In general, $\|\hat{\beta} - \beta_0\|_2^2 = \int_0^1 (\hat{\beta}(t) - \beta_0(t))^2\,dt$ may not converge to zero in probability, and to obtain the convergence of $\|\hat{\beta} - \beta_0\|_2^2$ one needs additional smoothness conditions linking β to the functional predictor X; see Crambes, Kneip and Sarda (2009) for a discussion of this phenomenon for functional linear models. On the other hand, in the presence of censoring, the Kullback–Leibler distance between two probability measures $\mathbb{P}_{h_0,\hat{\alpha}}$ and $\mathbb{P}_{h_0,\alpha_0}$ is equivalent to the L_2 distance d in (1.3). When failure times T^u are fully observed, that is, $\Delta = 1$ is true regardless of $X(s)$, the $\|\cdot\|_{C_\Delta}$ norm becomes $\|\cdot\|_C$, where $C(t,s) = \mathrm{Cov}(X(t), X(s))$ is the covariance function of X. This norm $\|\cdot\|_C$ has been widely used for functional linear models [e.g., Cai and Yuan (2012)].

Many people have studied parametric, nonparametric, or semiparametric modeling of the covariate effects using the Cox model [e.g., Sasieni (1992a, 1992b), Hastie and Tibshirani (1986, 1990), Huang (1999) and references therein] and Cox (1972) proposed to use partial likelihood to estimate θ in (1.1). The advantage

of using partial likelihood is that it estimates θ without knowing or involving the functional form of h_0. The asymptotic equivalence of the partial likelihood estimator and the maximum likelihood estimator has been established by several authors [Cox (1975), Tsiatis (1981), Andersen and Gill (1982), Johansen (1983), Jacobsen (1984)]. On the other hand, the literature on functional regression, in particular for functional linear models, is too vast to be summarized here. Hence, we only refer to the well-known monographs Ramsay and Silverman (2005) and Ferraty and Vieu (2006), and some recent developments such as James and Hastie (2002), Müller and Stadtmüller (2005), Hall and Horowitz (2007), Crambes, Kneip and Sarda (2009), Yuan and Cai (2010), Cai and Yuan (2012) for further references. Recently, Kong et al. (2014) studied a similar functional Cox model to establish some asymptotic properties but without investigating the optimality property. Moreover, their estimate of the parametric component converges at a rate which is slower than root-n. Thus, it is desirable to develop new theory to systematically investigate properties of the estimates and establish their optimal asymptotic properties. In addition, instead of assuming that both β_0 and X can be represented by the same set of basis functions, we adopt a more general reproducing kernel Hilbert space framework to estimate the coefficient function.

In this paper, we study the convergence of the estimator $\hat{\alpha} = (\hat{\theta}, \hat{\beta})$ under the framework of the reproducing kernel Hilbert space and the Cox model. The true coefficient function β_0 is assumed to reside in a reproducing kernel Hilbert space $\mathcal{H}(K)$ with the reproducing kernel K, which is a subspace of the collection of square integrable functions on $[0, 1]$. There are two main challenges for our asymptotic analysis, the nonlinear structure of the Cox model, and the fact that the reproducing kernel K and the covariance kernel C_Δ may not share a common ordered set of eigenfunctions, so β_0 cannot be represented effectively by the leading eigenfunctions of C_Δ. We obtain the estimator by maximizing a penalized partial likelihood and establish $\sqrt{n}$-consistency, asymptotic normality and semiparametric efficiency of the estimator $\hat{\theta}$ of the finite-dimensional regression parameter.

A second optimality result is on the estimator of the coefficient function, which achieves the minimax optimal rate of convergence under the weighted L_2-risk. The optimal rate of convergence is established in the following two steps. First, the convergence rate of the penalized partial likelihood estimator is calculated. Second, in the presence of the nuisance parameter h_0, the minimax lower bound on the risk is derived, which matches the convergence rate of the partial likelihood estimator. Therefore, the estimator is rate-optimal. Furthermore, an efficient algorithm is developed to estimate the coefficient function. Implementation of the estimation approach, selection of the smoothing parameter, as well as calculation of the information bound $I(\theta)$ are all discussed in detail.

The rest of the paper is organized as follows. Section 2 summarizes the main results regarding the asymptotic analysis of the penalized partial likelihood predictor. Implementation of the estimation approach is discussed in Section 3, including a GCV method to select the smoothing parameter and a method of calculating the

information bound of θ based on the alternating conditional expectations (ACE) algorithm. Section 4 contains numerical studies, including simulations and a data application. Proofs are relegated to Section 5, and more technical details are provided in the supplemental article [Qu, Wang and Wang (2016)].

2. Main results. We estimate $\alpha_0 = (\theta_0, \beta_0) \in \mathbb{R}^p \times \mathcal{H}(K)$ by maximizing the penalized log partial likelihood,

$$\hat{\alpha}_\lambda = \underset{\alpha \in \mathbb{R}^p \times \mathcal{H}(K)}{\arg\min} \; l_n(\alpha) + \lambda J(\beta), \tag{2.1}$$

where the negative log partial likelihood is given by

$$l_n(\alpha) = -\frac{1}{n} \sum_{i=1}^{n} \Delta_i \left\{ \eta_\alpha(W_i) - \log \sum_{T_j \geq T_i} \exp\bigl(\eta_\alpha(W_j)\bigr) \right\}, \tag{2.2}$$

J is a penalty function controlling the smoothness of β, and λ is a smoothing parameter that balances the fidelity to the model and the plausibility of β. The choice of the penalty function $J(\cdot)$ is a squared seminorm associated with $\mathcal{H}$ and its norm. In general, $\mathcal{H}(K)$ can be decomposed with respect to the penalty J as $\mathcal{H} = \mathcal{N}_J + \mathcal{H}_1$, where $\mathcal{N}_J$ is the null space defined as

$$\mathcal{N}_J = \bigl\{ \beta \in \mathcal{H}(K) : J(\beta) = 0 \bigr\},$$

and $\mathcal{H}_1$ is its orthogonal complement in $\mathcal{H}$. Correspondingly, the kernel K can be decomposed as $K = K_0 + K_1$, where K_0 and K_1 are kernels for the subspace $\mathcal{N}_J$ and $\mathcal{H}_1$, respectively. For example, for the Sobolev space,

$$\mathcal{W}_{2,m} = \bigl\{ f : [0,1] \to R | f, f', \ldots, f^{(m-1)} \text{ are absolutely continuous}, f^{(m)} \in L_2 \bigr\},$$

endowed with the norm

$$\|f\|_{\mathcal{W}_{2,m}} = \sum_{v=0}^{m-1} f^{(v)}(0) + \int_0^1 \bigl(f^{(m)}(s)\bigr)^2 \, ds, \tag{2.3}$$

where the penalty $J(\cdot)$ in this case can be assigned as $J(f) = \int_0^1 (f^{(m)}(s))^2 \, ds$.

We first present some main assumptions:

(A1) Assume $\mathbb{E}(\Delta Z) = 0$ and $\mathbb{E}(\Delta X(s)) = 0$, $s \in [0,1]$.

(A2) The failure time T^u and the censoring time T^c are conditionally independent given W.

(A3) The observed event time T_i, $1 \leq i \leq n$ is in a finite interval, say $[0, \tau]$, and there exists a small positive constant ε such that: (i) $\mathbb{P}(\Delta = 1 | W) > \varepsilon$, and (ii) $\mathbb{P}(T^c > \tau | W) > \varepsilon$ almost surely with respect to the probability measure of W.

(A4) The covariate Z takes values in a bounded subset of $\mathbb{R}^p$, and the L_2-norm $\|X\|_2$ of X is bounded almost surely.

(A5) Let $0 < c_1 < c_2 < \infty$ be two constants. The baseline joint density $f(t, \Delta = 1)$ of $(T, \Delta = 1)$ satisfies $c_1 < f(t, \Delta = 1) < c_2$ for all $t \in [0, \tau]$.

Condition (A1) requires Z and X to be suitably centered. Since the partial likelihood function (2.2) does not change when centering Z_i as $Z_i - \sum \Delta_i Z_i / \sum \Delta_i$ or X_i as $X_i - \sum \Delta_i X_i / \sum \Delta_i$, centering does not impose any real restrictions. In addition, centering by $\mathbb{E}(\Delta Z)$ and $\mathbb{E}(\Delta X)$, instead of centering by $\mathbb{E}(Z)$ and $\mathbb{E}(X)$, simplifies the asymptotic analysis. Conditions (A2) and (A3) are common assumptions for analyzing right-censored data, where (A2) guarantees the censoring mechanism to be noninformative while (A3) avoids the unboundedness of the partial likelihood at the end point of the support of the observed event time. This is a reasonable assumption since the experiment can only last for a certain amount of time in practice. Assumption (A3)(i) further ensures the probability of being uncensored to be positive regardless of the covariate and (A3)(ii) controls the censoring rate so that it will not be too heavy. Assumption (A4) places a boundedness restriction on the covariates. This assumption can be relaxed to the sub-Gaussianity of $\|X\|_2$, which implies that with a large probability, $\|X\|_2$ is bounded. Condition (A5) and condition (A1) together guarantee the identifiability of the model. Moreover the joint density $f(T, Z, X, \Delta = 1)$ is bounded away from zero and infinity under assumptions (A3)–(A5), which is used to calculate the information bound and convergence rate later in Theorem 2.1 and Theorem 2.2.

Let $r(W) = \exp(\eta_\alpha(W))$, then the counting process martingale associated with model (1) is

$$M(t) = M(t|W) = \Delta I\{T \le t\} - \int_0^t I\{T \ge u\} r(W)\, dH_0(u),$$

where $H_0(t) = \int_0^t h_0(u)\, du$ is the baseline cumulative hazard function. For two sequences $a_k : k \ge 1$ and $b_k : k \ge 1$ of positive real numbers, we write $a_k \asymp b_k$ if there are positive constants c and C independent of k such that $c \le a_k / b_k \le C$ for all $k \ge 1$.

THEOREM 2.1. *Under* (A1)–(A5), *the efficient score for the estimation of* θ *is*

$$l_\theta^*(T, \Delta, W) = \int_0^\tau \big(Z - a^*(t) - \eta_{g^*}(X)\big)\, dM(t),$$

where $(a^*, g^*) \in L_2 \times \mathcal{H}(K)$ *is a solution that minimizes*

$$\mathbb{E}\{\Delta \|Z - a(T) - \eta_g(X)\|^2\}.$$

Here, a^* *can be expressed as* $a^*(t) = \mathbb{E}[Z - \eta_{g^*}(X)|T = t, \Delta = 1]$. *The information bound for the estimation of* θ *is*

$$I(\theta) = \mathbb{E}\big[l_\theta^*(T, \Delta, W)\big]^{\otimes 2} = \mathbb{E}\big\{\Delta\big[Z - a^*(T) - \eta_{g^*}(X)\big]^{\otimes 2}\big\},$$

where $y^{\otimes 2} = yy'$ *for column vector* $y \in \mathbb{R}^d$.

Recall that K and C_Δ are two real, symmetric and nonnegative definite functions. Define a new kernel $K^{1/2}C_\Delta K^{1/2} : [0,1]^2 \to \mathbb{R}$, which is a real, symmetric, square integrable and nonnegative definite function. Let $L_{K^{1/2}C_\Delta K^{1/2}}$ be the corresponding linear operator $L_2 \to L_2$. Then Mercer's theorem [Riesz and Sz-Nagy (1990)] implies that there exists a set of orthonomal eigenfunctions $\{\phi_k : k \geq 1\}$ and a sequence of eigenvalues $s_1 \geq s_2 \geq \cdots > 0$ such that

$$K^{1/2}C_\Delta K^{1/2}(s,t) = \sum_{k=1}^{\infty} s_k \phi_k(s)\phi_k(t), \qquad L_{K^{1/2}C_\Delta K^{1/2}}(\phi_k) = s_k.$$

THEOREM 2.2. *Assume* (A1)–(A5) *hold*:

(i) (*Consistency*) $d(\hat{\alpha}, \alpha_0) \xrightarrow{p} 0$, *provided that* $\lambda \to 0$ *as* $n \to \infty$.

(ii) (*Convergence rate*) *If the eigenvalues* $\{s_k : k \geq 1\}$ *of* $K^{1/2}C_\Delta K^{1/2}$ *satisfy* $s_k \asymp k^{-2r}$ *for some constant* $0 < r < \infty$, *then for* $\lambda = O(n^{-2r/(2r+1)})$ *we have*

$$d(\hat{\alpha}, \alpha_0) = O_p\big(n^{-r/(2r+1)}\big).$$

(iii) *If* $I(\theta)$ *is nonsingular, then* $\|\hat{\theta} - \theta_0\|_2 = O_p(n^{-r/(2r+1)})$ *and*

$$\lim_{A\to\infty} \lim_{n\to\infty} \sup_{\beta_0 \in \mathcal{H}(K)} \mathbb{P}_{h_0\beta_0}\big\{\|\hat{\beta}_\lambda - \beta_0\|_{C_\Delta} \geq A n^{-r/(2r+1)}\big\} = 0.$$

Theorem 2.2 indicates that the convergence rate is determined by the decay rate of the eigenvalues of $K^{1/2}C_\Delta K^{1/2}$, which is jointly determined by the eigenvalues of both reproducing kernel K and the conditional covariance function C_Δ as well as by the alignment between K and C_Δ. When K and C_Δ are perfectly aligned, meaning that K and C_Δ have the same ordered eigenfunctions, the decay rate of $\{s_k : k \geq 1\}$ equals to the summation of the decay rates of the eigenvalues of K and C_Δ. Cai and Yuan (2012) established a similar result for functional linear models, for which the optimal prediction risk depends on the decay rate of the eigenvalues of $K^{1/2}CK^{1/2}$, where C is the covariance function of X.

The next theorem establishes the asymptotic normality of $\hat{\theta}$ with root-n consistency.

THEOREM 2.3. *Suppose* (A1)–(A5) *hold, and that the Fisher information* $I(\theta_0)$ *is nonsingular. Let* $\hat{\alpha} = (\hat{\theta}, \hat{\beta})$ *be the estimator given by* (2.1) *with* $\lambda = O(n^{-2r/(2r+1)})$. *Then*

$$\sqrt{n}(\hat{\theta} - \theta_0) = n^{-1/2} I^{-1}(\theta_0) \sum_{i=1}^{n} l^*_{\theta_0}(T_i, \Delta_i, W_i) + o_p(1) \xrightarrow{d} \mathcal{N}(0, \Sigma),$$

where $\Sigma = I^{-1}(\theta_0)$.

For the nonparametric coefficient function β, it is of interest to see whether the convergence rate of $\hat{\beta}$ in Theorem 2.2 is optimal. In the following, we derive a minimax lower bound for the risk.

THEOREM 2.4. *Assume that the baseline hazard function* $h_0 \in \mathcal{F} = \{h : H(t) = \int_0^t h(s)\,ds < \infty$, *for any* $0 < t < \infty\}$. *Suppose that the eigenvalues* $\{s_k : k \geq 1\}$ *of* $K^{1/2}C_\Delta K^{1/2}$ *satisfy* $s_k \asymp k^{-2r}$ *for some constant* $0 < r < \infty$. *Then*

$$\lim_{a\to 0}\lim_{n\to\infty}\inf_{\hat{\alpha}}\sup_{\alpha_0\in\mathbb{R}^p\times\mathcal{H}(K)}\sup_{h_0\in\mathcal{F}}\mathbb{P}_{\alpha_0,h_0}\{\|\hat{\beta}-\beta_0\|_{C_\Delta} \geq an^{-r/(2r+1)}\} = 1,$$

where the infimum is taken over all possible predictors $\hat{\alpha}$ *based on the observed data*.

Theorem 2.4 shows that the minimax lower bound of the convergence rate for estimating β_0 is $n^{-r/(2r+1)}$, which is determined by r and the decay rate of the eigenvalues of $K^{1/2}C_\Delta K^{1/2}$. We have shown that this rate is achieved by the penalized partial likelihood predictor and, therefore, this estimator is rate-optimal.

3. Computation of the estimator.

3.1. *Penalized partial likelihood.* In this section, we present an algorithm to compute the penalized partial likelihood estimator. Let $\{\xi_1, \ldots, \xi_m\}$ be a set of orthonormal basis of the null space with $m = \dim(\mathcal{N}_J)$. The next theorem provides a closed form representation of $\hat{\beta}$ from the penalized partial likelihood method.

THEOREM 3.1. *The penalized partial likelihood estimator of the coefficient function is given by*

$$\hat{\beta}_\lambda(t) = \sum_{k=1}^{m} d_k\xi_k(t) + \sum_{i=1}^{n} c_i\int_0^1 X_i(s)K_1(s,t)\,ds, \tag{3.1}$$

where d_k $(k = 1, \ldots, m)$ *and* c_i $(i = 1, \ldots, n)$ *are constant coefficients*.

Theorem 3.1 is a direct application of the generalized version of the well-known representer lemma for smoothing splines [see Wahba (1990) and Yuan and Cai (2010)]. We omit the proof here. In fact, the algorithm can be made more efficient without using all n bases $\int_0^1 X_i(s)K_1(s,t)\,ds$, $i = 1, \ldots, n$ in (3.1). Gu (2013) showed that, under some conditions, a more efficient estimator, denoted by β_λ^*, sharing the same convergence rate with $\hat{\beta}_\lambda$, can be calculated in the data-adaptive finite-dimensional space

$$\mathcal{H}^* = \mathcal{N}_J \oplus \{K_1(\tilde{X}_j, \cdot), j = 1, \ldots, q\},$$

where $\{\tilde{X}_j\}$ is a random subset of $\{X_i : \Delta_i = 1\}$ and

$$K_1(\tilde{X}_j, \cdot) = \int_0^1 \tilde{X}_j(s)K_1(s, \cdot)\,ds.$$

Here, $q = q_n \asymp n^{2/(ps+1)+\epsilon}$ for some $s > 1$ and $p \in [1, 2]$, and for any $\epsilon > 0$. Therefore, β_λ^* is given by

$$\beta_\lambda^*(t) = \sum_{k=1}^{m} d_k \xi_k(t) + \sum_{j=1}^{q} c_j K_1(\tilde{X}_j, t).$$

The computational efficiency is more prominent when n is large, as the number of coefficients is significantly reduced from $n + m$ to $q + m$.

For the Sobolev space $\mathcal{W}_{2,m}$, the penalty function $J(\cdot)$ is

$$J(f) = \int_0^1 \big(f^{(m)}(s)\big)^2 \, ds,$$

and (2.1) becomes

$$(\hat{\theta}, \hat{\beta}_\lambda) = \underset{\theta \in \mathbb{R}^p, \beta \in \mathcal{W}_{2,m}}{\arg\min} \; -\frac{1}{n} \sum_{i=1}^{n} \Delta_i \left\{ \eta_\alpha(W_i) - \log \sum_{T_j > T_i} \exp\big(\eta_\alpha(W_j)\big) \right\} + \lambda \int_0^1 \big(\beta^{(m)}(s)\big)^2 \, ds. \tag{3.2}$$

Let $\xi_\nu = t^{\nu-1}/(\nu-1)!$, $\nu = 1, \ldots, m$, be the orthonormal basis of the null space

$$\mathcal{N}_J = \left\{ \beta \in \mathcal{W}_{2,m}, \int_0^1 \big(\beta^{(m)}(s)\big)^2 \, ds = 0 \right\}.$$

Write $G_m(t, u) = (t-u)_+^{m-1}/(m-1)!$, then the kernels are in forms of

$$K_0(s, t) = \sum_{\nu=1}^{m} \xi_\nu(s)\xi_\nu(t) \quad \text{and} \quad K_1(s, t) = \int_0^1 G_m(s, u) G_m(t, u) \, du.$$

Hence, the estimator is given by

$$\hat{\beta}_\lambda(t) = \sum_{\nu=1}^{m} d_\nu \xi_\nu(t) + \sum_{i=1}^{n} c_i \int_0^1 X_i(s) K_1(s, t) \, ds. \tag{3.3}$$

We may obtain the constants c_i and d_j as well as the estimator $\hat{\theta}$ by maximizing the objective function (3.2) after plugging $\hat{\beta}_\lambda(t)$ back into the objective function.

3.2. *Choosing the smoothing parameter.* The choice of the smoothing parameter λ is always a critical but difficult question. In this section, we borrow ideas from Gu (2013) and provide a simple GCV method to choose λ. The key idea is to draw an analogy between the partial likelihood estimation and weighted density estimation, which then allows us to define a criterion analogous to the Kullback–Leibler distance to select the best performing smoothing parameter. Below we provide more details.

Let $i_1, \ldots, i_N$ be the index for the uncensored data, that is, $\Delta_{i_k} = 1$, for $k = 1, \ldots, N$ and $N = \sum_1^n \Delta_i$. Define weights $w_{i_k}(\cdot)$ as $w_{i_k}(t) = I\{t \geq T_{i_k}\}$ and

$$f_{\alpha|i_k}(t, w) = \frac{w_{i_k}(t)e^{\eta_\alpha(w)}}{\sum_{k=1}^N w_{i_k}(t)e^{\eta_\alpha(w)}}.$$

Following the suggestion in Section 8.5 of Gu (2013), we extend the Kullback–Leibler distance for density functions to the partial likelihood as follows:

$$\begin{aligned} KL(\hat{\alpha}_\lambda, \alpha) &= \frac{1}{N}\sum_{k=1}^N \mathbb{E}_{f_{\alpha_0|i_k}}\left\{\log \frac{f_{\alpha_0|i_k}(T_{i_k}, W_{i_k})}{f_{\hat{\alpha}|i_k}(T_{i_k}, W_{i_k})}\right\} \\ &= \frac{1}{N}\sum_{k=1}^N \mathbb{E}_{f_{\alpha_0|i_k}}\left\{\log \frac{e^{\eta_{\alpha_0}(W_{i_k})}}{\sum_{j=1}^n w_{i_k}(T_j)e^{\eta_{\alpha_0}(W_j)}}\right. \\ &\quad \left. - \log \frac{e^{\eta_{\hat{\alpha}_\lambda}(W_{i_k})}}{\sum_{j=1}^n w_{i_k}(T_j)e^{\eta_{\hat{\alpha}_\lambda}(W_j)}}\right\}. \end{aligned}$$

Dropping off terms not involving $\hat{\alpha}_\lambda$, we have a relative KL distance

$$RKL(\hat{\alpha}_\lambda, \alpha) = -\frac{1}{N}\sum_{k=1}^N \mathbb{E}_{f_{\alpha_0|i_k}} \eta_{\hat{\alpha}_\lambda}(W) + \frac{1}{N}\sum_{k=1}^N \log \sum_{j=1}^n w_{i_k}(T_j)e^{\eta_{\hat{\alpha}_\lambda}(W_j)}.$$

The second term is ready to be computed once we have an estimate $\hat{\alpha}_\lambda$, but the first term involves α_0 and needs to be estimated. We approximate the RKL by

$$\widehat{RKL}(\hat{\alpha}_\lambda, \alpha_0) = -\frac{1}{n}\sum_{i=1}^n \eta_{\hat{\alpha}_\lambda}^{[i]}(W_i) + \frac{1}{N}\sum_{i=1}^n \Delta_i \log \sum_{T_j \geq T_i} \exp\{\eta_{\hat{\alpha}_\lambda}(W_j)\}.$$

Based on this $\widehat{RKL}(\hat{\alpha}_\lambda, \alpha_0)$, a function $GCV(\lambda)$ can be derived analytically when replacing the penalized partial likelihood function by its quadratic approximation,

$$\begin{aligned} GCV(\lambda) &= -\frac{1}{n}\sum_{i=1}^n \eta_{\hat{\alpha}_\lambda}(W_i) + \frac{1}{n(n-1)}\mathrm{tr}\big[(SH^{-1}S)(\mathrm{diag}\Delta - \Delta\mathbf{1}'/n)\big] \\ &\quad + \frac{1}{N}\sum_{i=1}^n \Delta_i \log \sum_{T_j \geq T_i} \exp\{\eta_{\hat{\alpha}_\lambda}(W_j)\}. \end{aligned}$$

Details of deriving $GCV(\lambda)$ are given in the supplemental article [Qu, Wang and Wang (2016)].

3.3. *Calculating the information bound $I(\theta)$.* To calculate the information bound $I(\theta)$, we apply the ACE method [Breiman and Friedman (1985)], the estimator of which is shown to converge to (a^*, g^*). For simplicity, we take Z as a

one-dimensional scalar. When Z is a vector, we just need to apply the following procedure to all dimensions of Z separately.

Theorem 2.1 shows that

$$I(\theta) = \mathbb{E}\{\Delta[Z - a^*(t) - \eta_{g^*}(X)]^{\otimes 2}\}$$

with $(a^*, g^*) \in L_2 \times \mathcal{H}(K)$ being the unique solution that minimizes

$$\mathbb{E}\{\Delta \|Z - a(T) - \eta_g(X)\|^2\}.$$

Furthermore, the proof of Theorem 2.1 reveals that this is equivalent to the following: (a^*, g^*) is the unique solution to the equations:

$$\mathbb{E}(Z - a^* - \eta_{g^*} | T, \Delta = 1) = 0, \qquad \text{a.s. } P_T^{(u)},$$

$$\mathbb{E}(Z - a^* - \eta_{g^*} | X, \Delta = 1) = 0, \qquad \text{a.s. } P_X^{(u)},$$

where $P_T^{(u)}$ and $P_X^{(u)}$ represent, respectively, the measure space of $(T, \Delta = 1)$ and $(X, \Delta = 1)$.

The idea of ACE is to update a and g alternatively until the objective function $e(a, g) = \mathbb{E}\Delta \|Z - a(T) - \eta_g(X)\|^2$ stops to decrease. In our case, the procedure is as follows:

(i) Initialize a and g,

(ii) Update a by

$$a(T) = \mathbb{E}(Z - \eta_g | T, \Delta = 1) = 0,$$

(iii) Update g such that

$$\eta_g(X) = \mathbb{E}(Z - a | X, \Delta = 1) = 0, \qquad \text{a.s. } P_X^{(u)},$$

(iv) Calculate $e(a, g) = \mathbb{E}\Delta \|Z - a(T) - \eta_g(X)\|^2$ and repeat (ii) and (iii) until $e(a, g)$ fails to decrease.

In practice, we replace $\mathbb{E}\Delta \|Z - a(T) - \eta_g(X)\|^2$ by the sample mean

$$e(a, g) = \frac{1}{n} \sum_{i=1}^{n} \Delta_i \|Z_i - a(T_i) - \eta_g(X_i)\|^2.$$

As for a and g, we need to employ some smoothing techniques. For a given $g \in \mathcal{H}(K)$, we calculate

$$\tilde{a}_i = \sum_{T_j = T_i} \Delta_j [Z_j - \eta_g(X_j)] \Big/ \sum_{T_j = T_i} \Delta_j,$$

and update $a(t)$ as the local polynomial regression estimator for the data $(T_1, \tilde{a}_1), \ldots, (T_n, \tilde{a}_n)$. For a given $a \in L_2$, we calculate

$$y_i = Z_i - a(T_i), \qquad \text{for all } \Delta_i = 1,$$

and update g by fitting a functional linear regression

$$y = \int g(s)X(s)\,ds + \epsilon,$$

based on the data (y_i, X_i) with $\Delta_i = 1$. More details can be find in Yuan and Cai (2010). When (a^*, g^*) is obtained, $I(\theta)$ is estimated by

$$\widehat{I(\theta)} = \frac{1}{n}\sum_{i=1}^{n} \Delta_i \big[Z_i - a^*(T_i) - \eta_{g^*}(X_i)\big]^{\otimes 2}.$$

4. Numerical studies. In this session, we first carry out simulations under different settings to study the finite sample performance of the proposed method and to demonstrate practical implications of the theoretical results. In the second part, we apply the proposed method to data that were collected to study the effect of early reproduction history to the longevity of female Mexican fruit flies.

4.1. *Simulations*. We adopt a similar design as that in Yuan and Cai (2010). The functional covariate X is generated by a set of cosine basis functions, $\phi_1 = 1$ and $\phi_{k+1}(s) = \sqrt{2}\cos(k\pi s)$ for $k \geq 1$, such that

$$X(s) = \sum_{k=1}^{50} \zeta_k U_k \phi_k(s),$$

where the U_k are independently sampled from the uniform distribution on $[-3, 3]$ and $\zeta_k = (-1)^{k+1}k^{-v/2}$ with $v = 1, 1.5, 2, 2.5$. In this case, the covariance function of X is $C(s,t) = \sum_{k=1}^{50} 3k^{-v}\phi_k(s)\phi_k(t)$. The coefficient function β_0 is

$$\beta_0 = \sum_{i=1}^{50} (-1)^k k^{-3/2}\phi_k,$$

which is from a Sobolov space $\mathcal{W}_{2,2}$. The reproducing kernel takes the form:

$$K(s,t) = 1 + st + \int_0^1 (s-u)_+(t-u)_+\,du,$$

and $K_1 = \int_0^1 (s-u)_+(t-u)_+\,du$. The null space becomes $\mathcal{N}_J = \text{span}\{1, s\}$. The penalty function as mentioned before is $J(f) = \int (f'')^2$. The vector covariate Z is set to be univariate with distribution $\mathcal{N}(0,1)$ and corresponding slope $\theta = 1$. The failure time T^u is generated based on the hazard function

$$h(t) = h_0(t)\exp\left\{\theta' Z + \int_0^1 X(s)\beta_0(s)\,ds\right\},$$

where $h_0(t)$ is chosen as a constant or a linear function t. Given X, T^u follows an exponential distribution when h_0 is a constant, and follows a Weibull distribution when $h_0(t) = t$. The censoring time T^c is generated independently, following an exponential distribution with parameter γ which controls the censoring rate. When $h_0(t)$ is constant, $\gamma = 19$ and 3.4 lead to censoring rates around 10% and 30%,

respectively. Similar censoring rates result from $\gamma = 15$ and 3.9 for the case when $h_0(t) = t$. (T, Δ) is then generated by $T = \min\{T^u, T^c\}$ and $\Delta = I\{T^u \le T^c\}$.

The criterion to evaluate the performance of the estimators $\hat{\beta}$ is the mean squared error, defined as

$$MSE(\hat{\beta}) = \left\{ \frac{1}{\sum_{i=1}^n \Delta_i} \sum_{i=1}^n \Delta_i \big(\eta_{\hat{\beta}}(X_i) - \eta_{\beta_0}(X_i)\big) \right\}^{1/2},$$

which is an empirical version of $\|\hat{\beta} - \beta_0\|_{C_\Delta}$. To study the trend as the sample size increases, we vary the sample size n according to $n = 50, 100, 150, 200$ for each value $v = 1, 1.5, 2, 2.5$. For each combination of censoring rate, h_0, v and n, the simulation is repeated 1000 times, and the average mean squared error was obtained for each scenario.

Note that for a fixed γ, $\mathbb{E}(\Delta|X)$ is roughly a constant for different values of v. Therefore, $C_\Delta(s,t)$ is approximately proportional to $C(s,t) = \sum_{k=1}^{50} k^{-v}\phi_k(s) \times \phi_k(t)$. In this case, v controls the decay rate of the eigenvalues of C_Δ and $K^{1/2}C_\Delta K^{1/2}$. It follows from Theorem 2.2 that a faster decay rate of the eigenvalues leads to a faster convergence rate. Figure 1 displays the average MSE based on 1000 simulations. The simulation results are in agreement with Theorem 2.2; it is very clear that when v increases from 1 to 2.5 with the remaining parameters fixed, the average MSEs decrease steadily. The average MSEs also decrease with the sample sizes. Besides, for both the exponential and Weibull distribution, the average MSEs are lower for each setting at the 10% censoring rate comparing to the values for the 30% censoring rate. This is consistent with the expectation that the lower the censoring rate is, the more accurate the estimate will be.

Averages and standard deviations of the estimated $\hat{\theta}$, for each setting of v and n over 1000 repetition for the case of $h_0 = c$ and 30% censoring rate, are given in Table 1. For each case of v, as n increases, the average of $\hat{\theta}$ gets closer to the true value and the standard deviation decreases. Noting that the results do not vary much across different values of v, as v is specially designed to examine the estimation of β and has little effect on the estimation of θ.

For each simulated dataset, we also calculated the information bound $I(\theta)$ based on the ACE method proposed in Section 3.3. The inverse of this information bound, as suggested by Theorem 2.3, can be used to estimate the asymptotic variance of $\hat{\theta}$. We further used these asymptotic variance estimates to construct a 95% confidence interval for θ. Table 2 shows the observed percentage the constructed 95% confidence interval covered the true value 1 for the various settings. As expected, the covering rates increase toward 95% as n gets larger. Results for other choices of h_0 and censoring rates were about the same and are omitted.

4.2. *Mexican fruit fly data.* We now apply the proposed method to the Mexican fruit fly data in Carey et al. (2005). There were 1152 female flies in that paper

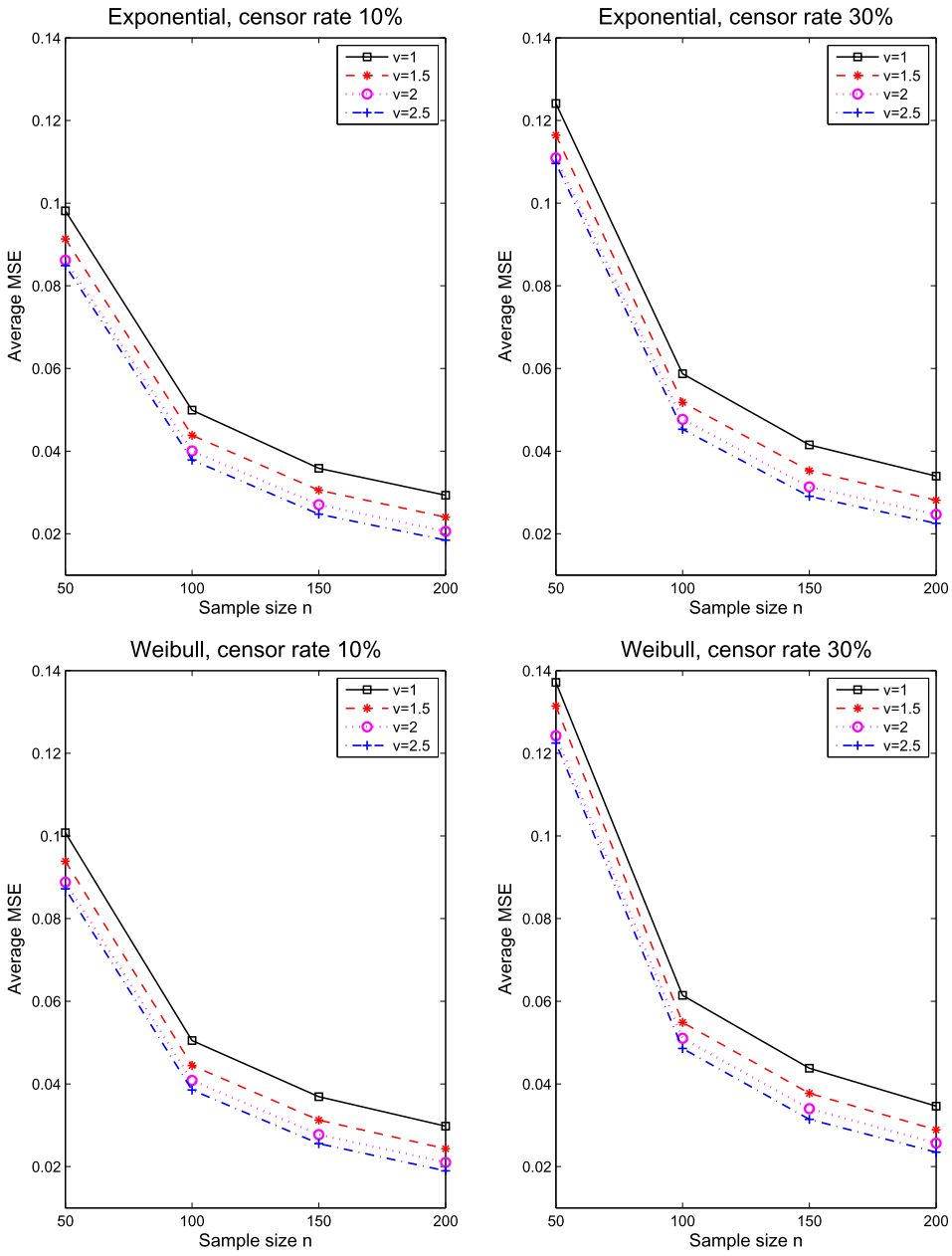

FIG. 1. *The average MSE based on* 1000 *simulations. The top panel is for the constant baseline hazard function and the bottom panel is for the linear baseline hazard function. For each panel, from left to right, the censoring rate is controlled to be around* 10% *and* 30%. *The sample sizes are* $n = 50, 100, 150, 200$ *and the decay rate parameters are* $v = 1, 1.5, 2, 2.5$.

TABLE 1
Average and standard deviation of $\hat{\theta}$
($h_0 = c$, 30% *censoring rate*)

n	$v = 1$	$v = 1.5$	$v = 2$	$v = 2.5$
50	1.061	1.064	1.064	1.065
	(0.264)	(0.265)	(0.264)	(0.265)
100	1.027	1.030	1.031	1.031
	(0.164)	(0.164)	(0.164)	(0.163)
150	1.013	1.016	1.017	1.018
	(0.133)	(0.132)	(0.131)	(0.131)
200	1.011	1.013	1.015	1.016
	(0.111)	(0.111)	(0.110)	(0.110)

coming from four cohorts; for illustration purposes, we are using the data from cohort 1 and cohort 2, which consist of the lifetime and daily reproduction (in terms of number of eggs laid daily) of 576 female flies.

We are interested in whether and how early reproduction will affect the lifetime of female Mexican fruit flies. For this reason, we exclude 28 infertile flies from cohort 1 and 20 infertile flies from cohort 2. The period for early reproduction is chosen to be from day 6 to day 30 based on the average reproduction curve (Figure 2), which shows that no flies laid any eggs before day 6 and the peak of reproduction was day 30. Once the period of early reproduction was determined to be [6, 30], we further excluded flies that died before day 30 to guarantee a fully observed trajectory for all flies and this leaves us with a total of 479 flies for further exploration of the functional Cox model. The mean and median lifetime of the remaining 224 flies in cohort 1 is 56.41 and 58 days, respectively; the mean and the median lifetime of the remaining 255 flies in cohort 2 is 55.78 and 55 days, respectively.

The trajectories of early reproduction for these 479 flies are of interest to researchers but they are very noisy, so for visualization we display the smoothed

TABLE 2
Covering rate of the 95% *confidence intervals for* θ
($h_0 = c$, 30% *censoring rate*)

n	$v = 1$	$v = 1.5$	$v = 2$	$v = 2.5$
50	91.5%	91.9%	92.0%	91.5%
100	93.3%	92.4%	92.4%	93.0%
150	93.5%	93.1%	93.9%	93.4%
200	93.6%	93.7%	93.9%	93.8%

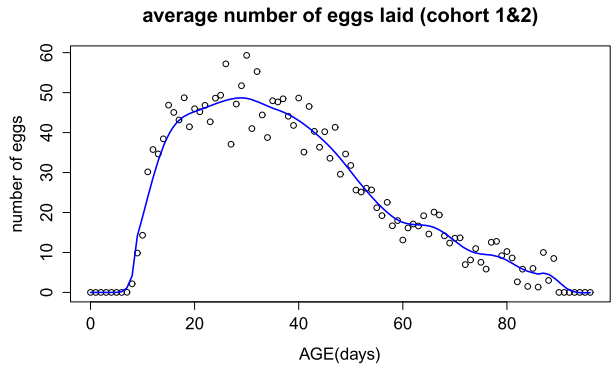

FIG. 2. *Average number of eggs laid daily for both cohorts.*

egg-laying curves for the first 100 flies (Figure 3). The data of these 100 flies were individually smoothed with a local linear smoother, but the subsequent data analysis for all 479 flies was based on the original data without smoothing.

Using the original egg-laying curves from day 6 to day 30 as the longitudinal covariates and the cohort indicator as a time-independent covariate, the functional Cox model resulted in an estimate $\hat{\theta} = 0.0562$ with 95% confidence interval $[-0.1235, 0.2359]$. Since zero is included in the interval, we conclude that the cohort effect is not significant. Figure 4 shows the estimated coefficient function $\hat{\beta}$ for the longitudinal covariate. The shaded area is the 95% pointwise bootstrap confidence interval. Under the functional Cox model, a positive $\hat{\beta}(s)$ yields a larger hazard function and a decreased probability of survival and vice versa for a negative $\hat{\beta}(s)$.

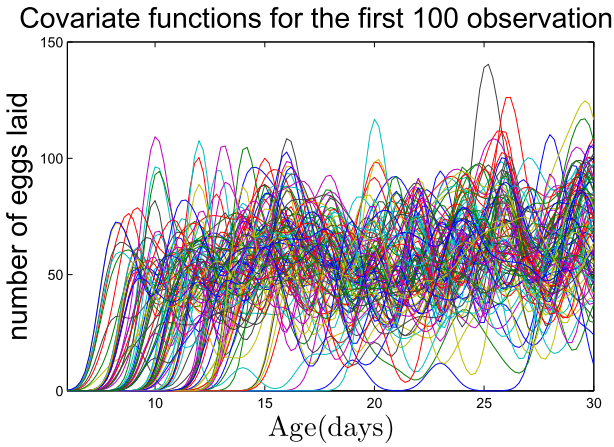

FIG. 3. *Pre-smoothed individual curves for the first* 100 *observations.*

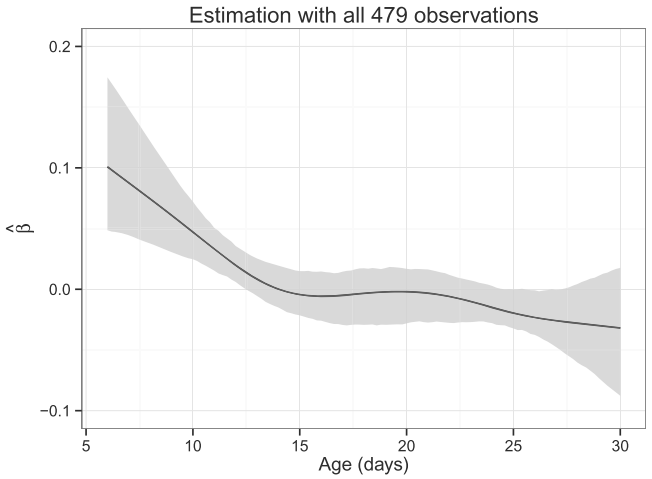

FIG. 4. *Estimated coefficient function* $\hat{\beta}(s)$ *using all* 479 *observations and* 95% *pointwise c.i. for* $\beta(s)$.

Checking the plot of $\hat{\beta}(s)$, we can see that $\hat{\beta}(s)$ starts with a large positive value, but decreases fast to near zero on day 13 and stays around zero until day 22, then declines again mildly towards day 30. The pattern of $\hat{\beta}(s)$ indicates that higher early reproduction before day 13 results in a much higher mortality rate suggesting the high cost of early reproduction, whereas a higher reproduction that occurs after day 22 tends to lead to a relatively lower mortality rate, suggesting that reproduction past day 22 might be sign of physical fitness. However, the latter effect is less significant than the early reproduction effect as indicated by the bootstrap confidence interval. Reproduction between day 13 and day 22 does not have a major effect on the mortality rate. In other words, flies that lay a lot of eggs in their early age (before day 13) and relatively fewer eggs after day 22 tend to die earlier, while those with the opposite pattern tend to have a longer life span.

The Mexfly data contains no censoring, so it is easy to check how the proposed method works in the presence of censored data. We artificially randomly censor the data by 10% and then again by 30% using an exponential censoring distribution with parameter $\gamma = 450$ and 150, respectively. See Table 3. The estimated coefficient $\hat{\theta}$ and corresponding 95% confidence intervals are given in Table 4. Regardless of the censoring conditions, all the confidence intervals contain zero and, therefore, indicate a insignificant cohort effect. This is consistent with the previous result for noncensored data. The estimated coefficient functions $\hat{\beta}$ and the corresponding pointwise bootstrap confidence intervals are displayed in Figure 5. Despite the slightly different results for different censoring proportions and choice of tuning parameters, all the $\hat{\beta}$ have a similar pattern. This indicates that

TABLE 3
Values of fixed cut-off point and parameters for generating random cut-off point, followed by the actual censored percentage for both cohorts and the whole data

	Fixed cut-off point		**Random cut-off point**	
	$T^c = 71$ **(10%)**	$T^c = 62$ **(30%)**	$T^c \sim \exp(450)$ **(10%)**	$T^c \sim \exp(150)$ **(30%)**
Cohort 1	0.138	0.339	0.071	0.353
Cohort 2	0.067	0.259	0.110	0.251
Total	0.100	0.296	0.092	0.300

the proposed method is quite stable with respect to right censorship, as long as the censoring rate is below 30%.

5. Proofs of theorems. We first introduce some notation by denoting $d(\beta_1, \beta_2) = \|\beta_1 - \beta_2\|_{C_\Delta}$, for any $\beta_1, \beta_2 \in \mathcal{H}(K)$; $Y(t) = 1_{\{T \geq t\}}$; $Y_j(t) = 1_{\{T_j \geq t\}}$, $1 \leq j \leq n$; and $\eta_\beta(X_i) = \int_0^1 \beta(s) X_i(s)\, ds$.

Recall that $W = (Z, X)$ represents the covariates, $\alpha = (\theta, \beta)$ represents the corresponding regression coefficient with θ the coefficient for Z and β the coefficient function for $X(\cdot)$, and the true coefficient is denoted as $\alpha_0 = (\theta_0, \beta_0)$. The index $\eta_\alpha(W) = \theta' Z + \int_0^1 \beta(s) X(s)\, ds$ summarizes the information carried by the covariate W. To measure the distance between two coefficients α_1 and α_2, we use

$$d(\alpha_1, \alpha_2)^2 = \mathbb{E}\big(\Delta\big[\eta_{\alpha_1}(W) - \eta_{\alpha_2}(W)\big]^2\big).$$

Furthermore, we denote

$$S_{0n}(t, \alpha) = \frac{1}{n} \sum_{j=1}^{n} Y_j(t) e^{\eta_\alpha(W_j)}, \qquad S_0(t, \alpha) = \mathbb{E}\big\{Y(t) e^{\eta_\alpha(W)}\big\},$$

TABLE 4
The estimated $\hat{\theta}$ and 95% confidence interval for θ under different censoring conditions

	10% censoring	**30% censoring**
Fixed cut-off point	0.0929 [−0.0914, 0.2772]	0.0757 [−0.1268, 0.2870]
Random cut-off point	0.0104 [−0.1705, 0.1913]	0.1863 [−0.0177, 0.3903]

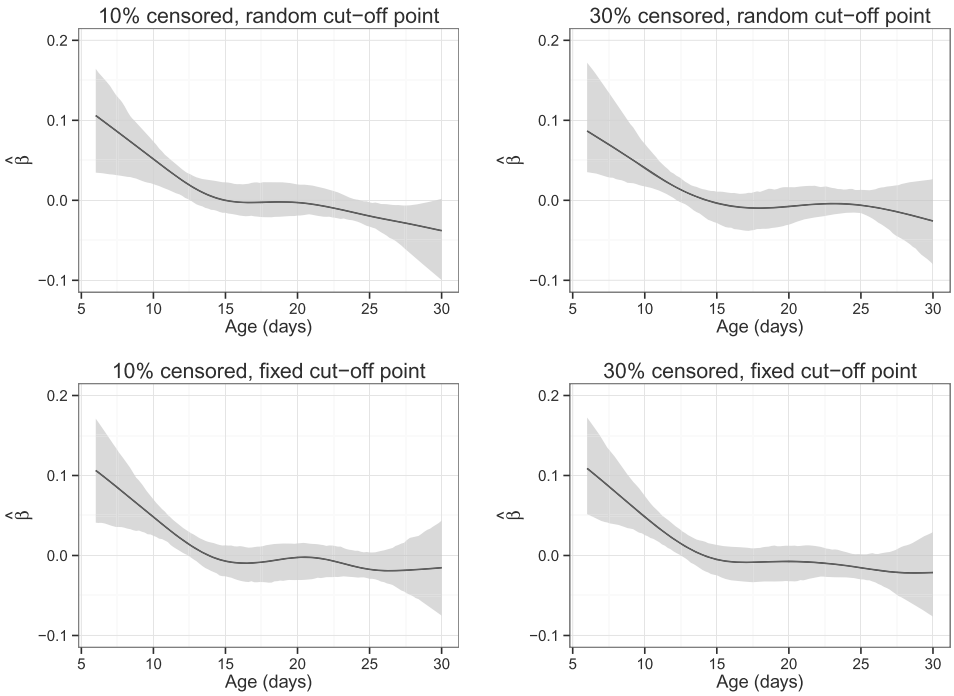

FIG. 5. *Estimation for $\beta(s)$ with censored data and* 95% *pointwise c.i.*

and for $\widetilde{\alpha} \in L_2 \times \mathcal{H}(K)$,

$$S_{1n}(t,\alpha)[\widetilde{\alpha}] = \frac{1}{n}\sum_{j=1}^{n} Y_j(t) e^{\eta_\alpha(W_j)} \eta_{\widetilde{\alpha}}(W_j),$$

$$S_1(t,\alpha)[\widetilde{\alpha}] = \mathbb{E}\big[Y(t) e^{\eta_\alpha(W)} \eta_{\widetilde{\alpha}}(W)\big].$$

Define

$$m_n(t,W,\alpha) = \big[\eta_\alpha(W) - \log S_{0n}(t,\alpha)\big] 1_{\{0\le t\le\tau\}},$$

and

$$m_0(t,W,\alpha) = \big[\eta_\alpha(W) - \log S_0(t,\alpha)\big] 1_{\{0\le t\le\tau\}}.$$

Let P_n and P be the empirical and probability measure of (T_i, Δ_i, W_i) and (T, Δ, W), respectively, and $P_{\Delta n}$ and P_Δ be the subprobability measure with $\Delta_i = 1$ and $\Delta = 1$ accordingly. The logarithm of the partial likelihood is $M_n(\alpha) = P_{\Delta n} m_n(\cdot,\alpha)$. Let $M_0(\alpha) = P_\Delta m_0(\cdot,\alpha)$. Note that P_Δ is restricted to $T \in [0,\tau]$ due to the $1\{0 \le t \le \tau\}$ term.

A useful identity due to Lemma 2 in Sasieni (1992b) is

$$\frac{S_1(t,\alpha)[\widetilde{\alpha}]}{S_0(t,\alpha)} = \mathbb{E}\big[\eta_{\widetilde{\alpha}}(W) | T = t, \Delta = 1\big]. \tag{5.1}$$

5.1. *Proof of Theorem 2.1.* The log-likelihood for a single sample $(t, \Delta, Z, X(\cdot))$ is

$$l(h_0, \theta, \beta) = \Delta\Big[\log h_0(t) + Z'\theta + \int_0^1 X(s)\beta(s)\,ds\Big] - H_0(t)\exp\Big[Z'\theta + \int_0^1 X(s)\beta(s)\,ds\Big],$$

where $H_0(t) = \int_0^t h_0(u)\,du$ is the baseline cumulative hazard function. Consider a parametric and smooth submodel $\{h_{(\mu_1)} : \mu_1 \in \mathbb{R}\}$ satisfying $h_{(0)} = h_0$ and

$$\frac{\partial \log h_{(\mu_1)}}{\partial \mu_1}(t)\bigg|_{\mu_1=0} = a(t).$$

Let $\eta_{(\mu_2)}(X) = \eta_\beta(X) + \eta_{\mu_2 g}(X)$, for $g \in \mathcal{H}(K)$. Therefore, $\eta_{(0)} = \eta_\beta(X)$ and

$$\frac{\partial \eta(\mu_2)}{\partial \mu_2}(X)\bigg|_{\mu_2=0} = \eta_g(X).$$

Recall that $r(W) = \exp(\eta_\alpha(W))$, and $M(t)$ is the counting process martingale associated with model (1),

$$M(t) = M(t|W) = \Delta I\{T \le t\} - \int_0^t I\{T \ge u\} r(W)\,dH_0(u).$$

The score operators for the cumulative hazard H_0, coefficient function β and the score vector for θ are the partial derivatives of the likelihood $l(h_{(\mu_1)}, \theta, \eta_{(\mu_2)})$ with respect to μ_1, μ_2 and θ evaluated at $\mu_1 = \mu_2 = 0$,

$$i_H a := \Delta a(T) - r(W)\int_0^\infty Y(t)a(t)\,dH_0(t) = \int_0^\infty a(t)\,dM(t),$$

$$i_\beta g := \eta_g(X)\big[\Delta - r(W)H_0(T)\big] = \int_0^\infty \eta_g(X)\,dM(t),$$

$$i_\theta := Z\big[\Delta - r(W)H_0(T)\big] = \int_0^\infty Z\,dM(t).$$

Define $L(P_T^{(u)}) := \{a \in \mathcal{L}_2 : \mathbb{E}[\Delta a^2(T)] < \infty\}$ and $L(P_X^{(u)}) := \{g \in \mathcal{H}(K) : \mathbb{E}[\Delta \eta_g(X)] = 0; \mathbb{E}[\Delta \eta_g^2(X) < \infty]\}$. Let

$$A_H = \big\{i_H a : a \in L\big(P_T^{(u)}\big)\big\},$$

and

$$G = \big\{i_\beta g : g \in L\big(P_X^{(u)}\big)\big\}.$$

To calculate the information bound for θ, we need to find the (least favorable) direction (a^*, g^*) such that $i_\theta - i_H a^* - i_\beta g^*$ is orthogonal to the sum space $\mathbf{A} =$

$A_H + G$. That is, (a^*, g^*) must satisfy

$$\mathbb{E}\big[(i_\theta - i_H a^* - i_\beta g^*) i_H a\big] = 0, \qquad a \in L\big(P_T^{(u)}\big),$$

$$\mathbb{E}\big[(i_\theta - i_H a^* - i_\beta g^*) i_\beta g\big] = 0, \qquad g \in L\big(P_X^{(u)}\big).$$

Following the proof of Theorem 3.1 in Huang (1999), we can show that (a^*, g^*) satisfies

$$\mathbb{E}\big[\Delta(Z - a^* - \eta_{g^*})a\big] = 0, \qquad a \in L\big(P_T^{(u)}\big), \tag{5.2}$$

$$\mathbb{E}\big[\Delta(Z - a^* - \eta_{g^*})\eta_g\big] = 0, \qquad g \in L\big(P_X^{(u)}\big). \tag{5.3}$$

Therefore, (a^*, g^*) is the solution to the following equations:

$$\mathbb{E}\big(Z - a^* - \eta_{g^*} | T, \Delta = 1\big) = 0, \qquad \text{a.s. } P_T^{(u)},$$

$$\mathbb{E}\big(Z - a^* - \eta_{g^*} | X, \Delta = 1\big) = 0, \qquad \text{a.s. } P_X^{(u)}.$$

So, $(a^*, g^*) \in L(P_T^{(u)}) \times L(P_X^{(u)})$ minimizes

$$\mathbb{E}\big\{\Delta \big\| Z - a(T) - \eta_g(X)\big\|^2\big\}. \tag{5.4}$$

It follows from Conditions A3 and A4 that the space $L(P_T^{(u)}) \times L(P_X^{(u)})$ is closed, so that the minimizer of (5.4) is well defined. Further, the solution can be obtained by the population version of the ACE algorithm of Breiman and Friedman (1985).

5.2. *Proof of Theorem 2.2.* For some large number M, such that $\|\theta_0\|_\infty < M$ and $\|\beta_0\|_K < M$, define $\mathbb{R}_M = \{\theta \in \mathbb{R}^p, \|\theta\|_\infty < M\}$ and $\mathcal{H}^M = \{\beta \in \mathcal{H}(K), \|\beta\|_K < M\}$. Let $\alpha^M = (\theta^M, \beta^M)$ be the penalized partial likelihood estimator with minimum taken over $L^M \times \mathcal{H}^M$, that is,

$$\begin{aligned}\alpha^M = \mathop{\arg\min}_{\alpha \in \mathbb{R}_M \times \mathcal{H}^M} & -n^{-1} \sum_{i=1}^n \Delta_i \Big\{ \eta_\alpha(W_i) - \log \sum_{T_j > T_i} \exp\{\eta_\alpha(W_j)\} \Big\} \\ & + \lambda \cdot J(\beta).\end{aligned} \tag{5.5}$$

We first prove that

$$\sup_{\alpha \in \mathbb{R}_M \times \mathcal{H}^M} \big| M_n(\alpha) - M_0(\alpha) \big| \xrightarrow{P} 0. \tag{5.6}$$

Observe that

$$\begin{aligned}&\big| M_n(\alpha) - M_0(\alpha) \big| \\ &\quad \le \big| P_{\Delta n} m_n(\cdot, \alpha) - P_{\Delta n} m_0(\cdot, \alpha) \big| + \big| P_{\Delta n} m_0(\cdot, \alpha) - P_\Delta m_0(\cdot, \alpha) \big| \\ &\quad \le P_{\Delta n} \big| \log S_{0n}(T, \alpha) - \log S_0(T, \alpha) \big| 1_{\{0 \le T \le \tau\}} + \big| (P_n - P) \Delta m_0(\cdot, \alpha) \big| \\ &\quad \lesssim \sup_{0 \le t \le \tau} \big| S_{0n}(t, \alpha) - S_0(t, \alpha) \big| + \big| (P_n - P) \Delta m_0(\cdot, \alpha) \big| \\ &\quad = \sup_{0 \le t \le \tau} \big| (P_n - P) Y(t) e^{\eta_\alpha(W)} \big| + \big| (P_n - P) \Delta m_0(\cdot, \alpha) \big|.\end{aligned}$$

Lemma 3 shows that $\mathcal{F}_1 = \{\Delta m_0(t, W, \alpha) : \alpha \in \mathbb{R}_M \times \mathcal{H}^M\}$ and $\mathcal{F}_2 = \{Y(t) \times e^{\eta_\alpha(W)} : \alpha \in \mathbb{R}_M \times \mathcal{H}^M, 0 \le t \le \tau\}$ are P-Glivenko–Cantelli, which means that both terms on the right-hand side above converge to zero in probability uniformly with respect to $\alpha \in \mathbb{R}_M \times \mathcal{H}^M$. Therefore, (5.6) holds.

The definition of α^M in (5.5) indicates that

$$-M_n(\alpha^M) + \lambda J(\beta^M) \le -M_n(\alpha_0) + \lambda J(\beta_0).$$

Rearranging the inequality with $M_n(\alpha^M)$ on one side and the fact that $\lambda \to 0$ as $n \to \infty$ lead to

$$M_n(\alpha^M) \ge M_n(\alpha_0) - o_p(1). \tag{5.7}$$

On the other hand, Lemma 2 implies that $\sup_{d(\alpha,\alpha_0)\ge\epsilon} M_0(\alpha) < M_0(\alpha_0)$. Combining this with (5.6) and (5.7) and by the consistency result in van der Vaart (2000), Theorem 5.7 on page 45, we can show that α^M is consistent, that is, $d(\alpha^M, \alpha_0) \xrightarrow{P} 0$.

Part (i) now follows from

$$d(\hat\alpha, \alpha_0) \le d(\hat\alpha, \alpha^M) + d(\alpha^M, \alpha_0),$$

and $P(\hat\alpha = \alpha^M) = P(\|\hat\beta\|_K < M, \|\hat\theta\|_\infty < M) \to 1$, as $M \to \infty$, that is, $d(\hat\alpha, \alpha^M) \to 0$ a.s.

For part (ii), we follow the proof of Theorem 3.4.1 in van der Vaart and Wellner (1996). We first show that

$$E^* \sup_{\delta/2 \le d(\alpha,\alpha_0) \le \delta} \sqrt{n}\,|(M_n - M_0)(\alpha - \alpha_0)| \lesssim \phi_n(\delta), \tag{5.8}$$

where $\phi_n(\delta) = \delta^{(2r-1)/(2r)}$. Direct calculation yields that

$$\begin{aligned}
&(M_n - M_0)(\alpha - \alpha_0)\\
&\quad = P_{\Delta n} m_n(\cdot, \alpha) - P_{\Delta n} m_n(\cdot, \alpha_0) - P_\Delta m_0(\cdot, \alpha) + P_\Delta m_0(\cdot, \alpha_0)\\
&\quad = (P_{\Delta n} - P_\Delta)\big(m_0(\cdot, \alpha) - m_0(\cdot, \alpha_0)\big)\\
&\qquad + P_{\Delta n}\big(m_n(\cdot, \alpha) - m_n(\cdot, \alpha_0) - m_0(\cdot, \alpha) + m_0(\cdot, \alpha_0)\big)\\
&\quad = (P_{\Delta n} - P_\Delta)\big(m_0(\cdot, \alpha) - m_0(\cdot, \alpha_0)\big)\\
&\qquad + P_{\Delta n}\left(\log \frac{S_0(T, \alpha)}{S_0(T, \alpha_0)} - \log \frac{S_{0n}(T, \alpha)}{S_{0n}(T, \alpha_0)}\right)\\
&\quad = I + II.
\end{aligned}$$

For the first term, $I = (P_{\Delta n} - P_\Delta)(m_0(\cdot, \alpha) - m_0(\cdot, \alpha_0))$. By Lemma 4, we have

$$\sup_{\delta/2 \le d(\alpha,\alpha_0) \le \delta} |I| = O\big(\delta^{(2r-1)/(2r)} n^{-1/2}\big).$$

For the second term II, we have

$$\begin{aligned}
&\sup_{\delta/2\le d(\alpha,\alpha_0)\le\delta} |II| \\
&\quad\le \sup_{\substack{\delta/2\le d(\alpha,\alpha_0)\le\delta \\ t\in[0,\tau]}} \left|\log\frac{S_0(t,\alpha)}{S_0(t,\alpha_0)} - \log\frac{S_{0n}(t,\alpha)}{S_{0n}(t,\alpha_0)}\right| \\
&\quad\le \sup_{\substack{\delta/2\le d(\alpha,\alpha_0)\le\delta \\ t\in[0,\tau]}} c\left|\frac{S_{0n}(t,\alpha)}{S_{0n}(t,\alpha_0)} - \frac{S_0(t,\alpha)}{S_0(t,\alpha_0)}\right| \\
&\quad= \sup_{\substack{\delta/2\le d(\alpha,\alpha_0)\le\delta \\ t\in[0,\tau]}} c\left|\frac{S_{0n}(t,\alpha)S_0(t,\alpha_0) - S_{0n}(t,\alpha_0)S_0(t,\alpha)}{S_0(t,\alpha_0)S_{0n}(t,\alpha_0)}\right|.
\end{aligned}$$

For $t\in[0,\tau]$, the denominator $S_0(t,\alpha_0)S_{0n}(t,\alpha_0)$ is bounded away from zero with probability tending to one. The numerator satisfies

$$\begin{aligned}
&S_{0n}(t,\alpha)S_0(t,\alpha_0) - S_{0n}(t,\alpha_0)S_0(t,\alpha) \\
&\quad= S_0(t,\alpha_0)\big[S_{0n}(t,\alpha) - S_{0n}(t,\alpha_0) - S_0(t,\alpha) + S_0(t,\alpha_0)\big] \\
&\qquad - \big[S_{0n}(t,\alpha_0) - S_0(t,\alpha_0)\big]\big[S_0(t,\alpha) - S_0(t,\alpha_0)\big].
\end{aligned}$$

For the first term on the right-hand side, we have $S_0(t,\alpha_0) = O(1)$ and

$$\begin{aligned}
&\big[S_{0n}(t,\alpha) - S_{0n}(t,\alpha_0) - S_0(t,\alpha) + S_0(t,\alpha_0)\big] \\
&\quad= (P_n - P)\big\{Y(t)\big[\exp\big(\eta_\alpha(W)\big) - \exp\big(\eta_{\alpha_0}(W)\big)\big]\big\}.
\end{aligned}$$

Define the above $(P_n - P)\{Y(t)[\exp(\eta_\alpha(W)) - \exp(\eta_{\alpha_0}(W))]\} \stackrel{\text{def}}{=} III$.
Lemma 4 implies that

$$\sup_{\delta/2\le d(\alpha,\alpha_0)\le\delta} |III| = O\big(\delta^{(2r-1)/(2r)}n^{-1/2}\big).$$

For the second term, the central limit theorem implies $S_{0n}(t,\alpha_0) - S_0(t,\alpha_0) = O_p(n^{-1/2})$, and

$$\begin{aligned}
\big|S_0(t,\alpha) - S_0(t,\alpha_0)\big| &\le E\big\{Y(t)\big|\exp\big(\eta_\alpha(W)\big) - \exp\big(\eta_{\alpha_0}(W)\big)\big|\big\} \\
&\lesssim \big(E\big[\eta_\alpha(W) - \eta_{\alpha_0}(W)\big]^2\big)^{1/2} \\
&\lesssim d(\alpha,\alpha_0).
\end{aligned}$$

Therefore,

$$\sup_{\delta/2\le d(\alpha,\alpha_0)\le\delta} |II| \le O\big(\delta^{(2r-1)/(2r)}n^{-1/2}\big) + O\big(\delta n^{-1/2}\big) = O\big(\delta^{(2r-1)/(2r)}n^{-1/2}\big).$$

Combining I and II yields

$$E^* \sup_{\delta/2\le d(\alpha,\alpha_0))\le\delta} \sqrt{n}\big|(M_n - M_0)(\alpha-\alpha_0)\big| \lesssim O\big(\delta^{(2r-1)/(2r)}\big).$$

Furthermore, Lemma 2 implies

$$\sup_{\delta/2\le d(\alpha,\alpha_0)\le\delta} P_\Delta m_0(\cdot,\alpha) - P_\Delta m_0(\cdot,\alpha_0) \lesssim -\delta^2.$$

Let $r_n = n^{r/(2r+1)}$. It is easy to check that r_n satisfies $r_n^2\phi_n(\frac{1}{r_n}) \le \sqrt{n}$, and

$$M_n(\hat\alpha_\lambda) \ge M_n(\alpha_0) + \lambda\big[J(\hat\beta_\lambda) - J(\beta_0)\big] \ge M_n(\alpha_0) - O_p\big(r_n^{-2}\big)$$

with $\lambda = O(r_n^{-2}) = O(n^{-2r/(2r+1)})$.

So far, we have verified all the conditions in Theorem 3.4.1 of van der Vaart and Wellner (1996), and thus conclude that

$$d(\hat\alpha,\alpha_0) = O_p\big(r_n^{-1}\big) = O_p\big(n^{-r/(2r+1)}\big).$$

For part (iii), recall the projections a^* and g^* defined in Theorem 2.1, then

$$\begin{aligned}
d(\hat\alpha,\alpha_0)^2 &= \mathbb{E}\Delta\big[\eta_{\hat\alpha}(W) - \eta_{\alpha_0}(W)\big]^2 \\
&= \mathbb{E}\Delta\big[Z'(\hat\theta-\theta_0) + \big(\eta_{\hat\beta}(X) - \eta_{\beta_0}(X)\big)\big]^2 \\
&= \mathbb{E}\Delta\big[\big(Z - a^*(T) - \eta_{g^*}(X)\big)'(\hat\theta-\theta_0) + \big(a^*(T) + \eta_{g^*}(X)\big)(\hat\theta-\theta_0) \\
&\quad + \big(\eta_{\hat\beta}(X) - \eta_{\beta_0}(X)\big)\big]^2 \\
&= \mathbb{E}\Delta\big[\big(Z - a^*(T) - \eta_{g^*}(X)\big)'(\hat\theta-\theta_0)\big]^2 \\
&\quad + \mathbb{E}\Delta\big[\big(a^*(T) + \eta_{g^*}(X)\big)(\hat\theta-\theta_0) + \big(\eta_{\hat\beta}(X) - \eta_{\beta_0}(X)\big)\big]^2.
\end{aligned} \tag{5.9}$$

Since $I(\theta)$ is nonsingular, it follows that $\|\hat\theta-\theta_0\|^2 = O_p(n^{-2r/(2r+1)})$. This in turn implies

$$d(\hat\beta,\beta_0)^2 = O_p\big(n^{-2r/(2r+1)}\big).$$

5.3. *Proof of Theorem 2.3.* Let $u = (t, Z, X(\cdot))$. For $g \in \mathcal{H}(K)$, define

$$s_n(u,\alpha)[g] = \eta_g(X) - \frac{S_{1n}(t,\alpha)[g]}{S_{0n}(t,\alpha)}, \qquad s(u,\alpha)[g] = \eta_g(X) - \frac{S_1(t,\alpha)[g]}{S_0(t,\alpha)},$$

and for $Z \in \mathbb{R}^d$ and the identify map $I(Z) = Z$, define

$$s_n(u,\alpha)[Z] = Z - \frac{S_{1n}(t,\alpha)[I]}{S_{0n}(t,\alpha)}, \qquad s(u,\alpha)[Z] = \eta_g(X) - \frac{S_1(t,\alpha)[I]}{S_0(t,\alpha)},$$

where $S_{1n}(t,\alpha)[I] = \frac{1}{n}\sum_{j=1}^n Y_j(t)e^{\eta_\alpha(W_j)}Z_j$ and $S_1(t,\alpha)[I] = \mathbb{E}Y(t)e^{\eta_\alpha(W)}Z$.

By analogy to the score function, we call the derivatives of the partial likelihood with respect to the parameters the partial score functions. The partial score function based on the partial likelihood for θ is

$$i_{n\theta}(\alpha) = P_{\Delta n} s_n(\cdot, \alpha)[Z].$$

The partial score function based on the partial likelihood for β in a direction $g \in \mathcal{H}(K)$ is

$$i_{n\beta}(\alpha)[g] = P_{\Delta n} s_n(\cdot, \alpha)[g].$$

Recall that $(\hat{\theta}, \hat{\beta})$ is defined to maximize the penalized partial likelihood, that is,

$$-P_{\Delta n} m_n(\cdot, \hat{\theta}, \hat{\beta}) + \lambda J(\hat{\beta}) \le -P_{\Delta n} m_n(\cdot, \theta, \beta) + \lambda J(\beta),$$

for all $\theta \in \mathbb{R}^p$ and $\beta \in \mathcal{H}(K)$. Since the penalty term is unrelated to θ, the partial score function should satisfy

$$i_{n\theta}(\hat{\alpha}) = P_{\Delta n} s_n(\cdot, \hat{\alpha})[Z] = 0.$$

On the other hand, the partial score function for β satisfies

$$i_{n\beta}(\hat{\alpha})[g] = P_{\Delta n} s_n(\cdot, \alpha)[g] = O(\lambda) = o_p\big(n^{-1/2}\big), \qquad \text{for all } g \in \mathcal{H}(K).$$

Combining this with Lemma 5 and Lemma 6, we have

$$n^{1/2} P_\Delta \big\{ s(\cdot, g_0)\big[Z - h^*\big] \big\}^{\otimes 2} (\hat{\theta} - \theta_0) = -n^{1/2} P_{\Delta n} s_n(\cdot, \alpha_0)\big[Z - g^*\big] + o_p(1).$$

Let

$$M_i(t) = \Delta_i I\{T_i \le t\} - \int_0^t Y_i(u) \exp\big(\eta_{\alpha_0}(W_i)\big)\, dH_0(u), \qquad 1 \le i \le n.$$

We can write

$$\begin{aligned} &n^{1/2} P_{\Delta n} s_n(\cdot, \alpha_0)\big[Z - g^*\big] \\ &\quad = n^{-1/2} \sum_{i=1}^n \int_0^\tau \left[Z_i - \eta_{h^*}(X_i) - \frac{S_{1n}(t, \alpha_0)[Z - g^*]}{S_{0n}(t, \alpha_0)} \right] dM_i(t). \end{aligned}$$

Thus,

$$\begin{aligned} &n^{1/2} P_{\Delta n} s_n(\cdot, \alpha_0)\big[Z - g^*\big] - n^{-1/2} \sum_{i=1}^n \int_0^\tau \left[Z_i - \eta_{h^*}(X_i) \right. \\ &\qquad \left. - \frac{S_1(t, \alpha_0)[Z - g^*]}{S_0(t, \alpha_0)} \right] dM_i(t) \\ &\quad = n^{-1/2} \sum_{i=1}^n \int_0^\tau \left[\frac{S_1(t, \alpha_0)[Z - g^*]}{S_0(t, \alpha_0)} - \frac{S_{1n}(t, \alpha_0)[Z - g^*]}{S_{0n}(t, \alpha_0)} \right] dM_i(t). \end{aligned}$$

Because

$$n^{-1}\sum_{i=1}^{n}\int_0^\tau\left[\frac{S_1(t,\alpha_0)[Z-g^*]}{S_0(t,\alpha_0)}-\frac{S_{1n}(t,\alpha_0)[Z-g^*]}{S_{0n}(t,\alpha_0)}\right]Y_i(t)$$
$$\times\exp\bigl[\eta_{\alpha_0}(W_i)\bigr]\,dH_i(t)\xrightarrow{P}0,$$

by Lenglart's inequality, as stated in Theorem 3.4.1 and Corollary 3.4.1 of Fleming and Harrington (1991), we have

$$n^{1/2}P_{\Delta n}s_n(\cdot,\alpha_0)\bigl[Z-g^*\bigr]$$
$$=n^{-1/2}\sum_{i=1}^{n}\int_0^\tau\left[Z_i-\eta_{h^*}(X_i)-\frac{S_1(t,\alpha_0)[Z-g^*]}{S_0(t,\alpha_0)}\right]dM_i(t)+o_p(1).$$

Recall that

$$\frac{S_1(t,\alpha_0)[Z-g^*]}{S_0(t,\alpha_0)}=E\bigl[\bigl[Z-\eta_{g^*}(W)\bigr]|T=t,\Delta=1\bigr]=a^*(t).$$

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

$$n^{1/2}P_{\Delta n}s_n(\cdot,\alpha_0)\bigl[Z-g^*\bigr]=n^{-1/2}\sum_{i=1}^{n}l_\theta^*(T_i,\Delta_i,W_i)+o_p(1)\xrightarrow{d}\mathcal{N}\bigl(0,I(\theta_0)\bigr).$$

5.4. *Proof of Theorem 2.4.* To get the minmax lower bound, it suffices to show that, when the true baseline hazard function h_0 and the true θ_0 are fixed and known, for a subset $\mathcal{H}^*$ of $\mathcal{H}(K)$,

$$\lim_{a\to0}\lim_{n\to\infty}\inf_{\hat\beta}\sup_{\beta_0\in\mathcal{H}^*}\mathbb{P}_{h_0,\theta_0,\beta_0}\bigl\{d(\hat\beta,\beta_0)\geq an^{-r/(2r+1)}\bigr\}=1. \tag{5.10}$$

If we can find a subset $\{\beta^{(0)},\ldots,\beta^{(N)}\}\subset\mathcal{H}^*$ with N increasing with n, such that for some positive constant c and all $0\leq i<j\leq N$,

$$d^2\bigl(\beta^{(i)},\beta^{(j)}\bigr)\geq c\gamma^{2r/(2r+1)}n^{-2r/(2r+1)}, \tag{5.11}$$

and

$$\frac{1}{N}\sum_{j=1}^{N}KL(P_j,P_0)\leq\gamma\log N, \tag{5.12}$$

then we can conclude, according to Tsybakov (2009), Theorem 2.5 on page 99, that

$$\inf_{\hat\beta}\sup_{\beta\in\mathcal{H}^*}\mathbb{P}\bigl(d^2\bigl(\beta^{(i)},\beta^{(j)}\bigr)\geq c\gamma^{2r/(2r+1)}n^{-2r/(2r+1)}\bigr)$$
$$\geq\frac{\sqrt{N}}{1+\sqrt{N}}\biggl(1-2\gamma-\sqrt{\frac{2\gamma}{\log N}}\biggr),$$

which yields

$$\lim_{a\to 0}\lim_{n\to\infty}\inf_{\hat{\beta}}\sup_{\beta_0\in\mathcal{H}^*}\mathbb{P}\big(d\big(\beta^{(i)},\beta^{(j)}\big)\geq an^{-r/(2r+1)}\big)\geq 1.$$

Hence, Theorem 2.4 will be proved.

Next, we are going to construct the set $\mathcal{H}^*$ and the subset $\{\beta^{(0)},\ldots,\beta^{(N)}\}\subset\mathcal{H}^*$, and then show that both (5.11) and (5.12) are satisfied.

Consider the function space

$$\mathcal{H}^*=\left\{\beta=\sum_{k=M+1}^{2M}b_kM^{-1/2}L_{K^{1/2}}\varphi_k:(b_{M+1},\ldots,b_{2M})\in\{0,1\}^M\right\},\tag{5.13}$$

where $\{\varphi_k:k\geq 1\}$ are the orthonomal eigenfunctions of $T(s,t)=K^{1/2}C_\Delta\times K^{1/2}(s,t)$ and M is some large number to be decided later.

For any $\beta\in\mathcal{H}^*$, observe that

$$\begin{aligned}\|\beta\|_K^2&=\left\|\sum_{k=M+1}^{2M}b_kM^{-1/2}L_{K^{1/2}}\varphi_k\right\|_K^2\\&=\sum_{k=M+1}^{2M}b_k^2M^{-1}\|L_{K^{1/2}}\varphi_k\|_K^2\\&\leq\sum_{k=M+1}^{2M}M^{-1}\|L_{K^{1/2}}\varphi_k\|_K^2\\&=1,\end{aligned}$$

which follows from the fact that

$$\langle L_{K^{1/2}}\varphi_k,L_{K^{1/2}}\varphi_l\rangle_K=\langle L_K\varphi_k,\varphi_l\rangle_K=\langle\varphi_k,\varphi_l\rangle_{L_2}=\delta_{kl}.$$

Therefore, $\mathcal{H}^*\subset\mathcal{H}(K)=\{\beta:\|\beta\|_k<\infty\}$.

The Varshamov–Gilbert bound shows that for any $M\geq 8$, there exists a set $\mathcal{B}=\{b^{(0)},b^{(1)},\ldots,b^{(N)}\}\subset\{0,1\}^M$ such that:

1. $b^{(0)}=(0,\ldots,0)'$;
2. $H(b,b')>M/8$ for any $b\neq b'\in\mathcal{B}$, where $H(\cdot,\cdot)=\frac{1}{4}\sum_{i=1}^M(b_i-b_i')^2$ is the Hamming distance;
3. $N\geq 2^{M/8}$.

The subset $\{\beta^{(0)},\ldots,\beta^{(N)}\}\subset\mathcal{H}^*$ is chosen as $\beta^{(i)}=\sum_{k=M+1}^{2M}b_{k-M}^{(i)}M^{-1/2}\times L_{K^{1/2}}\varphi_k$, $i=0,\ldots,N$.

For any $0 \le i < j \le N$, observe that

$$\begin{aligned} d^2\big(\beta^{(i)}, \beta^{(j)}\big) &= \mathbb{E}\Delta\big(\eta_{\beta^{(i)}}(X) - \eta_{\beta^{(j)}}(X)\big)^2 \\ &= \Big\| L_{C_\Delta^{1/2}} \sum_{k=M+1}^{2M} \big(b_{k-M}^{(i)} - b_{k-M}^{(j)}\big) M^{-1/2} L_{K^{1/2}} \varphi_k \Big\|_{L_2}^2 \\ &= \sum_{k=M+1}^{2M} \big(b_{k-M}^{(i)} - b_{k-M}^{(j)}\big)^2 M^{-1} \big\| L_{C_\Delta^{1/2}} L_{K^{1/2}} \varphi_k \big\|_{L_2}^2 \\ &= \sum_{k=M+1}^{2M} \big(b_{k-M}^{(i)} - b_{k-M}^{(j)}\big)^2 M^{-1} s_k. \end{aligned}$$

On one hand, we have

$$\begin{aligned} d^2\big(\beta^{(i)}, \beta^{(j)}\big) &= \sum_{k=M+1}^{2M} \big(b_{k-M}^{(i)} - b_{k-M}^{(j)}\big)^2 M^{-1} s_k \\ &\ge s_{2M} M^{-1} \sum_{k=1}^{M} \big(b_k^{(i)} - b_k^{(j)}\big)^2 \\ &= 4 s_{2M} M^{-1} H\big(b^{(i)}, b^{(j)}\big) \\ &\ge s_{2M}/2. \end{aligned}$$

On the other hand, we have

$$\begin{aligned} d^2\big(\beta^{(i)}, \beta^{(j)}\big) &= \sum_{k=M+1}^{2M} \big(b_{k-M}^{(i)} - b_{k-M}^{(j)}\big)^2 M^{-1} s_k \\ &\le s_M M^{-1} \sum_{k=1}^{M} \big(b_k^{(i)} - b_k^{(j)}\big)^2 \\ &\le s_M. \end{aligned}$$

So altogether,

$$s_{2M}/2 \le d^2\big(\beta^{(i)}, \beta^{(j)}\big) \le s_M. \tag{5.14}$$

Let P_j, $j = 1, \ldots, N$, be the likelihood function with data $\{(T_i, \Delta_i, W_i(s)), i = 1, \ldots, n\}$ and $\beta^{(j)}$, that is,

$$P_j = \prod_{i=1}^{n} \big[f_{T^u|W}(T_i) S_{T^c|W}(T_i) \big]^{\Delta_i} \cdot \big[f_{T^c|W}(T_i) S_{T^u|W}(T_i) \big]^{1-\Delta_i}.$$

Let $c_{T^c} = \prod_{i=1}^n [S_{T^c|W}(T_i)]^{\Delta_i} [S_{T^u|W}(T_i)]^{1-\Delta_i}$, which does not depend on $\beta^{(j)}$, then

$$P_j = c_{T^c} \prod_{i=1}^n \big[h_0(T_i) \exp\big(\theta_0' Z_i + \eta_{\beta^{(j)}}(X_i)\big)\big]^{\Delta_i} \cdot \exp\big\{-H_0(T_i) \cdot e^{\theta_0' Z_i + \eta_{\beta^{(j)}}(X_i)}\big\}.$$

We calculate the Kullback–Leibler distance between P_j and P_0 as

$$\begin{aligned}
KL(P_j, P_0) &= \mathbb{E}_{P_j} \log \frac{P_j}{P_0} \\
&= \mathbb{E}_{P_j} \Bigg\{ \Delta_i \sum_{i=1}^n \big\{\eta_{\beta^{(j)} - \beta^{(0)}}(X_i)\big\} \\
&\qquad + \sum_{i=1}^n H_0(T_i) e^{\theta_0' Z_i} \big[\exp\big(\eta_{\beta^{(0)}}(X_i)\big) - \exp\big(\eta_{\beta^{(j)}}(X_i)\big)\big] \Bigg\} \\
&= n \mathbb{E}_{P_j} \Delta \big[\eta_{\beta^{(j)} - \beta^{(0)}}(X)\big] \\
&\qquad + n \mathbb{E}_{P_j} H_0(T) e^{\theta_0' Z} \big[\exp\big(\eta_{\beta^{(0)}}(X)\big) - \exp\big(\eta_{\beta^{(j)}}(X)\big)\big] \\
&= n \mathbb{E}_{P_j}^W \mathbb{E}_{P_j}^{T,\Delta} \big\{H_0(T) | W\big\} e^{\theta_0' Z} \big[\exp\big(\eta_{\beta^{(0)}}(X)\big) - \exp\big(\eta_{\beta^{(j)}}(X)\big)\big],
\end{aligned}$$

where

$$\begin{aligned}
&\mathbb{E}_{p_j}^{T,\Delta}\big(H_0(T)|W\big) \\
&\quad = \mathbb{E}^{T^c}\big\{\mathbb{E}_{p_j}^{T,\Delta}\big(H_0(T)|T^c, W\big)\big|W\big\} \\
&\quad = \mathbb{E}^{T^c}\Big\{\int_0^{T^c} H_0(t) f_{T^u|W}(t)\, dt + H_0(T^c)\mathbb{P}\big(T^u > T^c|T^c, W\big)\big|W\Big\}, \\
&\int_0^{T^c} H_0(t) f_{T^u|W}(t)\, dt \\
&\quad = \int_0^{T^c} H_0(t) \cdot h_0(t) \exp\big[\theta_0' Z + \eta_{\beta^{(j)}}(X)\big] \exp\big\{-H_0(t) e^{\theta_0' Z + \eta_{\beta^{(j)}}(X)}\big\}\, dt \\
&\quad = e^{-\theta_0' Z - \eta_{\beta^{(j)}}(X)} \int_0^{T^c} e^{\theta_0' Z + \eta_{\beta^{(j)}}(X)} H_0(T) \\
&\qquad \times \exp\big\{-H_0(T) \cdot e^{\theta_0' Z + \eta_{\beta^{(j)}}(X)}\big\}\, d e^{\theta_0' Z + \eta_{\beta^{(j)}}(X)} H_0(T) \\
&\quad = \exp\big(-\theta_0' Z + \eta_{\beta^{(j)}}(X)\big) \int_0^a u e^{-u}\, du \Big|_{a = e^{\theta_0' Z + \eta_{\beta^{(j)}}(X)} H_0(T^c)} \\
&\quad = \exp\big(-\theta_0' Z - \eta_{\beta^{(j)}}(X)\big)\big[1 - e^{-a} - a e^{-a}\big]\big|_{a = e^{\theta_0' Z + \eta_{\beta^{(j)}}(X)} H_0(T^c)},
\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}\big(T^u > T^c | T^c, W\big) &= S_{T^u|W}\big(T^c\big)\\ &= \exp\big\{-H_0\big(T^c\big)e^{\theta_0' Z+\eta_{\beta^{(j)}}(X)}\big\}.\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{E}_{p_j}^{T,\Delta}\big(H_0(T)|T^c, W\big)\\ &\quad= e^{-\theta_0' Z-\eta_{\beta^{(j)}}(X)}\big[1-\exp\big\{-H_0\big(T^c\big)e^{\theta_0' Z+\eta_{\beta^{(j)}}(X)}\big\}\big] - H_0\big(T^c\big)\\ &\qquad\times\exp\big\{-H_0\big(T^c\big)e^{\theta_0' Z+\eta_{\beta^{(j)}}(X)}\big\} + H_0\big(T^c\big)\exp\big\{-H_0\big(T^c\big)e^{\theta_0' Z+\eta_{\beta^{(j)}}(X)}\big\}\\ &\quad= e^{-\theta_0' Z-\eta_{\beta^{(j)}}(X)}\big[1-\exp\big\{-H_0\big(T^c\big)e^{\theta_0' Z+\eta_{\beta^{(j)}}(X)}\big\}\big]\\ &\quad= e^{-\theta_0' Z-\eta_{\beta^{(j)}}(X)}\big[F_{T^u|W}\big(T^c\big)\big]\\ &\quad= e^{-\theta_0' Z-\eta_{\beta^{(j)}}(X)}\mathbb{P}\big(T^u \le T^c | T^c, W\big),\end{aligned}$$

and further

$$\begin{aligned}\mathbb{E}_{p_j}^{T,\Delta}\big(H_0(T)|W\big) &= \mathbb{E}^{T^c}\big\{\mathbb{E}_{p_j}^{T,\Delta}\big(H_0(T)|T^c, W\big)|W\big\}\\ &= e^{-\theta_0' Z-\eta_{\beta^{(j)}}(X)}\mathbb{P}\big(T^u \le T^c|W\big)\\ &= e^{-\theta_0' Z-\eta_{\beta^{(j)}}(X)}\mathbb{E}[\Delta|W].\end{aligned}$$

Then the KL distance becomes

$$\begin{aligned}KL(P_j, P_0) &= n\mathbb{E}_{P_j}^{W}\mathbb{E}[\Delta|W]e^{-\theta_0' Z-\eta_{\beta^{(j)}}(X)}e^{\theta_0' Z}\big[\exp\big(\eta_{\beta^{(0)}}(X)\big) - \exp\big(\eta_{\beta^{(j)}}(X)\big)\big]\\ &= n\mathbb{E}_{P_j}^{W,\Delta}\Delta\big[\exp\big(\eta_{\beta^{(0)}}(X) - \eta_{\beta^{(j)}}(X)\big) - 1\big]\\ &= n\mathbb{E}_{P_j}^{W,\Delta}\big[\tfrac{1}{2}\Delta\big(\eta_{\beta^{(0)}}(X) - \eta_{\beta^{(j)}}(X)\big)^2 + o\big(\Delta\big(\eta_{\beta^{(0)}}(X) - \eta_{\beta^{(j)}}(X)\big)^2\big)\big]\\ &\le n\mathbb{E}_{P_j}^{X}\big[\tfrac{1}{2}\big(\eta_{\beta^{(0)}}(X) - \eta_{\beta^{(j)}}(X)\big)^2 + o\big(\big(\eta_{\beta^{(0)}}(X) - \eta_{\beta^{(j)}}(X)\big)^2\big)\big]\\ &\lesssim nd^2\big(\beta^{(j)}, \beta^{(0)}\big)\\ &\lesssim ns_M.\end{aligned}$$

Therefore, for some positive constant c_1, we have shown that

$$KL(P_j, P_0) \le c_1 n M^{-2r}.$$

By taking M to be the smallest integer greater than $c_2\gamma^{-1/(2r+1)}n^{1/(2r+1)}$ with $c_2 = (c_1 \cdot 8\log 2)^{1/(1+2r)}$, we verified (5.12) that

$$\frac{1}{N}\sum_{j=1}^{N} KL(P_j, P_0) \le \gamma \log N.$$

Meanwhile, since $d^2(\beta^{(i)}, \beta^{(j)}) \geq s_{2M}/2$ and $s_{2M} \asymp (2M)^{-2r}$, condition (5.11) is verified by plugging in M.

SUPPLEMENTARY MATERIAL

Supplement to "Optimal estimation for the functional Cox model" (DOI: 10.1214/00-AOS1441SUPP; .pdf). Due to space constraint, the derivation of $GCV(\lambda)$ and proofs of lemmas are relegated to the supplementary file [Qu, Wang and Wang (2016)].

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. MR0673646

BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–619 (with discussion and with a reply by the authors). MR0803258

CAI, T. T. and YUAN, M. (2012). Minimax and adaptive prediction for functional linear regression. *J. Amer. Statist. Assoc.* **107** 1201–1216. MR3010906

CAREY, J. R., LIEDO, P., MÜLLER, H.-G., WANG, J.-L., SENTURK, D. and HARSHMAN, L. (2005). Biodemography of a long-lived tephritid: Reproduction and longevity in a large cohort of female Mexican fruit flies, Anastrepha ludens. *Exp. Gerontol.* **40** 793–800.

CHEN, K., CHEN, K., MÜLLER, H.-G. and WANG, J.-L. (2011). Stringing high-dimensional data for functional analysis. *J. Amer. Statist. Assoc.* **106** 275–284. MR2816720

COX, D. R. (1972). Regression models and life-tables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **34** 187–220. MR0341758

COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276. MR0400509

CRAMBES, C., KNEIP, A. and SARDA, P. (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist.* **37** 35–72. MR2488344

FERRATY, F. and VIEU, P. (2006). *Nonparametric Functional Data Analysis. Theory and Practice. Springer Series in Statistics*. Springer, New York. MR2229687

FLEMING, T. R. and HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York. MR1100924

GU, C. (2013). *Smoothing Spline ANOVA Models*, 2nd ed. *Springer Series in Statistics* **297**. Springer, New York. MR3025869

HALL, P. and HOROWITZ, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35** 70–91. MR2332269

HASTIE, T. and TIBSHIRANI, R. (1986). Generalized additive models. *Statist. Sci.* **1** 297–318 (with discussion). MR0858512

HASTIE, T. J. and TIBSHIRANI, R. (1990). Exploring the nature of covariate effects in proportional hazards model. *Biometrics* **46** 1005–1016.

HUANG, J. (1999). Efficient estimation of the partly linear additive Cox model. *Ann. Statist.* **27** 1536–1563. MR1742499

JACOBSEN, M. (1984). Maximum likelihood estimation in the multiplicative intensity model: A survey. *Int. Stat. Rev.* **52** 193–207. MR0967210

JAMES, G. M. and HASTIE, T. J. (2002). Generalized linear models with functional predictors. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 533–550.

JOHANSEN, S. (1983). An extension of Cox's regression model. *Int. Stat. Rev.* **51** 165–174. MR0715533

KONG, D., IBRAHIM, J., LEE, E. and ZHU, H. (2014). FLCRM: Functional linear Cox regression models. Submitted.

MÜLLER, H.-G. and STADTMÜLLER, U. (2005). Generalized functional linear models. *Ann. Statist.* **33** 774–805. MR2163159

QU, S., WANG, J. and WANG, X. (2016). Supplement to "Optimal estimation for the functional Cox model." DOI:10.1214/16-AOS1441SUPP.

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York. MR2168993

RIESZ, F. and SZ-NAGY, B. (1990). *Functional Analysis*. Ungar, New York.

SASIENI, P. (1992a). Information bounds for the conditional hazard ratio in a nested family of regression models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **54** 617–635. MR1160487

SASIENI, P. (1992b). Nonorthogonal projections and their application to calculating the information in a partly linear Cox model. *Scand. J. Stat.* **19** 215–233. MR1183198

TSIATIS, A. A. (1981). A large sample study of Cox's regression model. *Ann. Statist.* **9** 93–108. MR0600535

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York. Revised and extended from the 2004 French original, translated by Vladimir Zaiats. MR2724359

VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. Cambridge Univ. Press, Cambridge.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics*. Springer, New York. MR1385671

WAHBA, G. (1990). *Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. SIAM, Philadelphia, PA. MR1045442

YUAN, M. and CAI, T. T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *Ann. Statist.* **38** 3412–3444. MR2766857

S. QU
X. WANG
DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47907
USA
E-MAIL: qu20@purdue.edu
wangxiao@purdue.edu

J.-L. WANG
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
DAVIS, CALIFORNIA 95616
USA
E-MAIL: janelwang@ucdavis.edu