The Annals of Statistics
2010, Vol. 38, No. 4, 2092–2117
DOI: 10.1214/09-AOS780

PENALIZED VARIABLE SELECTION PROCEDURE FOR COX MODELS WITH SEMIPARAMETRIC RELATIVE RISK

BY PANG DU, SHUANGGE MA AND HUA LIANG[1]

Virginia Tech, Yale University and University of Rochester

We study the Cox models with semiparametric relative risk, which can be partially linear with one nonparametric component, or multiple additive or nonadditive nonparametric components. A penalized partial likelihood procedure is proposed to simultaneously estimate the parameters and select variables for both the parametric and the nonparametric parts. Two penalties are applied sequentially. The first penalty, governing the smoothness of the multivariate nonlinear covariate effect function, provides a smoothing spline ANOVA framework that is exploited to derive an empirical model selection tool for the nonparametric part. The second penalty, either the smoothly-clipped-absolute-deviation (SCAD) penalty or the adaptive LASSO penalty, achieves variable selection in the parametric part. We show that the resulting estimator of the parametric part possesses the oracle property, and that the estimator of the nonparametric part achieves the optimal rate of convergence. The proposed procedures are shown to work well in simulation experiments, and then applied to a real data example on sexually transmitted diseases.

1. Introduction. In survival analysis, a problem of interest is to identify relevant risk factors and evaluate their contributions to survival time. Cox proportional hazards (PH) model is a popular approach to study the influence of covariates on survival outcome. Conventional PH models assume that covariates have a log-linear effect on the hazard function. These PH models have been studied by numerous authors; see, for example, the references in [15]. The log-linear assumption can be too rigid in practice, especially when continuous covariates are present. This limitation motivates PH models with nonparametric relative risk. Some examples are [6, 11, 12, 23, 35]. However, nonparametric models may suffer from the curse of dimensionality. They also lack the easy interpretation in parametric risk models. PH models with semiparametric relative risk strike a good balance by allowing nonparametric risk for some covariates and parametric risk for others. The benefits of such models are two-folds. First, they have the merits of models with parametric risk, including easy interpretation, easy estimation and easy inference. Second, their nonparametric part allows a flexible form for some continuous

Received August 2009; revised December 2009.
[1]Supported in part by NIH/NIAID Grant AI59773 and NSF Grant DMS-08-06097.
AMS 2000 subject classifications. Primary 62N01, 62N03; secondary 62N02.
Key words and phrases. Backfitting, partially linear models, penalized variable selection, proportional hazards, penalized partial likelihood, smoothing spline ANOVA.

covariates whose patterns are unexplored and whose contribution cannot be assessed by simple parametric models. For example, [10] proposed efficient estimation for a partially linear Cox model with additive nonlinear covariate effects. Reference [3] studied partially linear hazard regression for multivariate survival data with time-dependent covariates via a profile pseudo-partial likelihood approach, where the only nonlinear covariate effect was estimated by local polynomials. But these models are limited to one nonparametric component or additive nonparametric components, ignoring the possible interactions between different nonparametric components. Reference [30] proposed a partially linear additive hazard model whose nonlinear varying coefficients represent the interaction between the time-dependent nonlinear covariate and other covariates.

Variable selection in survival data has drawn much attention in the past decade. Traditional procedures such as Akaike information criterion (AIC) and Bayesian information criterion (BIC), as noted by [2], suffer from the lack of stability and lack of incorporating stochastic errors inherited in the stage of variable selection. References [26] and [32] extended, respectively, the LASSO and the adaptive LASSO variable selection procedures to the Cox model. Reference [8] extended the nonconcave penalized likelihood approach [7] to the Cox PH models. Reference [4] studied variable selection for multivariate survival data. The Cox models considered in these three papers all assumed a linear form of covariate effects in the relative risk. More recently, [13] and [14] proposed procedures for selecting variables in semiparametric linear regression models for censored data, where the dependence of response over covariates was also assumed to be of linear form. Hence, the aforementioned variable selection procedures are limited in their rigid assumption of parametric covariate effects which may not be realistic in practice. We will fill in these gaps in three aspects: (i) our models are flexible with semiparametric relative risk, which allows nonadditive nonparametric components, without limiting to single or additive nonlinear covariate effects; and (ii) our approach can simultaneously estimate the parametric coefficient vector and select contributing parametric components; and (iii) our approach also provides a model selection tool for the nonparametric components.

Let the hazard function for a subject be

$$h(t) = h_0(t) \exp[\boldsymbol{\beta}^T U + \eta(W)], \tag{1.1}$$

where h_0 is the unknown baseline hazard, $Z^T = (U^T, W^T)$ is the covariate vector, $\boldsymbol{\beta}$ is the unknown coefficient vector, and $\eta(w) = \eta(w_1, \ldots, w_q)$ is an unknown multivariate smooth function. We propose a doubly penalized profile partial likelihood approach for estimation, following the general profile likelihood framework set up by [20]. Given $\boldsymbol{\beta}$, η is estimated by smoothing splines through the minimization of a penalized log partial likelihood. Then the smoothing spline ANOVA decomposition not only allows the natural inclusion of interaction effects but also provides the basis for deriving an empirical model selection tool. After substituting the estimate of η, we obtain a profile partial likelihood, which is then penalized to get an estimate of $\boldsymbol{\beta}$. To achieve variable selection in $\boldsymbol{\beta}$, we use the

smoothly clipped absolute deviation (SCAD) penalty. We show that our estimate of η achieves the optimal convergence rate, and our estimate of $\boldsymbol{\beta}$ possesses the oracle property such that the true zero coefficients are automatically estimated as zeros and the remaining coefficients are estimated as well as if the correct submodel were known in advance. Our numerical studies reveal that the proposed method is promising in both estimation and variable selection. We then apply it to a study on sexually transmitted diseases with 877 subjects.

The rest of the article is organized as follows. Section 2 gives the details of the proposed method, in the order of model description and estimation procedure (Section 2.1), model selection in the nonparametric part (Section 2.2), asymptotic properties (Section 2.3), and miscellaneous issues (Section 2.4) like standard error estimates and smoothing parameter selection. Section 3 presents the empirical studies, and Section 4 gives an application study. Remarks in Section 5 conclude the article.

2. Method. Let T be the failure time and C be the right-censoring time. Assume that T and C are conditionally independent given the covariate. The observable random variable is (X, Δ, Z), where $X = \min(T, C)$, $\Delta = I_{[T \le C]}$, and $Z = (U, W)$ is the covariate vector with $U \in \mathbb{R}^d$ and $W \in \mathbb{R}^q$. With n i.i.d. $(X_i, \Delta_i, Z_i), i = 1, \ldots, n$, we assume a Cox model for the hazard function as in (1.1).

2.1. *Estimation and variable selection for parametric parts.* Let $Y_i(t) = I_{[X_i \ge t]}$. We propose to estimate $(\boldsymbol{\beta}, \eta)$ through a penalized profile partial likelihood approach. Given $\boldsymbol{\beta}$, η is estimated as the minimizer of the penalized partial likelihood

$$l_{\boldsymbol{\beta}}(\eta) \equiv -\frac{1}{n}\sum_{i=1}^{n}\Delta_i\left\{U_i^T\boldsymbol{\beta} + \eta(W_i) - \log\sum_{k=1}^{n}Y_k(X_i)\exp[U_k^T\boldsymbol{\beta} + \eta(W_k)]\right\} + \lambda J(\eta), \tag{2.1}$$

where the summation is the negative log partial likelihood representing the goodness-of-fit, $J(\eta)$ is a roughness penalty specifying the smoothness of η, and $\lambda > 0$ is a smoothing parameter controlling the tradeoff. A popular choice for J is the L_2-penalty which yields tensor product cubic splines (see, e.g., [9]) for multivariate W. Note that η in (2.1) is identifiable up to a constant, so we use the constraint $\int \eta = 0$.

Once an estimate $\hat{\eta}$ of η is obtained, the estimator of $\boldsymbol{\beta}$ is then the maximizer of the penalized profile partial likelihood

$$l_{\hat{\eta}}(\boldsymbol{\beta}) \equiv \sum_{i=1}^{n}\Delta_i\left\{U_i^T\boldsymbol{\beta} + \hat{\eta}(W_i) - \log\sum_{k=1}^{n}Y_k(X_i)\exp[U_k^T\boldsymbol{\beta} + \hat{\eta}(W_k)]\right\} - n\sum_{j=1}^{d}p_{\theta_j}(|\beta_j|), \tag{2.2}$$

where $p_{\theta_j}(|\cdot|)$ is the SCAD penalty on $\boldsymbol{\beta}$ [7].

The detailed algorithm for our estimation procedure is as follows.

Step 1. Find a proper initial estimate $\hat{\beta}^{(0)}$. We note that, as long as the initial estimate is reasonable, convergence to the true optimizer can be achieved. Difference choices of the initial estimate will affect the number of iterations needed but not the convergence itself.

Step 2. Let $\hat{\boldsymbol{\beta}}^{(k-1)}$ be the estimate of $\boldsymbol{\beta}$ before the kth iteration. Plug $\hat{\boldsymbol{\beta}}^{(k-1)}$ into (2.1) and solve for η by minimizing the penalized partial likelihood $l_{\hat{\boldsymbol{\beta}}^{(k-1)}}(\eta)$. Let $\hat{\eta}^{(k)}$ be the estimate thus obtained.

Step 3. Plug $\hat{\eta}^{(k)}$ into (2.2) and solve for $\boldsymbol{\beta}$ by maximizing the penalized profile partial likelihood $l_{\hat{\eta}^{(k)}}(\boldsymbol{\beta})$. Let $\hat{\boldsymbol{\beta}}^{(k)}$ be the estimate thus obtained.

Step 4. Replace $\hat{\boldsymbol{\beta}}^{(k-1)}$ in step 2 by $\hat{\boldsymbol{\beta}}^{(k)}$ and repeat steps 2 and 3 until convergence to obtain the final estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\eta}$.

Our experience shows that the algorithm usually converges quickly within a few iterations. As in the classical Cox proportional hazards model, the estimation of baseline hazard function is of less interest and not required in our estimation procedure.

In step 3, we use a one-step approximation to the SCAD penalty [34]. It transforms the SCAD penalty problem to a LASSO-type optimization, where the celebrated LARS algorithm proposed in [5] can be used. Let $l_{\hat{\eta}}(\boldsymbol{\beta})$ be the profile log partial likelihood in step 3, and $I(\boldsymbol{\beta}) = -\nabla^2 l_{\hat{\eta}}(\boldsymbol{\beta})$ be the Hessian matrix, where the derivative is with respect to $\boldsymbol{\beta}$ treating $\hat{\eta}$ as fixed. Compute the Cholesky decomposition of $I(\hat{\boldsymbol{\beta}}^{(k-1)})$ such that $I(\hat{\boldsymbol{\beta}}^{(k-1)}) = V^T V$. Let $A = \{j : p'_{\theta_j}(|\hat{\beta}_j^{(k-1)}|) = 0\}$ and $B = \{j : p'_{\theta_j}(|\hat{\beta}_j^{(k-1)}|) > 0\}$. Decompose V and the new estimate $\hat{\boldsymbol{\beta}}^{(k)}$ accordingly such that $V = [V_A, V_B]$ and $\hat{\boldsymbol{\beta}}^{(k)} = (\hat{\boldsymbol{\beta}}_A^{(k)^T}, \hat{\boldsymbol{\beta}}_B^{(k)^T})^T$.

(Step 3a) Let $y = V\hat{\boldsymbol{\beta}}^{(k-1)}$. Then for each $j \in B$, replace the jth column of V by setting $v_j = v_j \frac{\theta_j}{p'_{\theta_j}(|\hat{\beta}_j^{(k-1)}|)}$.

(Step 3b) Let $H_A = V_A(V_A^T V_A)^{-1} V_A^T$ be the projection matrix to the column space of V_A. Compute $y^* = y - H_A y$ and $V_B^* = V_B - H_A V_B$.

(Step 3c) Apply the LARS algorithm to solve

$$\hat{\boldsymbol{\beta}}_B^* = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \|y^* - V_B^* \boldsymbol{\beta}\|^2 + n \sum_{j \in B} \theta_j |\beta_j| \right\}.$$

(Step 3d) Compute $\hat{\boldsymbol{\beta}}_A^* = (V_A^T V_A)^{-1} V_A^T (y - V_B \hat{\boldsymbol{\beta}}_B^*)$ to obtain $\hat{\boldsymbol{\beta}}^* = (\hat{\boldsymbol{\beta}}_A^{*T}, \hat{\boldsymbol{\beta}}_B^{*T})^T$.

(Step 3e) For $j \in A$, set $\hat{\beta}_j^{(k)} = \hat{\beta}_j^*$. For $j \in B$, set $\hat{\beta}_j^{(k)} = \hat{\beta}_j^* \frac{\theta_j}{p'_{\theta_j}(|\hat{\beta}_j^{(k-1)}|)}$.

2.2. *Model selection for nonparametric component.* While the SCAD penalty takes care of variable selection for the parametric components, we still need an approach to assess the structure of the nonparametric components. In this section, we will first transform the profile partial likelihood problem in (2.1) to a density estimation problem with biased sampling, and then derive a model selection tool based on the Kullback–Leibler geometry. In this part, we treat $\boldsymbol{\beta}$ as fixed, taking the value from the previous step in the algorithm.

Let $(i_1, \ldots, i_N)$ be the indices for the failed subjects. Then the profile partial likelihood in (2.1) for estimating η is

$$\prod_{i=1}^{n}\Bigg[\frac{e^{U_i^T\boldsymbol{\beta}+\eta(W_i)}}{\sum_{k=1}^{n} Y_k(X_i)e^{U_k^T\boldsymbol{\beta}+\eta(W_k)}}\Bigg]^{\Delta_i} = \prod_{p=1}^{N}\Bigg[\frac{e^{U_{i_p}^T\boldsymbol{\beta}+\eta(W_{i_p})}}{\sum_{k=1}^{n} Y_k(X_{i_p})e^{U_k^T\boldsymbol{\beta}+\eta(W_k)}}\Bigg].$$

Consider the empirical measure P_n^w on the discrete domain $\mathcal{W}_n = \{W_1, \ldots, W_n\}$ such that $\int f\, d\mathrm{P}_n^w = \frac{1}{n}\sum_{i=1}^{n} f(W_i)$. Then $e^{\eta}/\int e^{\eta}\, d\mathrm{P}_n^w$ defines a density function on $\mathcal{W}_n$. Let $a_1(\cdot), \ldots, a_N(\cdot)$ be weight functions defined on the discrete domain $\mathcal{W}_n$ such that $a_p(W_k) = Y_k(X_{i_p})e^{U_k^T\boldsymbol{\beta}}$, $p = 1, \ldots, N$. Alternatively, one can think of a_p's as vectors of weights with length n. Then each term in the profile partial likelihood, with the constant n ignored, becomes $a_p(W_{i_p})e^{\eta(W_{i_p})}/\int a_p(w) \times e^{\eta(w)}\, d\mathrm{P}_n^w$. Thus, this resembles a density estimation problem with bias introduced by the known weight function $a_p(\cdot)$.

For two density estimates η_1 and η_2 in the above pseudo biased sampling density estimation problem, define their Kullback–Leibler distance as

$$\begin{aligned}\mathrm{KL}(\eta_1, \eta_2) = \frac{1}{N}\sum_{p=1}^{N}\Bigg\{&\frac{\int(\eta_1(w)-\eta_2(w))a_p(w)e^{\eta_1(w)}\, d\mathrm{P}_n^w}{\int a_p(w)e^{\eta_1(w)}\, d\mathrm{P}_n^w}\\ &- \log\int a_p(w)e^{\eta_1(w)}\, d\mathrm{P}_n^w\\ &+ \log\int a_p(w)e^{\eta_2(w)}\, d\mathrm{P}_n^w\Bigg\}.\end{aligned} \tag{2.3}$$

Let η_0 be the true function. Suppose the estimation of η_0 has been done in a space $\mathcal{H}_1$, but in fact $\eta_0 \in \mathcal{H}_2 \subset \mathcal{H}_1$. Let $\hat{\eta}$ be the estimate of η_0 in $\mathcal{H}_1$. Let $\tilde{\eta}$ be the Kullback–Leibler projection of $\hat{\eta}$ in $\mathcal{H}_2$, that is, the minimizer of $\mathrm{KL}(\hat{\eta}, \eta)$ for $\eta \in \mathcal{H}_2$, and η_c be the estimate from the constant model. Set $\eta = \tilde{\eta} + \alpha(\tilde{\eta} - \eta_c)$ for α real. Differentiating $\mathrm{KL}(\hat{\eta}, \eta)$ with respect to α and evaluating at $\alpha = 0$, one has

$$\sum_{p=1}^{N}\frac{\int(\tilde{\eta}(w)-\eta_c(w))a_p(w)e^{\hat{\eta}(w)}\, d\mathrm{P}_n^w}{\int a_p(w)e^{\hat{\eta}(w)}\, d\mathrm{P}_n^w} = \sum_{p=1}^{N}\frac{\int(\tilde{\eta}(w)-\eta_c(w))a_p(w)e^{\tilde{\eta}(w)}\, d\mathrm{P}_n^w}{\int a_p(w)e^{\tilde{\eta}(w)}\, d\mathrm{P}_n^w},$$

which, through straightforward calculation, yields

$$\mathrm{KL}(\hat{\eta}, \eta_c) = \mathrm{KL}(\hat{\eta}, \tilde{\eta}) + \mathrm{KL}(\tilde{\eta}, \eta_c).$$

Hence, the ratio $\mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c)$ can be used to diagnose the feasibility of a reduced model $\eta \in \mathcal{H}_2$: the smaller the ratio is, the more feasible the reduced model is.

2.3. *Asymptotic results.* Denote by $\mathcal{H}^m(\mathcal{W})$ the Sobolev space of functions on $\mathcal{W}$ whose mth order partial derivatives are square integrable. Let

$$\mathcal{H} = \left\{ \eta \in \mathcal{H}^m(\mathcal{W}), \int_{\mathcal{W}} \eta(w)\, dw = 0 \right\},$$

and $\hat{\eta}^*$ be the estimate of η_0 in $\mathcal{H}$ that minimizes the penalized partial likelihood

$$-\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}\left\{U_i^T\boldsymbol{\beta} + \eta(W_i) - \log\sum_{k=1}^{n} Y_k(t)\exp[U_k^T\boldsymbol{\beta} + \eta(W_k)]\right\} dN_i(t) + \frac{\lambda}{2}J(\eta). \tag{2.4}$$

Note that $\mathcal{H}$ is an infinite-dimensional function space. Hence, in practice, the minimization of (2.4) is usually performed in a data-adaptive finite-dimensional space

$$\mathcal{H}_n = \mathcal{N}_J \oplus \operatorname{span}\{R_J(W_{i_l}, \cdot) : l = 1, \ldots, q_n\},$$

where $\mathcal{N}_J = \{\eta \in \mathcal{H}, J(\eta) = 0\}$ is the null space of J, and R_J is the *reproducing kernel* (see, e.g., [28]) in its complement space $\mathcal{H}_J = \mathcal{H} \ominus \mathcal{N}_J$, and $\{W_{i_1}, \ldots, W_{i_{q_n}}\}$ is a random subset of $\{W_i : i = 1, \ldots, n\}$. When $q_n = n$, one selects all the W_i, $i = 1, \ldots, n$ as the knots. This is the number of knots used in conventional smoothing splines. However, under the regression setting, [16] showed that a q_n of the order $n^{2/(r+1)+\varepsilon}$, $\forall \varepsilon > 0$ is sufficient to yield an estimate with the optimal convergence rate. Here r is a constant associated with the Sobolev space $\mathcal{H}$, for example, $r = 2m$ for splines of order m (one-dimension w) and $r = 2m - \delta$, $\forall \delta > 0$ for tensor product splines (multi-dimension w). We shall show that such an order for q_n also works for the η estimation in our partially linear Cox model.

Let $s_n[f; \boldsymbol{\beta}, \eta](t) = \frac{1}{n}\sum_{k=1}^{n} Y_k(t) f(U_k, W_k)\exp(U_k^T\boldsymbol{\beta} + \eta(W_k))$ and $s_n[\boldsymbol{\beta}, \eta](t) = s_n(1; \boldsymbol{\beta}, \eta)(t)$. Define

$$\begin{aligned} s[f; \boldsymbol{\beta}, \eta](t) &= E\big[Y(t) f(U, W)\exp\big(U^T\boldsymbol{\beta} + \eta(W)\big)\big] \\ &= \iint f(u, w) e^{u^T\boldsymbol{\beta} + \eta(w)} q(t, u, w)\, du\, dw. \end{aligned}$$

For any functions f and g, define

$$V(f, g) = \int_{\mathcal{T}}\left\{\frac{s[fg; \boldsymbol{\beta}, \eta_0](t)}{s[\boldsymbol{\beta}, \eta_0](t)} - \frac{s[f; \boldsymbol{\beta}, \eta_0](t)}{s[\boldsymbol{\beta}, \eta_0](t)}\frac{s[g; \boldsymbol{\beta}, \eta_0](t)}{s[\boldsymbol{\beta}, \eta_0](t)}\right\} \times s[\boldsymbol{\beta}, \eta_0](t)\, d\Lambda_0(t). \tag{2.5}$$

Write $V(f) \equiv V(f, f)$. Let $\hat{\eta}$ be the estimate that minimizes (2.4) in $\mathcal{H}_n$. Then we have the following theorem.

THEOREM 2.1. *Under conditions* A1–A7 *in the Appendix*,

$$(V+\lambda J)(\hat{\eta}^*-\eta_0)=O_p\big(n^{-r/(r+1)}\big) \quad \text{and} \quad (V+\lambda J)(\hat{\eta}-\eta_0)=O_p\big(n^{-r/(r+1)}\big).$$

This is the optimal convergence rate for estimate of a nonparametric function. In the view of Lemma A.1, this theorem also indicates the same convergence rate in terms of the L_2-norm. Also note that although a higher order of q_n such as $O(n)$ would yield the same convergence rate for $\hat{\eta}$, it will make the function space $\mathcal{H}_n$ too big to apply an entropy bound result that is critical in the proof of Theorem 2.2.

Let $\mathcal{L}_P(\boldsymbol{\beta})=l_p(\boldsymbol{\beta})-n\sum_{j=1}^{d}p_{\theta_j}(|\beta_j|)$, where

$$l_p(\boldsymbol{\beta})=\sum_{i=1}^{n}\int\{U_i^T\boldsymbol{\beta}+\hat{\eta}(W_i)-\log s_n[\boldsymbol{\beta},\hat{\eta}](t)\}\,dN_i(t).$$

Let $\boldsymbol{\beta}_0=(\beta_{10},\ldots,\beta_{d0})^T=(\boldsymbol{\beta}_{10}^T,\boldsymbol{\beta}_{20}^T)^T$ be the true coefficient vector. Without loss of generality, assume that $\boldsymbol{\beta}_{20}=\mathbf{0}$. Let s be the number of nonzero components in $\boldsymbol{\beta}_0$. Define $a_n=\max_j\{|p'_{\theta_j}(|\beta_{j0}|)|:\beta_{j0}\neq 0\}$, $b_n=\max_j\{|p''_{\theta_j}(|\beta_{j0}|)|:\beta_{j0}\neq 0\}$, and

$$\mathbf{b}=(p'_{\theta_1}(\beta_{10})\operatorname{sgn}(\beta_{10}),\ldots,p'_{\theta_s}(\beta_{s0})\operatorname{sgn}(\beta_{s0}))^T,$$
$$\Sigma_{\boldsymbol{\theta}}=\operatorname{diag}\big(p'_{\theta_1}(|\beta_{10}|)/|\beta_{10}|,\ldots,p'_{\theta_s}(|\beta_{s0}|)/|\beta_{s0}|\big).$$

Define $\pi_1:\mathbb{R}^{d+q}\to\mathbb{R}^s$ such that $\pi_1(u,w)=u_1$, where u_1 is the vector of the first s components of u. Let $V_0(\pi_1)$ be defined like $V(\pi_1)$ in (2.5) but with $\boldsymbol{\beta}$ replaced by $\boldsymbol{\beta}_0$.

THEOREM 2.2. *Under conditions* A1–A7 *in the Appendix*, *if* $a_n=O(n^{-1/2})$, $b_n=o(1)$ *and* $q_n=o(n^{1/2})$, *then*:

(i) *There exists a local maximizer* $\hat{\boldsymbol{\beta}}$ *of* $\mathcal{L}_P(\boldsymbol{\beta})$ *such that* $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|=O_p(n^{-1/2})$.

(ii) *Further assume that for all* $1\le j\le d$, $\theta_j=o(1)$, $\theta_j^{-1}=o(n^{1/2})$, *and*

$$\liminf_{n\to\infty}\liminf_{u\to 0^+}\theta_j^{-1}p'_{\theta_j}(u)>0.$$

With probability approaching one, the root-n consistent estimator $\hat{\boldsymbol{\beta}}$ *in* (i) *must satisfy* $\hat{\boldsymbol{\beta}}_2=\mathbf{0}$ *and*

$$\sqrt{n}\big(V_0(\pi_1)+\Sigma_{\boldsymbol{\theta}}\big)\big\{\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10}+\big(V_0(\pi_1)+\Sigma_{\boldsymbol{\theta}}\big)^{-1}\mathbf{b}\big\}\to N(\mathbf{0},V_0(\pi_1)).$$

2.4. *Miscellaneous issues.* In this section, we will propose the standard error estimates for both the parametric and the nonparametric components, and discuss the selection of the smoothing parameters θ and λ.

2.4.1. *Standard error estimates.* Let $l_p(\boldsymbol{\beta})$ be the profile log partial likelihood in the last iteration of step 3 and

$$\Sigma_{\boldsymbol{\theta}}(\boldsymbol{\beta}) = \operatorname{diag}\{p'_{\theta_j}(|\beta_1|)/|\beta_1|, \ldots, p'_{\theta_j}(|\beta_p|)/|\beta_p|\}.$$

Then the standard errors for the nonzero coefficients of $\hat{\boldsymbol{\beta}}$ are given by the sandwich formula

$$\hat{\operatorname{cov}}(\hat{\boldsymbol{\beta}}) = \{\nabla^2 l_p(\hat{\boldsymbol{\beta}}) - n\Sigma_{\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}})\}^{-1}\hat{\operatorname{cov}}\{\nabla l_p(\hat{\boldsymbol{\beta}})\}\{\nabla^2 l_p(\hat{\boldsymbol{\beta}}) - n\Sigma_{\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}})\}^{-1}.$$

Sometimes the standard errors for zero coefficients are also of interest. A discussion of this problem is in Section 5.

In (2.1), η can be decomposed as $\eta = \eta^{[0]} + \eta^{[1]}$ where $\eta^{[0]}$ lies in the null space of the penalty J representing the lower order part and $\eta^{[1]}$ lies in the complement space representing the higher order part. A Bayes model interprets (2.1) as a posterior likelihood when $\eta^{[0]}$ is assigned an improper constant prior and $\eta^{[1]}$ is assigned a Gaussian prior with zero mean and certain covariance matrix. The minimizer $\hat{\eta}$ of (2.1) then becomes the posterior mode. When the minimization is carried out in a data-adaptive function space $\mathcal{H}_n$ with basis functions $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_{q_n})^T$, we can write $\hat{\eta} = \boldsymbol{\psi}^T\hat{\mathbf{c}}$. Then a quadratic approximation to (2.1) yields an approximate posterior covariance matrix for $\mathbf{c}$, which can be used to construct point-wise confidence intervals for η.

2.4.2. *Smoothing parameter selection.* As shown in [33], the effective degrees of freedom for l_1-penalty model is well approximated by the number of nonzero coefficients. Note that our SCAD procedure is implemented by a LASSO approximation at each step. Hence, if we let $\hat{\mathcal{A}}$ be the set of nonzero coefficients, the AIC score for selecting θ in step 3 is

$$\mathrm{AIC} = 2l_p(\hat{\boldsymbol{\beta}}) + 2|\hat{\mathcal{A}}|,$$

where $|\hat{\mathcal{A}}|$ is the cardinality of $\hat{\mathcal{A}}$.

As illustrated in Section 2.2, the estimation of η in step 2 can be cast as a density estimation problem with biased sampling. Let $\mathrm{KL}(\eta_0, \hat{\eta}_\lambda)$ be the Kullback–Leibler distance, as defined in (2.3), between the true "density" $e^{\eta_0}/\int e^{\eta_0}\,d\mathrm{P}_n^w$ and the estimate $e^{\hat{\eta}_\lambda}/\int e^{\hat{\eta}_\lambda}\,d\mathrm{P}_n^w$. An optimal λ should minimize $\mathrm{KL}(\eta_0, \hat{\eta}_\lambda)$ or the relative Kullback–Leibler distance

$$\mathrm{RKL}(\eta_0, \hat{\eta}_\lambda) = \frac{1}{N}\sum_{p=1}^{N}\left\{\frac{\int(\eta_0(w) - \hat{\eta}_\lambda(w))a_p(w)e^{\eta_0(w)}\,d\mathrm{P}_n^w}{\int a_p(w)e^{\eta_0(w)}\,d\mathrm{P}_n^w} + \log\int a_p(w)e^{\hat{\eta}_\lambda(w)}\,d\mathrm{P}_n^w\right\}. \tag{2.6}$$

The second term of (2.6) is directly computable from the estimate $\hat{\eta}_\lambda$. But the first term needs to be estimated. Let $\boldsymbol{\psi}$ be the vector of spline basis functions as

in the previous subsection and $\eta = \boldsymbol{\psi}^T \mathbf{c}$. Through a delete-one cross-validation approximation, a proxy for (2.6) can be derived as

$$-\frac{1}{N}\sum_{p=1}^{N}\left\{\eta(W_i) - \log\int a_p(w)e^{\eta(w)}\,d\mathrm{P}_n^w\right\} + \frac{\mathrm{tr}(P_{\mathbf{1}}Q^T H^{-1} Q P_{\mathbf{1}})}{N(N-1)},$$

where $P_{\mathbf{1}} = I - \mathbf{1}\mathbf{1}^T/N$, $Q = (\boldsymbol{\psi}(W_{i_1}), \ldots, \boldsymbol{\psi}(W_{i_N}))$, and H is the hessian matrix for minimizing (2.1) with respect to the coefficient vector $\mathbf{c}$. λ is chosen to minimize this score.

3. Numerical studies. In the simulations, we generated failure times from the exponential hazard model with $h(t|U, W) = \exp[U^T\boldsymbol{\beta}_0 + \eta_0(W)]$. We used the same settings for the parametric component, which consists of eight covariates U_j, $j = 1, \ldots, 8$. The U_j's were generated from a multivariate normal distribution with zero mean and $\mathrm{Cov}(U_j, U_k) = 0.5^{|j-k|}$. The true coefficient vector was $\boldsymbol{\beta}_0 = (0.8, 0, 0, 1, 0, 0, 0.6, 0)^T$.

The theory in Section 2.3 gives the sufficient order for q_n, the number of knots in our smoothing spline estimation of η. In practice, [16] suggested $q_n = kn^{2/(2m+1)}$ with $k = 10$ if the tensor product splines of order m are used. Since we use tensor product cubic splines in all the simulations below, our choice is $q_n = 10n^{2/5}$.

3.1. *Variable selection for parametric components.* The nonparametric part had one covariate W generated from Uniform(0, 1). Two different η_0 were used:

$$\eta_{0a}(w) = 1.5\sin\left(2\pi w - \frac{\pi}{2}\right) \quad \text{or} \quad \eta_{0b}(w) = 4(w-0.3)^2 + 4.7e^{-w} - 3.4643.$$

Note that both functions satisfies $\int_0^1 \eta_0(w)\,dw = 0$. Given U and W, the censoring times were generated from exponential distributions such that the censoring rates are, respectively, 23% and 40%. Sample sizes $n = 150$ and 500 were considered. One thousand data replicates were generated for each of the four combinations of η_0 and n.

For a prediction procedure $\mathcal{M}$ and the estimator $(\hat{\boldsymbol{\beta}}_{\mathcal{M}}, \hat{\eta}_{\mathcal{M}})$ yielded from the procedure, an appropriate measure for the goodness-of-fit under Cox model with $h_0(t) \equiv 1$ is the model error: $\mathrm{ME}(\hat{\boldsymbol{\beta}}_{\mathcal{M}}, \hat{\eta}_{\mathcal{M}}) = E[(\exp(-U^T\hat{\boldsymbol{\beta}}_{\mathcal{M}} - \hat{\eta}_{\mathcal{M}}(W)) - \exp(-U^T\boldsymbol{\beta}_0 - \eta_0(W)))^2]$. The relative model error (RME) of $\mathcal{M}_1$ versus $\mathcal{M}_2$ is defined as the ratio $\mathrm{ME}(\hat{\boldsymbol{\beta}}_{\mathcal{M}_1}, \hat{\eta}_{\mathcal{M}_1})/\mathrm{ME}(\hat{\boldsymbol{\beta}}_{\mathcal{M}_2}, \hat{\eta}_{\mathcal{M}_2})$. The procedure $\mathcal{M}_0$ with complete oracle is used as our benchmark. In $\mathcal{M}_0$, (U_1, U_4, U_7, W) are known to be the only contributing covariates, the exact form of η_0 is known, and the only parameters to be estimated are the coefficients of U_1, U_4, U_7. Note that $\mathcal{M}_0$ can be implemented only in simulations, but is unrealistic in practice since neither the contributing covariates nor the form of η_0 would be known. We then compare the performance of the following four procedures, including the proposed procedures, through their RMEs versus $\mathcal{M}_0$:

$\mathcal{M}_A$: procedure with partial oracle and misspecified parametric η_0, that is, (U_1, U_4, U_7, W) are known to be the only contributing covariates but η_0 is misspecified to be of the parametric form $\eta_0(W) = \beta_W W$ and β_W is estimated together with the coefficients for (U_1, U_4, U_7);

$\mathcal{M}_B$: procedure with partial oracle and estimated η_0, that is, (U_1, U_4, U_7, W) are known to be the only contributing covariates but the form of η_0 is unknown, and η_0 is estimated together with the coefficients for (U_1, U_4, U_7) by penalized profile partial likelihood;

$\mathcal{M}_C$: the proposed partial linear procedure with the SCAD penalty on $\boldsymbol{\beta}$;

$\mathcal{M}_D$: the proposed partial linear procedure with the adaptive LASSO penalty on $\boldsymbol{\beta}$.

Procedure $\mathcal{M}_A$ has a misspecified covariate effect. We intend to show that the estimation results can be unsatisfactory if the semiparametric form of covariate effect is mistakenly specified as parametric. Procedure $\mathcal{M}_B$ is "partial oracle" and expected to have equal or better performance than procedures $\mathcal{M}_C$ and $\mathcal{M}_D$. Note, however, $\mathcal{M}_B$ is unrealistic in practice since the contributing covariates would not be known. $\mathcal{M}_C$ and $\mathcal{M}_D$ are two versions of the proposed partial linear procedure with different penalties on $\boldsymbol{\beta}$.

For each combination of η_0 and n, we computed the following quantities out of the 1000 data replicates: the median RMEs of the complete oracle procedure $\mathcal{M}_0$ versus the procedures $\mathcal{M}_A$ to $\mathcal{M}_D$, the average number of correctly selected nonzero coefficients (CC), the average number of incorrectly selected nonzero coefficients (IC), the proportion of under-fit replicates that excluded any nonzero coefficients, the proportion of correct-fit replicates that selected the exact subset model, and the proportion over-fit replicates that included all three significant variables and some noise variables. The results are summarized in Table 1. In general, a partial oracle with misspecified parametric η_0 (procedure $\mathcal{M}_A$) has much inferior performance when comparing with the other three procedures; the proposed procedure with the SCAD penalty (procedure $\mathcal{M}_C$) or the adaptive LASSO penalty (procedure $\mathcal{M}_D$) is competitive to the partial oracle with estimated η_0 (procedure $\mathcal{M}_B$); the SCAD penalty performs slightly better than the adaptive LASSO penalty. Also, the proposed procedure generally performs as well as the complete oracle. For procedure $\mathcal{M}_C$, we also did some extra computation to evaluate the proposed standard error estimate of $\boldsymbol{\beta}$. In Table 2, SD is the median absolute deviation divided by 0.6745 of the 1000 nonzero $\hat{\boldsymbol{\beta}}$'s that can be regarded as the true standard error, SD_m is the median of the 1000 estimated SDs, and $\mathrm{SD}_{\mathrm{mad}}$ is the median absolute deviation of the 1000 estimated SDs divided by 0.6745. The standard errors were set to 0 for the coefficients estimated as 0s. The results in Table 2 suggests a good performance of the proposed standard error formula for $\boldsymbol{\beta}$.

To examine the estimation of η_0, we computed the point-wise estimates at the grid $w = (0, 1, \text{by} = 0.01)$ for each data replicate. Then at each grid point, the

TABLE 1
Variable selection for parametric components (*Section* 3.1)

Procedure	MRME	No. of nonzeros CC	No. of nonzeros IC	Proportion of Under-fit	Proportion of Correct-fit	Proportion of Over-fit
			$n = 150$, $\eta_0 = \eta_{0a}$ (23% censoring)			
$\mathcal{M}_A$	0.168	–	–	–	–	–
$\mathcal{M}_B$	0.475	–	–	–	–	–
$\mathcal{M}_C$	0.409	2.998	0.825	0.002	0.476	0.522
$\mathcal{M}_D$	0.387	2.998	0.959	0.002	0.444	0.554
			$n = 150$, $\eta_0 = \eta_{0b}$ (40% censoring)			
$\mathcal{M}_A$	0.167	–	–	–	–	–
$\mathcal{M}_B$	0.711	–	–	–	–	–
$\mathcal{M}_C$	0.518	2.996	0.949	0.004	0.430	0.566
$\mathcal{M}_D$	0.563	2.998	1.131	0.002	0.378	0.620
			$n = 500$, $\eta_0 = \eta_{0a}$ (23% censoring)			
$\mathcal{M}_A$	0.056	–	–	–	–	–
$\mathcal{M}_B$	0.431	–	–	–	–	–
$\mathcal{M}_C$	0.396	3.000	0.717	0.000	0.525	0.475
$\mathcal{M}_D$	0.375	3.000	0.736	0.000	0.540	0.460
			$n = 500$, $\eta_0 = \eta_{0b}$ (40% censoring)			
$\mathcal{M}_A$	0.057	–	–	–	–	–
$\mathcal{M}_B$	0.712	–	–	–	–	–
$\mathcal{M}_C$	0.619	3.000	0.749	0.000	0.512	0.488
$\mathcal{M}_D$	0.628	3.000	0.776	0.000	0.529	0.471

mean, the 0.025 and the 0.975 quantiles of the 1000 estimates, together with the mean of the 1000 95% confidence intervals were computed. The results are in Figure 1. The plots show satisfactory nonparametric fits and standard error estimates.

TABLE 2
Standard deviations for $\hat{\boldsymbol{\beta}}$*'s in the partial linear SCAD procedure* $\mathcal{M}_C$ (*Section* 3.1)

n, censor %	$\hat{\beta}_1$ SD	$\hat{\beta}_1$ SD_m (SD_{mad})	$\hat{\beta}_4$ SD	$\hat{\beta}_4$ SD_m (SD_{mad})	$\hat{\beta}_7$ SD	$\hat{\beta}_7$ SD_m (SD_{mad})
150, 23%	0.124	0.113 (0.015)	0.141	0.121 (0.017)	0.135	0.109 (0.015)
150, 40%	0.159	0.135 (0.017)	0.188	0.145 (0.021)	0.155	0.128 (0.019)
500, 23%	0.065	0.059 (0.005)	0.073	0.063 (0.005)	0.062	0.057 (0.005)
500, 40%	0.075	0.070 (0.006)	0.088	0.076 (0.006)	0.078	0.066 (0.006)

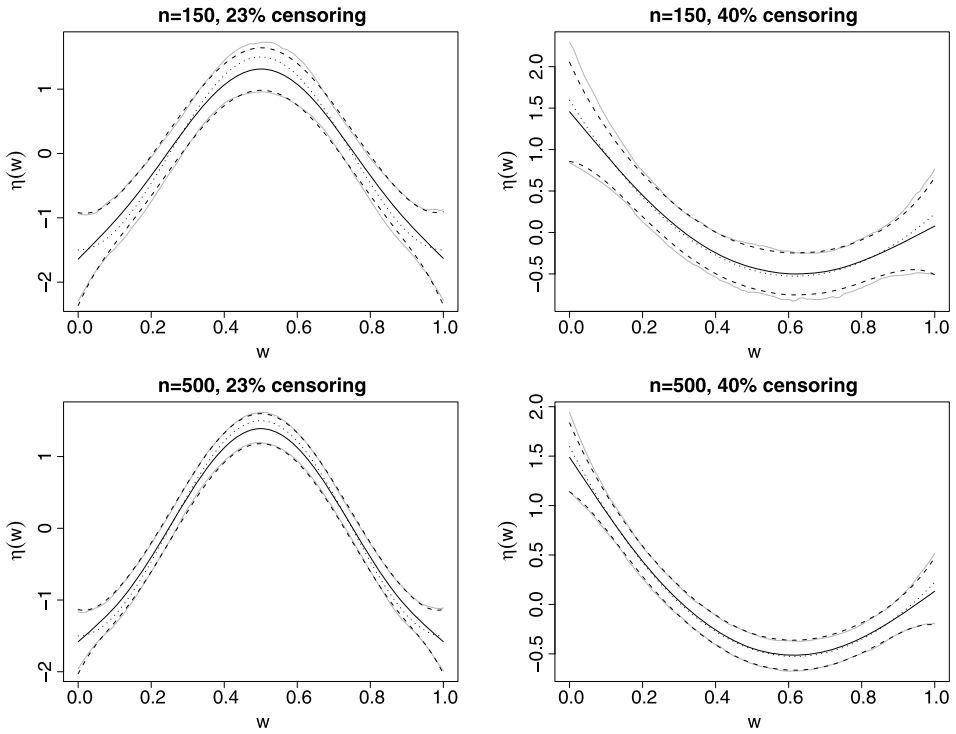

FIG. 1. *Estimates for nonparametric components* (*Section* 3.1). *Dotted lines are true function*, *solid lines are connected point-wise mean estimates*, *faded lines are connected* 0.025 *and* 0.975 *quantiles of the point-wise estimates*, *and dashed lines are the connected point-wise* 95% *confidence intervals*.

3.2. *Model selection for nonparametric components*. In this section, we present some simulations to evaluate the model selection tool for nonparametric part introduced in Section 2.2. We used the SCAD penalty on the parametric components in this section. Two covariates W_1 and W_2, independently generated from Uniform(0, 1), were used. We considered two scenarios for the true model of the nonparametric part: (i) nonparametric univariate model $\eta_0(W) = \eta_{01}(W_1)$ and (ii) nonparametric bivariate additive model $\eta_0(W) = \eta_{01}(W_1) + \eta_{02}(W_2)$. For scenario (i), the data sets generated in the last section were used, with W_1 being the existing W covariate and W_2 being an additional noise covariate. The fitted model was nonparametric additive in W_1 and W_2. The ratios $\mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c)$ for the projections to the univariate models $\eta_0(W) = \eta_{01}(W_1)$ and $\eta_0(W) = \eta_{02}(W_2)$ were computed. For scenario (ii), we considered two sample sizes $n = 150$ and 300. The true η_0 was

$$\eta_0(w_1, w_2) = 0.7\eta_{0a}(w_1) + 0.3\eta_{0b}(w_2)$$

or

$$\eta_0(w_1, w_2) = \eta_{0a}(w_1) + \eta_{0b}(w_2),$$

TABLE 3
Model selection for nonparametric components (Section 3.2)

Sample size	Proportion of selecting			Proportion of		
	W_1	W_2	$W_1 : W_2$	Under-fit	Correct-fit	Over-fit
True model: $\eta_0(w_1, w_2) = \eta_{0a}(w_1)$, 23% censoring						
$n = 150$	1.000	0.036	–	0.000	0.964	0.036
$n = 500$	1.000	0.002	–	0.000	0.998	0.002
True model: $\eta_0(w_1, w_2) = \eta_{0b}(w_1)$, 40% censoring						
$n = 150$	1.000	0.304	–	0.000	0.696	0.304
$n = 500$	1.000	0.062	–	0.000	0.938	0.062
True model: $\eta_0(w_1, w_2) = 0.7\eta_{0a}(w_1) + 0.3\eta_{0b}(w_2)$, 25% censoring						
$n = 150$	1.000	0.998	0.084	0.002	0.914	0.084
$n = 300$	1.000	1.000	0.013	0.000	0.987	0.013
True model: $\eta_0(w_1, w_2) = \eta_{0a}(w_1) + \eta_{0b}(w_2)$, 39% censoring						
$n = 150$	1.000	0.672	0.201	0.328	0.471	0.201
$n = 300$	1.000	0.616	0.096	0.384	0.520	0.096

where η_{0a} and η_{0b} are as defined in Section 3.1. The censoring times were generated from exponential distributions such that the resulting censoring rates were, respectively, 25% and 39%. Note that both choices of η_0 are additive in w_1 and w_2. The fitted model was the nonparametric bivariate full model with both the main effects and the interaction. Then the ratios $\mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c)$ for the projections to the bivariate additive model and the two univariate models were computed. In both scenarios, we claim a reduced model is feasible when the corresponding ratio $\mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c) < 0.05$.

For each of the eight combinations of η_0 and n, we simulated 1000 data replicates and computed the proportions of replicates that produced the following results in the reduced model: selected the main effect of W_1, selected the main effect of W_2, selected the interaction $W_1 : W_2$, under-fitted the model by excluding at least one truly significant effect, correctly fitted the model by reducing to the exact subset model, and over-fitted the model by including all the truly significant effects and some irrelevant effects. These proportion results are summarized in Table 3. It shows that the variable selection tool for the nonparametric component works very well. The better performance appears to be associated with bigger sample sizes and lower censoring rates.

4. Example. An example in [17] is a study on two sexually transmitted diseases: gonorrhea and chlamydia. The purpose of the study was to identify factors that are related to time until reinfection by gonorrhea or chlamydia given an initial infection of either disease. A sample of 877 individuals with an initial diagnosis of

gonorrhea or chlamydia were followed for reinfection. Recorded for each individual were follow-up time, indicator of reinfection, demographic variables including race (white or black, U_1), marital status (divorced/separated, married or single, U_2 and U_3), age at initial infection (W_1), years of schooling (W_2) and type of initial infection (gonorrhea, chlamydia or both, U_4 and U_5), behavior factors at the initial diagnosis including number of partners in the last 30 days (U_6), indicators of oral sex within past 12 months and within past 30 days (U_7 and U_8), indicators of rectal sex within past 12 months and within past 30 days (U_9 and U_{10}) and condom use (always, sometimes or never, U_{11} and U_{12}), symptom variables at time of initial infection including presence of abdominal pain (U_{13}), sign of discharge (U_{14}), sign of dysuria (U_{15}), sign of itch (U_{16}), sign of lesion (U_{17}), sign of rash (U_{18}) and sign of lymph involvement (U_{19}) and symptom variables at time of examination including involvement vagina at exam (U_{20}), discharge at exam (U_{21}) and abnormal node at exam (U_{22}).

We used $q_n = 10 \cdot 877^{2/5} = 151$ knots in all the analysis below. We first considered the partial linear Cox model

$$h_i(t|Z) = h_0(t) \exp\left\{ \sum_{j=1}^{3} \eta_j(W_{ji}) + \sum_{k=1}^{22} U_{ki}\beta_k \right\},$$

where $\eta_3(W_{3i}) = \eta_3(W_{1i}, W_{2i})$ is the interaction term between W_1 and W_2. However, the interaction term was found to be negligible with the ratio $\mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c) = 0.003$. Hence, we took out this interaction term and refitted the model. In this model, neither W_1 (age) nor W_2 (years of schooling) in the nonparametric component were found to be negligible, with the ratios $\mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c)$ equal to 0.633 for removing W_1 and 0.259 for removing W_2. Their effects are plotted in Figure 2 together with the 95% point-wise confidence interval. We can see that the hazard increased with age at both ends of the age domain (between age 13

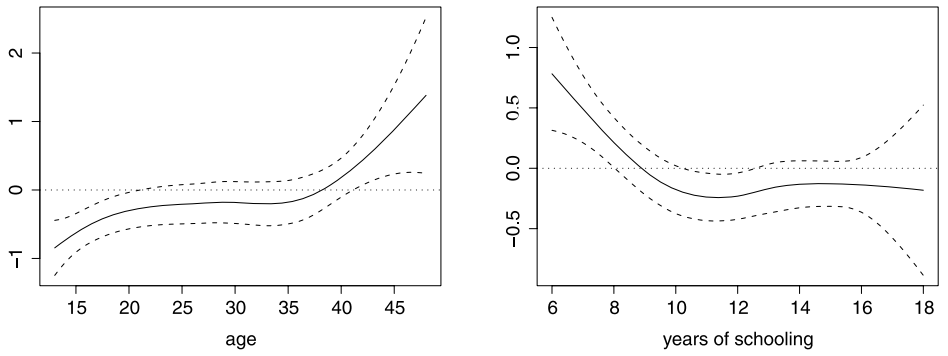

FIG. 2. *Nonparametric component estimates for sexually transmitted diseases data. Left*: *effect of age at initial infection. Right*: *effect of years of schooling. Solid lines are the estimates, dashed lines are* 95% *confidence intervals and dotted lines are the reference zero lines.*

TABLE 4

Fitted coefficients and their standard errors for sexually transmitted diseases data. (*Models from top to bottom*: *semiparametric relative risk with SCAD penalty and with adaptive LASSO penalty, parametric relative risk with SCAD penalty and with adaptive LASSO penalty*)

age	***yschool***	**npart**	**raceW**	**maritalM**	**maritalS**
– (–)	– (–)	0 (–)	0 (–)	0 (–)	0.487 (0.212)
– (–)	– (–)	0.060 (0.048)	−0.127 (0.097)	0 (–)	0.448 (0.186)
0 (–)	−0.059 (0.018)	0 (–)	0 (–)	0 (–)	0.332 (0.213)
0 (–)	−0.119 (0.031)	0.026 (0.024)	0 (–)	0 (–)	0.210 (0.119)
typeC	**typeB**	**oralY**	**oralM**	**rectY**	**rectM**
−0.412 (0.149)	−0.337 (0.144)	−0.336 (0.201)	−0.341 (0.235)	0 (–)	0 (–)
−0.349 (0.137)	−0.300 (0.130)	−0.330 (0.155)	−0.318 (0.173)	0 (–)	0 (–)
−0.376 (0.149)	−0.249 (0.145)	−0.236 (0.202)	−0.348 (0.235)	0 (–)	0 (–)
−0.228 (0.096)	−0.083 (0.065)	−0.110 (0.058)	−0.371 (0.117)	0 (–)	0 (–)
abdom	**disc**	**dysu**	**condS**	**condN**	**itch**
0.253 (0.151)	0 (–)	0.193 (0.152)	0 (–)	−0.327 (0.114)	0 (–)
0.177 (0.120)	0 (–)	0.089 (0.074)	0.152 (0.114)	−0.291 (0.106)	0 (–)
0.285 (0.148)	0 (–)	0 (–)	0 (–)	−0.296 (0.114)	0 (–)
0.184 (0.094)	0 (–)	0 (–)	0 (–)	−0.223 (0.092)	0 (–)
lesion	**rash**	**lymph**	**involve**	**discE**	**node**
0 (–)	0 (–)	0 (–)	0.423 (0.166)	−0.460 (0.220)	0 (–)
0 (–)	0 (–)	0 (–)	0.327 (0.159)	−0.407 (0.209)	0 (–)
0 (–)	0 (–)	0 (–)	0.392 (0.168)	−0.443 (0.221)	0 (–)
0 (–)	0 (–)	0 (–)	0.289 (0.133)	−0.280 (0.163)	0 (–)

and 20, and between age 38 and 48) and stayed flat in the middle, and that the hazard decreased with years of school from 6 years to 10 years but stayed flat afterwards. The fitted coefficients from the proposed method with the SCAD penalty are in Table 4 together with their standard error estimates. For comparisons, Table 4 also lists the fitted coefficients and standard errors for three other models, namely the proposed semiparametric relative risk model with the adaptive LASSO penalty, and the parametric relative risk models with the SCAD and the adaptive LASSO penalties. We can see that the SCAD penalty yielded sparser models than the adaptive LASSO penalty, and that both parametric models missed the age effect. Common factors identified by all the four procedures to be associated with reinfection risk are marital status, type of infection, oral sex behavior, condom use, sign of abdominal pain, sign of lymph involvement and sign of discharge at exam.

5. Discussion. We have proposed a Cox PH model with semiparametric relative risk. The nonparametric part of the risk is estimated by smoothing spline ANOVA model and model selection procedure derived based on a Kullback–Leibler geometry. The parametric part of the risk is estimated by penalized profile partial likelihood and variable selection achieved by choosing a nonconcave penalty. Both theoretical and numerical studies show promising results for the proposed method. An important question in using the method in practice is which covariate effects should be treated as parametric. We suggest the following guideline for making choices. As a starting point, the effects of all the continuous covariates are put in the nonparametric part and those of the discrete covariates in the parametric part. If the estimation results show that some of the continuous covariate effects can be described by certain parametric forms such as linear form, then a new model can be fitted with those continuous covariate effects moved to the parametric part. In this way, one can take full advantage of the flexible exploratory analysis provided by the proposed method.

We thank a referee for raising the interesting question on the standard error estimates for the coefficients estimated to be 0 in $\hat{\boldsymbol{\beta}}$. References [25] and [7] suggested to set these standard errors to 0s based on the belief that those covariates with zero coefficient estimates are not important. This is the approach adopted here. When such a belief is in doubt, nonzero standard errors are preferred even for coefficients estimated to be 0's. This problem has been addressed only in a few papers. Reference [22] looked at the problem for LASSO but it is based on a smooth approximation. Reference [24] presented a Bayesian approach and pointed out that no fully satisfactory frequentist solution had been proposed so far, no matter LASSO or SCAD variable selection procedure is considered. This problem presents an interesting challenge that we hope to address in some future work.

Another choice of $p_\theta(|\cdot|)$ is the adaptive LASSO penalty [31]. Our simulations in Section 3.1 indicates a similar performance when compared to the SCAD penalty. So we decided not to present the details here.

Although our method is presented for time-independent covariates, a lengthier argument modifying [23] can yield similar theoretical results for external time-dependent covariates [15]. However, the implementation of such extension is more complicated and not pursued here.

A recently proposed nonparametric component selection procedure in a penalized likelihood framework is the COSSO method in [19] where the penalty switches from $J(\eta)$ to $J^{1/2}(\eta)$. Taking advantage of the smoothing spline ANOVA decomposition, the COSSO method does model selection by applying a soft thresholding type operation to the function components. An extension of COSSO to the Cox proportional hazards model with nonparametric relative risk is available in [18]. Although a similar extension to our proportional hazards model with semiparametric relative risk is of interest, it is not clear whether the theoretical properties of the COSSO method such as the existence and the convergence rate

of the COSSO estimator can be transferred to the estimation of η under our semiparametric setting. Furthermore, the dimension of the function space in COSSO is $O(n)$, too big to allow an entropy bound that is critical in deriving the asymptotic properties of $\hat{\boldsymbol{\beta}}$.

APPENDIX: PROOFS

For $z = (u, w)$, let $p_z(t) = \exp(u^T\boldsymbol{\beta} + \eta_0(w))q(t, u, w)/s[\boldsymbol{\beta}, \eta_0](t)$ and $\bar{f}_t \equiv \iint f(u, w)p_z(t)\,du\,dw$. Let $\tilde{S}(t, u, w) = E[Y(t) = 1|U = u, W = w] = P(Y(t) = 1|U = u, W = w)$ and $q(t, u, w) = \tilde{S}(t, u, w)p(u, w)$, where $p(u, w)$ is the density function of (U, W). Let $\mathcal{Z} = \mathcal{U} \times \mathcal{W}$ be the domain of the covariate $Z = (U^T, W^T)^T$. We need the following conditions.

A1. The true coefficient β_0 is an interior point of a bounded subset of $\mathbb{R}^d$.
A2. The domain $\mathcal{Z}$ of covariate is a compact set in $\mathbb{R}^{d+q}$.
A3. Failure time T and censoring time C are conditionally independent given the covariate Z.
A4. Assume the observations are in a finite time interval $[0, \tau]$. Assume that the baseline hazard function $h_0(t)$ is bounded away from zero and infinity.
A5. Assume that there exist constants $k_2 > k_1 > 0$ such that $k_1 < q(t, u, w) < k_2$ and $|\frac{\partial}{\partial t}q(t, u, w)| < k_2$.
A6. Assume the true function $\eta_0 \in \mathcal{H}$. For any η in a sufficiently big convex neighborhood B_0 of η_0, there exist constants $c_1, c_2 > 0$ such that $c_1 e^{\eta_0(w)} \le e^{\eta(w)} \le c_2 e^{\eta_0(w)}$ for all w.
A7. The smoothing parameter $\lambda \asymp n^{-r/(r+1)}$.

Condition A1 requires that β_0 is not on the boundary of the parameter space. Condition A2 is also a common boundedness assumption on covariate. Condition A3 assumes noninformative censoring. Condition A4 is the common boundedness assumption on the baseline hazard. Condition A5 bounds the joint density of (T, Z) and thus also the derivatives of the partial likelihood. Condition A6 assumes that η_0 has proper level of smoothness and integrates to zero. The neighborhood B_0 in condition A6 should be big enough to contain all the estimates of η_0 considered below. When the members of B_0 are all uniformly bounded, condition A6 is automatically satisfied. The order for λ in condition A7 matches that in standard smooth spline problems.

We first show the equivalence between $V(\cdot)$ and the L_2-norm $\|\cdot\|_2^2$.

LEMMA A.1. *Let $f \in \mathcal{H}$. Then there exist constants $0 < c_3 \le c_4 < \infty$ such that*

$$c_3\|f\|_2^2 \le V(f) \le c_4\|f\|_2^2.$$

PROOF. For $z = (u, w)$, let $p_z(t) = \exp(u^T \boldsymbol{\beta} + \eta_0(w))q(t, u, w)/s[\boldsymbol{\beta}, \eta_0](t)$ and $\bar{f}_t \equiv \iint f(u, w) p_z(t)\, du\, dw$. Simple algebraic manipulation yields

$$V(f) = \int_{\mathcal{T}} \left\{ \iint (f(u, w) - \bar{f}_t)^2 p_z(t)\, du\, dw \right\} s[\boldsymbol{\beta}, \eta_0](t)\, d\Lambda_0(t).$$

By conditions A4 and A5, there exist positive constants c_1 and c_2 such that

$$c_1 \int_{\mathcal{T}} \left\{ \iint (f(u, w) - \bar{f}_t)^2\, du\, dw \right\} d\Lambda_0(t)$$
$$\le V(f) \le c_2 \int_{\mathcal{T}} \left\{ \iint (f(u, w) - \bar{f}_t)^2\, du\, dw \right\} d\Lambda_0(t).$$

Let $m(\mathcal{Z}) < \infty$ be the Lebesgue measure of $\mathcal{Z}$. Then

$$\int_{\mathcal{T}} \left\{ \iint (f(u, w) - \bar{f}_t)^2\, du\, dw \right\} d\Lambda_0(t)$$
$$= \Lambda_0(\tau) \iint f^2(u, w)\, du\, dw + m(\mathcal{Z}) \iint [\bar{f}_t]^2\, d\Lambda_0(t).$$

The lemma follows from the Cauchy–Schwarz inequality and condition A4. □

PROOF OF THEOREM 2.1. We will prove the results using an eigenvalue analysis of three steps. In the first step (linear approximation), we show the convergence rate $O_p(n^{-r/(r+1)})$ for the minimizer $\tilde{\eta}$ of a quadratic approximation to (2.4). In the second step (approximation error), we show that the difference between $\tilde{\eta}$ and the estimate $\hat{\eta}^*$ in $\mathcal{H}$ is also $O_p(n^{-r/(r+1)})$, and so is the convergence rate of $\hat{\eta}^*$. In the third step (semiparametric approximation), we show that the projection η^* of $\hat{\eta}^*$ in $\mathcal{H}_n$ is not so different from either $\hat{\eta}^*$ or the estimate $\hat{\eta}$ in $\mathcal{H}_n$, and then the convergence rate of $\hat{\eta}$ follows.

A quadratic function B is said to be *completely continuous* with respect to another quadratic functional A, if for any $\varepsilon > 0$, there exists a finite number of linear functionals $L_1, \ldots, L_k$ such that $L_j f = 0$, $j = 1, \ldots, k$, implies that $B(f) \le \varepsilon A(f)$; When a quadratic functional B is *completely continuous* with respect to another quadratic functional A, there exists eigenfunctions $\{\phi_\nu, \nu = 1, 2, \ldots\}$ such that $B(\phi_\nu, \phi_\mu) = \delta_{\nu\mu}$ and $A(\phi_\nu, \phi_\mu) = \rho_\nu \delta_{\nu\mu}$, where $\delta_{\nu\mu}$ is the Kronecker delta and $0 \le \rho_\nu \uparrow \infty$. And functions satisfying $A(f) < \infty$ can be expressed as a *Fourier series expansion* $f = \sum_\nu f_\nu \phi_\nu$, where $f_\nu = B(f, \phi_\nu)$ are the *Fourier coefficients*. See, for example, [9] and [29].

We first present two lemmas without proof. The first one follows directly from the results in Section 8.1 of [9] and Lemma A.1. The second one is exactly Lemma 8.1 in [9].

LEMMA A.2. *V is completely continuous to J and the eigenvalues ρ_ν of J with respect to V satisfy that as $\nu \to \infty$, $\rho_\nu^{-1} = O(\nu^r)$.*

LEMMA A.3. *As* $\lambda \to 0$, *the sums* $\sum_\nu \frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2}$, $\sum_\nu \frac{1}{(1+\lambda\rho_\nu)^2}$, *and* $\sum_\nu \frac{1}{1+\lambda\rho_\nu}$ *are all of order* $O(\lambda^{-1/r})$.

STEP 1 (Linear approximation). A linear approximation $\tilde\eta$ to $\hat\eta^*$ is the minimizer of a quadratic approximation to (2.4),

$$\begin{aligned} -\frac{1}{n}\sum_{i=1}^n \int_{\mathcal{T}} \{\eta(W_i) - s_n^{-1}[\boldsymbol{\beta}, \eta_0](t) s_n[\eta-\eta_0; \boldsymbol{\beta}, \eta_0](t)\}\, dN_i(t) \\ + \frac{1}{2}V(\eta-\eta_0) + \frac{\lambda}{2}J(\eta). \end{aligned} \tag{A.1}$$

Let $\eta = \sum_\nu \eta_\nu \phi_\nu$ and $\eta_0 = \sum_\nu \eta_{\nu,0}\phi_\nu$ be the Fourier expansions of η and η_0. Plugging them into (A.1) and dropping the terms not involving η yield

$$\sum_\nu \left\{ -\eta_\nu \gamma_\nu + \frac{1}{2}(\eta_\nu - \eta_{\nu,0})^2 + \frac{\lambda}{2}\rho_\nu \eta_\nu^2 \right\}, \tag{A.2}$$

where $\gamma_\nu = \frac{1}{n}\sum_{i=1}^n \int_{\mathcal{T}} \{\phi_\nu(W_i) - s_n^{-1}[\boldsymbol{\beta}, \eta_0](t) s_n[\phi_\nu; \boldsymbol{\beta}, \eta_0](t)\}\, dN_i(t)$. The Fourier coefficients that minimize (A.2) are $\tilde\eta_\nu = (\gamma_\nu + \eta_{\nu,0})/(1+\lambda\rho_\nu)$. Note that $\int \phi_\nu(w)\, dw = 0$ and $V(\phi_\nu) = 1$. Straightforward calculation gives $E[\gamma_\nu] = 0$ and $E[\gamma_\nu^2] = n^{-1}$. Then

$$\begin{aligned} E[V(\tilde\eta - \eta_0)] &= \frac{1}{n}\sum_\nu \frac{1}{(1+\lambda\rho_\nu)^2} + \lambda \sum_\nu \frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2}\rho_\nu \eta_{\nu,0}^2, \\ E[\lambda J(\tilde\eta - \eta_0)] &= \frac{1}{n}\sum_\nu \frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2} + \lambda \sum_\nu \frac{(\lambda\rho_\nu)^2}{(1+\lambda\rho_\nu)^2}\rho_\nu \eta_{\nu,0}^2. \end{aligned} \tag{A.3}$$

Combining Lemma A.3 and (A.3), we obtain that $(V + \lambda J)(\tilde\eta - \eta_0) = O_p(\lambda + n^{-1}\lambda^{-1/r})$, as $n \to \infty$ and $\lambda \to 0$.

STEP 2 (Approximation error). We now investigate the approximation error $\hat\eta^* - \tilde\eta$ and prove the convergence rate of $\hat\eta^*$. Define $A_{f,g}(\alpha)$ and $B_{f,g}(\alpha)$, respectively, as the resulting functionals from setting $\eta = f + \alpha g$ in (2.4) and (A.1). Differentiating them with respect to α and then setting $\alpha = 0$ yields

$$\begin{aligned} \dot A_{f,g}(0) = -\frac{1}{n}\sum_{i=1}^n \int_{\mathcal{T}} \{g(W_i) - s_n^{-1}[\boldsymbol{\beta}, f](t) s_n[g; \boldsymbol{\beta}, f](t)\}\, dN_i(t) \\ + \lambda J(f, g), \end{aligned} \tag{A.4}$$

$$\begin{aligned} \dot B_{f,g}(0) = -\frac{1}{n}\sum_{i=1}^n \int_{\mathcal{T}} \{g(W_i) - s_n^{-1}[\boldsymbol{\beta}, \eta_0](t) s_n[g; \boldsymbol{\beta}, \eta_0](t)\}\, dN_i(t) \\ + V(f-\eta_0, g) + \lambda J(f, g). \end{aligned} \tag{A.5}$$

Set $f = \hat{\eta}^*$ and $g = \hat{\eta}^* - \tilde{\eta}$ in (A.4), and set $f = \tilde{\eta}$ and $g = \hat{\eta}^* - \tilde{\eta}$ in (A.5). Then subtracting the resulted equations gives

$$\begin{aligned}(A.6)\qquad &\mu_{\hat{\eta}^*}(\hat{\eta}^* - \tilde{\eta}) - \mu_{\tilde{\eta}}(\hat{\eta}^* - \tilde{\eta}) + \lambda J(\hat{\eta}^* - \tilde{\eta}) \\ &\qquad = V(\tilde{\eta} - \eta_0, \hat{\eta}^* - \tilde{\eta}) + \mu_{\eta_0}(\hat{\eta}^* - \tilde{\eta}) - \mu_{\tilde{\eta}}(\hat{\eta}^* - \tilde{\eta}),\end{aligned}$$

where $\mu_f(g) \equiv \frac{1}{n}\sum_{i=1}^n \int_{\mathcal{T}} s_n^{-1}[\boldsymbol{\beta}, f](t) s_n[g; \boldsymbol{\beta}, f](t)\, dN_i(t)$. Define

$$S_n[f, g](t) = \frac{s_n[fg; \boldsymbol{\beta}, \eta_0](t)}{s_n[\boldsymbol{\beta}, \eta_0](t)} - \frac{s_n[f; \boldsymbol{\beta}, \eta_0](t)}{s_n[\boldsymbol{\beta}, \eta_0](t)} \frac{s_n[g; \boldsymbol{\beta}, \eta_0](t)}{s_n[\boldsymbol{\beta}, \eta_0](t)}$$

and $S[f, g](t)$ be its limit. The following lemma is needed to proceed.

LEMMA A.4.

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \int_{\mathcal{T}} S_n[f, g](t)\, dN_i(t) \\ &\qquad = V(f, g) + o_p\big(\{(V + \lambda J)(f)(V + \lambda J)(g)\}^{1/2}\big).\end{aligned}$$

PROOF. Let $f = \sum_\nu f_\nu \phi_\nu$ and $g = \sum_\mu g_\mu \phi_\mu$ be the Fourier series expansion of f and g. Reference [1] shows that $\sup_t |s_n[\boldsymbol{\beta}, \eta_0] - s[\boldsymbol{\beta}, \eta_0]|$ converges to zero in probability. Note that

$$M(t) \equiv M(t|Z) = N(t) - \int_0^t s[\boldsymbol{\beta}, \eta_0](\tau)\, d\Lambda_0(\tau)$$

defines a local martingale with mean zero. Combining the above uniform convergence result and the martingale property with the boundedness condition, we obtain that for any ν and μ,

$$E\left[\left\{\int_{\mathcal{T}} S[\phi_\nu, \phi_\mu](t)\, dN(t) - V(\phi_\nu, \phi_\mu)\right\}^2\right] < \infty.$$

Then from the Cauchy–Schwarz inequality and Lemma A.3,

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^n \int_{\mathcal{T}} S_n[f, g](t)\, dN_i(t) - V(f, g)\right| \\ &\quad = \left|\sum_\nu \sum_\mu f_\nu g_\mu \left\{\frac{1}{n}\sum_{i=1}^n \int_{\mathcal{T}} S_n[\phi_\nu, \phi_\mu](t)\, dN_i(t) - V(\phi_\nu, \phi_\mu)\right\}\right| \\ &\quad \le \left\{\sum_\nu \sum_\mu \frac{1}{1 + \lambda\rho_\nu}\frac{1}{1 + \lambda\rho_\mu}\right. \\ &\qquad\qquad \left.\times \left\{\frac{1}{n}\sum_{i=1}^n \int_{\mathcal{T}} S_n[\phi_\nu, \phi_\mu](t)\, dN_i(t) - V(\phi_\nu, \phi_\mu)\right\}^2\right\}^{1/2}\end{aligned}$$

$$\times \left\{ \sum_{\nu} \sum_{\mu} (1+\lambda\rho_\nu)(1+\lambda\rho_\mu) f_\nu^2 g_\mu^2 \right\}^{1/2}$$

$$= O_p(n^{-1/2}\lambda^{-1/r})\{(V+\lambda J)(f)(V+\lambda J)(g)\}^{1/2}. \qquad \square$$

A Taylor expansion at η_0 gives

$$\mu_{\hat\eta^*}(\hat\eta^* - \tilde\eta) - \mu_{\tilde\eta}(\hat\eta^* - \tilde\eta) = \frac{1}{n}\sum_{i=1}^{n} \int_{\mathcal{T}} S_n[\hat\eta^* - \tilde\eta, \hat\eta^* - \tilde\eta](t)\, dN_i(t)\big(1 + o_p(1)\big),$$

$$\mu_{\tilde\eta}(\hat\eta^* - \tilde\eta) - \mu_{\eta_0}(\hat\eta^* - \tilde\eta) = \frac{1}{n}\sum_{i=1}^{n} \int_{\mathcal{T}} S_n[\tilde\eta - \eta_0, \hat\eta^* - \tilde\eta](t)\, dN_i(t)\big(1 + o_p(1)\big).$$

Then by the mean value theorem, condition A6, Lemma A.4 and (A.6),

$$(c_1 V + \lambda J)(\hat\eta^* - \tilde\eta)\big(1 + o_p(1)\big)$$
$$\le \{(|1-c|V + \lambda J)(\hat\eta^* - \tilde\eta)\}^{1/2} O_p\big(\{(|1-c|V + \lambda J)(\tilde\eta - \eta_0)\}^{1/2}\big)$$

for some $c \in [c_1, c_2]$. Then the convergence rate of $\hat\eta^*$ follows from that of $\tilde\eta$ proved in the previous step.

STEP 3 (Semiparametric approximation). Our last goal is the convergence rate for the minimizer $\hat\eta$ in the space $\mathcal{H}_n$. For any $h \in \mathcal{H} \ominus \mathcal{H}_n$, one has $h(W_{i_l}) = J(R_J(W_{i_l}, \cdot), h) = 0$, so $s_{q_n}[h^j; \boldsymbol{\beta}, \eta_0](t) = \frac{1}{q_n}\sum_{k=1}^{q_n} Y_{i_k}(t) h^j(W_{i_k}) \exp(U_{i_k}^T \boldsymbol{\beta} + \eta(W_{i_k})) = 0$ for $j = 1, 2$ and $\sum_{l=1}^{q_n} \int_{\mathcal{T}} S_{q_n}[h, h](t)\, dN_{i_l}(t) = 0$. Hence, by the same arguments used in the proof of Lemma A.4,

$$\begin{aligned} V(h) &= \left| \frac{1}{q_n} \sum_{l=1}^{q_n} \int_{\mathcal{T}} S_{q_n}[h, h](t)\, dN_{i_l}(t) - V(h) \right| \\ &= O_p(q_n^{-1/2}\lambda^{-1/r})(V + \lambda J)(h) = o_p(\lambda J(h)), \end{aligned} \tag{A.7}$$

where the last equality follows from $q_n \asymp n^{2/(r+1)+\varepsilon}$ and condition A7.

Let η^* be the projection of $\hat\eta^*$ in $\mathcal{H}_n$. Setting $f = \hat\eta^*$ and $g = \hat\eta^* - \eta^*$ in (A.4) and noting that $J(\eta^*, \hat\eta^* - \eta^*) = 0$, some algebra yields

$$\begin{aligned} \lambda J(\hat\eta^* - \eta^*) = &\left\{ \frac{1}{n}\sum_{i=1}^{n} \int_{\mathcal{T}} (\hat\eta^* - \eta^*)(W_i)\, dN_i(t) - \mu_{\eta_0}(\hat\eta^* - \eta^*) \right\} \\ &- \{\mu_{\hat\eta^*}(\hat\eta^* - \eta^*) - \mu_{\eta_0}(\hat\eta^* - \eta^*)\}. \end{aligned} \tag{A.8}$$

Recall that $\gamma_\nu = \frac{1}{n}\sum_{i=1}^{n} \int_{\mathcal{T}} \phi_\nu(W_i)\, dN_i(t) - \mu_{\eta_0}(\phi_\nu)$ with $E[\gamma_\nu] = 0$ and $E[\gamma_\nu^2] = 1/n$. An application of the Cauchy–Schwarz inequality and Lemma A.3 shows that the first term in (A.8) is of order $\{(V + \lambda J)(\hat\eta^* - \eta^*)\}^{1/2} O_p(n^{-1/2}\lambda^{-1/2r})$. By the mean value theorem, condition A6, Lemma A.4 and (A.7), the remaining term in

(A.8) is of order $o_p(\{\lambda J(\hat{\eta}^* - \eta^*)(V + \lambda J)(\hat{\eta}^* - \eta_0)\}^{1/2})$. These, combined with (A.8) and the convergence rates of $\hat{\eta}^*$, yield $\lambda J(\hat{\eta}^* - \eta^*) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$ and $V(\hat{\eta}^* - \eta^*) = o_p(n^{-1}\lambda^{-1/r} + \lambda)$.

Note that $J(\hat{\eta}^* - \eta^*, \eta^*) = J(\hat{\eta}^* - \eta^*, \hat{\eta}) = 0$, so $J(\hat{\eta}^*, \hat{\eta}^* - \hat{\eta}) = J(\hat{\eta}^* - \eta^*) + J(\eta^*, \eta^* - \hat{\eta})$. Set $f = \hat{\eta}$ and $g = \hat{\eta} - \eta^*$ in (A.4), and set $f = \hat{\eta}^*$ and $g = \hat{\eta}^* - \hat{\eta}$ in (A.4). Adding the resulted equations yields

$$\begin{aligned}
&\mu_{\hat{\eta}}(\hat{\eta} - \eta^*) - \mu_{\eta_0}(\hat{\eta} - \eta^*) + \lambda J(\hat{\eta} - \eta^*) + \lambda J(\hat{\eta}^* - \eta^*) \\
&\quad = \left\{ \frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{T}} (\hat{\eta}^* - \eta^*)(W_i)\, dN_i(t) - \mu_{\eta_0}(\hat{\eta}^* - \eta^*) \right\} \\
&\qquad - \{\mu_{\hat{\eta}^*}(\hat{\eta}^* - \eta^*) - \mu_{\eta_0}(\hat{\eta}^* - \eta^*)\} \\
&\qquad + \{\mu_{\hat{\eta}^*}(\hat{\eta} - \eta^*) - \mu_{\eta_0}(\hat{\eta} - \eta^*)\}.
\end{aligned} \tag{A.9}$$

By the mean value theorem, condition A6, and Lemma A.4, the left-hand side of (A.9) is bounded from below by

$$(c_1 V + \lambda J)(\hat{\eta} - \eta^*)\big(1 + o_p(1)\big) + \lambda J(\hat{\eta}^* - \eta^*).$$

For the right-hand side, the terms in the first and second brackets are, respectively, of the orders $\{(V + \lambda J)(\hat{\eta}^* - \eta^*)\}^{1/2} O_p(n^{-1/2}\lambda^{-1/2r})$ and $o_p(\{\lambda J(\hat{\eta}^* - \eta^*)(V + \lambda J)(\hat{\eta}^* - \eta_0)\}^{1/2})$ by similar arguments for (A.8), and the terms in the third bracket is of the order

$$\{(V + \lambda J)(\hat{\eta} - \eta^*)\}^{1/2} o_p\big(\{\lambda J(\hat{\eta}^* - \eta^*)\}^{1/2}\big)$$

by condition 3, Lemma A.4 and (A.7). Putting all these together, one obtains $(V + \lambda J)(\hat{\eta} - \eta^*) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$ and hence $(V + \lambda J)(\hat{\eta} - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$. And an application of condition A7 yields the final convergence rates. □

PROOF FOR THE ASYMPTOTIC PROPERTIES OF $\hat{\boldsymbol{\beta}}$. Let P_n be the empirical measure of $(X_i, \Delta_i = 1, Z_i), i = 1, \ldots, n$ such that it is related to the empirical measure Q_n of $(X_i, \Delta_i, Z_i), i = 1, \ldots, n$ by $P_n f = \int f\, dP_n = \int \Delta f\, dQ_n = n^{-1} \sum_{i=1}^{n} \Delta_i f(T_i, \Delta_i, Z_i)$. Let P be its corresponding (sub)probability measure. Let $L_2(P) = \{f : \int f^2\, dP < \infty\}$ and $\|\cdot\|_2$ be the usual L_2-norm. For any subclass $\mathcal{F}$ of $L_2(P)$ and any $\varepsilon > 0$, let $\mathcal{N}_{[\cdot]}(\varepsilon, \mathcal{F}, L_2(P))$ be the bracketing number and $J_{[\cdot]}(\delta, \mathcal{F}, L_2(P)) = \int_0^\delta \sqrt{1 + \log \mathcal{N}_{[\cdot]}(\varepsilon, \mathcal{F}, L_2(P))}\, d\varepsilon$.

LEMMA A.5. *Let* $m_0(t, u, w; \boldsymbol{\beta}, \eta) = u^T \boldsymbol{\beta} + \eta(w) - \log s[\boldsymbol{\beta}, \eta](t)$, $m_1(t, u, w;\ s, \boldsymbol{\beta}, \eta) = 1_{[s \le t]} \exp(u^T \boldsymbol{\beta} + \eta(w))$, *and* $m_2(t, u, w; s, \boldsymbol{\beta}, \eta, f) = 1_{[s \le t]} f(u, w) \exp(u^T \boldsymbol{\beta} + \eta(w))$. *Define the classes of functions*

$$\begin{aligned}
\mathcal{M}_0(\delta) &= \{m_0 : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le \delta, \|\eta - \eta_0\|_2 \le \delta\}, \\
\mathcal{M}_1(\delta) &= \{m_1 : s \in \mathcal{T}, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le \delta, \|\eta - \eta_0\|_2 \le \delta\}, \\
\mathcal{M}_2(\delta) &= \{m_2 : s \in \mathcal{T}, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \le \delta, \|\eta - \eta_0\|_2 \le \delta, \|h\|_2 \le \delta\}.
\end{aligned}$$

Then $J_{[\cdot]}(\delta, \mathcal{M}_0, L_2(P)) \le c_0 q_n^{1/2}\delta$ *and* $J_{[\cdot]}(\delta, \mathcal{M}_j, L_2(P)) = c_j\delta\{q_n + \log(1/\delta)\}^{1/2}$ $j = 1, 2$.

PROOF. The proof is similar to that of Corollary A.1 in [10] and thus omitted here. □

LEMMA A.6.

$$\sup_{t\in\mathcal{T}} |s^{-1}[\boldsymbol{\beta}_0, \hat{\eta}](t)s[U; \boldsymbol{\beta}_0, \hat{\eta}](t) - s_n^{-1}[\boldsymbol{\beta}_0, \hat{\eta}](t)s_n[U; \boldsymbol{\beta}_0, \hat{\eta}](t)| = o_p(n^{-1/2}).$$

PROOF. Write

$$\begin{aligned} &s^{-1}[\boldsymbol{\beta}_0, \hat{\eta}](t)s[U; \boldsymbol{\beta}_0, \hat{\eta}](t) - s_n^{-1}[\boldsymbol{\beta}_0, \hat{\eta}](t)s_n[U; \boldsymbol{\beta}_0, \hat{\eta}](t) \\ &\quad = \frac{s[\boldsymbol{\beta}_0, \hat{\eta}](t)A_{1n}(t) - s[U; \boldsymbol{\beta}_0, \hat{\eta}](t)A_{2n}(t)}{s[\boldsymbol{\beta}_0, \hat{\eta}](t)s_n[\boldsymbol{\beta}_0, \hat{\eta}](t)}, \end{aligned}$$

where $A_{1n} = s_n[U; \boldsymbol{\beta}_0, \hat{\eta}](t) - s[U; \boldsymbol{\beta}_0, \hat{\eta}](t)$ and $A_{2n} = s_n[\boldsymbol{\beta}_0, \hat{\eta}](t) - s[\boldsymbol{\beta}_0, \hat{\eta}](t)$. Note that $q_n = o(n^{1/2})$, hence the result follows from Lemma 3.4.2 of [27] and Lemma A.5. □

PROOF OF THEOREM 2.2. Let $\gamma_n = n^{-1/2}$. To prove 2.2(i), we need to show that $\forall\delta > 0$, there exists a large constant C such that

$$P\Big\{\sup_{\|v\|=C} \mathcal{L}_P(\boldsymbol{\beta}_0 + \gamma_n v) < \mathcal{L}_P(\boldsymbol{\beta}_0)\Big\} \ge 1 - \delta.$$

Consider $\mathcal{L}_P(\boldsymbol{\beta}_0 + \gamma_n v) - \mathcal{L}_P(\boldsymbol{\beta}_0)$. We can decompose it to the sum of $D_{n1} = l_p(\boldsymbol{\beta}_0 + \gamma_n v) - l_p(\boldsymbol{\beta}_0)$ and the penalty difference D_{n2}. As shown in [7], under the assumption of $a_n = O(n^{-1/2})$ and $b_n = o(1)$, $n^{-1}D_{n2}$ is bounded by

$$\sqrt{s}\gamma_n a_n\|v\| + \gamma_n^2 b_n\|v\|^2 = C\gamma_n^2(\sqrt{s} + b_n C), \tag{A.10}$$

where s is the number of nonzero elements in $\boldsymbol{\beta}_0$.

Applying the second order Taylor expansion to $n^{-1}D_{n1}$, gives

$$n^{-1}D_{n1} = \gamma_n \mathbf{v}^T \mathbf{J}_{1n} - \tfrac{1}{2}\gamma_n^2 \mathbf{v}^T \mathbf{J}_{2n}\mathbf{v} + o_p(n^{-1}) \tag{A.11}$$

with $\mathbf{J}_{1n} = P_n\{U - s_n^{-1}[\boldsymbol{\beta}_0, \hat{\eta}](t)s_n[U; \boldsymbol{\beta}_0, \hat{\eta}](t)\}$, and $\mathbf{J}_{2n} = P_n\{s_n^{-2}[\boldsymbol{\beta}_0, \hat{\eta}](t) \times [s_n[UU^T; \boldsymbol{\beta}_0, \hat{\eta}](t)s_n[\boldsymbol{\beta}_0, \hat{\eta}](t) - s_n[U; \boldsymbol{\beta}_0, \hat{\eta}](t)s_n[U; \boldsymbol{\beta}_0, \hat{\eta}](t)^T]\}$, where $U(u, w) \equiv u$.

Let $\bar{U}_n = \sum_{i=1}^n U_i/n$. Note that $s_n^{-1}[\boldsymbol{\beta}_0, \hat{\eta}](t)s_n[\bar{U}_n; \boldsymbol{\beta}_0, \hat{\eta}](t) = \bar{U}_n$. We have

$$\begin{aligned} \mathbf{J}_{1n} &= P_n\{U - \bar{U}_n - s_n^{-1}[\boldsymbol{\beta}_0, \hat{\eta}](t)s_n[U - \bar{U}_n; \boldsymbol{\beta}_0, \hat{\eta}](t)\} \\ &\equiv I_{1n} + I_{2n} + I_{3n}, \end{aligned} \tag{A.12}$$

where

$$I_{1n} = (P_n - P)\{U - \bar{U}_n - s^{-1}[\boldsymbol{\beta}_0, \hat{\eta}](t)s[U - \bar{U}_n; \boldsymbol{\beta}_0, \hat{\eta}](t)\},$$
$$I_{2n} = P_n\{s^{-1}[\boldsymbol{\beta}_0, \hat{\eta}](t)s[U; \boldsymbol{\beta}_0, \hat{\eta}](t) - s_n^{-1}[\boldsymbol{\beta}_0, \hat{\eta}](t)s_n[U; \boldsymbol{\beta}_0, \hat{\eta}](t)\},$$
$$I_{3n} = P\{U - \bar{U}_n - s^{-1}[\boldsymbol{\beta}_0, \hat{\eta}](t)s[U - \bar{U}_n; \boldsymbol{\beta}_0, \hat{\eta}](t)\}.$$

Lemma 3.4.2 of [27] and Lemma A.5 indicate that $(P_n - P)\{s^{-1}[\boldsymbol{\beta}_0, \hat{\eta}](t)s[U - \bar{U}_n; \boldsymbol{\beta}_0, \hat{\eta}](t)\} = o_p(n^{-1/2})$, where the fact that $q_n = o_p(n^{-1/2})$ is used again. Also $(P_n - P)\{U - \bar{U}_n\} = O_p(n^{-1/2})$ by the LLN. Hence, we have $I_{1n} = O_p(n^{-1/2})$. Lemma A.6 gives $I_{2n} = o_p(n^{-1/2})$. Finally, by the boundedness assumption $I_{3n} = P\{s^{-1}[\boldsymbol{\beta}_0, \hat{\eta}](t)s[\bar{U}_n - U; \boldsymbol{\beta}_0, \hat{\eta}](t)\} = O_p(E|\bar{U}_n - U|) = O_p(n^{-1/2})$. Hence, $\mathbf{J}_{1n} = O_p(n^{-1/2})$. Also, $\mathbf{J}_{2n}$ converges to $V(U) > 0$. Thus, when C is sufficiently large, the second term in (A.11) dominates both terms in (A.10). Theorem 2.2(i) follows.

Next, we shall show the sparsity of $\hat{\boldsymbol{\beta}}$. It suffices to show that for any given $\boldsymbol{\beta}_1$ satisfying $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_p(n^{-1/2})$ and any $j = s+1, \ldots, d$, $\partial \mathcal{L}_P(\boldsymbol{\beta})/\partial\beta_j > 0$ for $0 < \beta_j < Cn^{-1/2}$ and $\partial \mathcal{L}_P(\boldsymbol{\beta})/\partial\beta_j < 0$ for $-Cn^{-1/2} < \beta_j < 0$. For $\beta_j \neq 0$ and $j = s+1, \ldots, d$,

$$n^{-1}\partial \mathcal{L}_P(\boldsymbol{\beta})/\partial\beta_j = P_n\{U_j - s_n^{-1}[\boldsymbol{\beta}, \hat{\eta}](t)s_n[U_j; \boldsymbol{\beta}, \hat{\eta}](t)\} - p'_{\theta_j}(|\beta_j|)\operatorname{sgn}(\beta_j).$$

Similar to bounding $\mathbf{J}_{1n}$, the first term can be shown to be $O_p(n^{-1/2})$. Recall that $\theta_j^{-1} = o(n^{1/2})$ and $\liminf_{n\to\infty} \liminf_{u\to 0^+} \theta_j^{-1} p'_{\theta_j}(u) > 0$. Hence, the sign of $\partial \mathcal{L}_P(\boldsymbol{\beta})/\partial\beta_j$ is completely determined by that of β_j. Then $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$.

Lastly, we show the asymptotic normality of $\hat{\boldsymbol{\beta}}_1$ using the result in [21]. Let $z_i = (X_i, \Delta_i, Z_i)$. Note that $\hat{\boldsymbol{\beta}}_1$ is the solution of the estimating equation

$$\sum_{i=1}^{n} M(z_i, \boldsymbol{\beta}_1, \hat{\eta}) - n\boldsymbol{\zeta}_1 = \mathbf{0}, \tag{A.13}$$

where $M(z, \boldsymbol{\beta}_1, \eta) = \int\{U_1 - s_n^{-1}[\boldsymbol{\beta}_1, \eta](t)s_n[U_1; \boldsymbol{\beta}_1, \eta](t)\}\,dN(t)$ and $\boldsymbol{\zeta}_1 = (p'_{\theta_1}(|\beta_1|)\operatorname{sgn}(\beta_1), \ldots, p'_{\theta_s}(|\beta_s|)\operatorname{sgn}(\beta_s))^T$. Let

$$D(z, h) = \int \left\{ \frac{s_n[U_1 h; \boldsymbol{\beta}_{10}, \eta_0](t)}{s_n[U_1 h; \boldsymbol{\beta}_{10}, \eta_0](t)} - \frac{s_n[U_1; \boldsymbol{\beta}_{10}, \eta_0](t)}{s_n[\boldsymbol{\beta}_{10}, \eta_0](t)} \frac{s_n[h; \boldsymbol{\beta}_{10}, \eta_0](t)}{s_n[\boldsymbol{\beta}_{10}, \eta_0](t)} \right\} dN(t)$$

be the Fréchet derivative of $M(z, \boldsymbol{\beta}_{10}, \eta)$ at η_0. Since the convergence rate of $\hat{\eta}$ is $n^{-r/[2(r+1)]} = o(n^{-1/4})$, the linearization assumption (Assumption 5.1) in [21] is satisfied. A derivation similar to bounding (A.12) can verify the stochastic assumption (Assumption 5.2) in [21]. Direct calculation yields $E[D(z, \eta - \eta_0)] = 0$ for η close to η_0. Then the mean-square continuity assumption (Assumption 5.3) in [21] also holds with $\alpha(z) \equiv 0$. By Lemma 5.1 in [21], $\hat{\boldsymbol{\beta}}_1$ thus has the same distribution as the solution to the equation

$$\sum_{i=1}^{n} M(z_i, \boldsymbol{\beta}_1, \eta_0) - n\boldsymbol{\zeta}_1 = \mathbf{0}.$$

A straightforward simplification yields the result. □

Acknowledgments. We would like to thank the Associate Editor and two referees for their insightful comments that have improved the article.

REFERENCES

[1] 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

[2] BREIMAN, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.* **24** 2350–2383. MR1425957

[3] CAI, J., FAN, J., JIANG, J. and ZHOU, H. (2007). Partially linear hazard regression for multivariate survival data. *J. Amer. Statist. Assoc.* **102** 538–551. MR2370851

[4] CAI, J., FAN, J., LI, R. and ZHOU, H. (2005). Variable selection for multivariate failure time data. *Biometrika* **92** 303–316. MR2201361

[5] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32** 407–499. MR2060166

[6] FAN, J., GIJBELS, I. and KING, M. (1997). Local likelihood and local partial likelihood in hazard regression. *Ann. Statist.* **25** 1661–1690. MR1463569

[7] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[8] FAN, J. and LI, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30** 74–99. MR1892656

[9] GU, C. (2002). *Smoothing Spline ANOVA Models*. Springer, New York, MR1876599

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

[11] HUANG, J. Z., KOOPERBERG, C., STONE, C. J. and TRUONG, Y. K. (2000). Functional ANOVA modeling for proportional hazards regression. *Ann. Statist.* **28** 961–999. MR1810916

[12] HUANG, J. Z. and LIU, L. (2006). Polynomial spline estimation and inference of proportional hazards regression models with flexible relative risk form. *Biometrics* **62** 793–802. MR2247208

[13] JOHNSON, B. A. (2008). Variable selection in semiparametric linear regression with censored data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 351–370. MR2424757

[14] JOHNSON, B. A., LIN, D. Y. and ZENG, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *J. Amer. Statist. Assoc.* **103** 672–680. MR2435469

[15] KALBFLEISCH, J. D. and PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley, New York. MR1924807

[16] KIM, Y.-J. and GU, C. (2004). Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 337–356. MR2062380

[17] KLEIN, J. P. and MOESCHBERGER, M. L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York.

[18] LENG, C. and ZHANG, H. H. (2006). Model selection in nonparametric hazard regression. *J. Nonparametr. Stat.* **18** 417–429. MR2311796

[19] LIN, Y. and ZHANG, H. H. (2006). Component selection and smoothing in smoothing spline analysis of variance models. *Ann. Statist.* **34** 2272–2297. MR2291500

[20] MURPHY, S. A. and VAN DER VAART, A. W. (2000). On profile likelihood (with discussion). *J. Amer. Statist. Assoc.* **95** 449–465. MR1803168

[21] NEWEY, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica* **62** 1349–1382. MR1303237
[22] OSBORNE, M. R., PRESNELL, B. and TURLACH, B. A. (2000). On the LASSO and its dual. *J. Comput. Graph. Statist.* **9** 319–337. MR1822089
[23] O'SULLIVAN, F. (1993). Nonparametric estimation in the Cox model. *Ann. Statist.* **21** 124–145. MR1212169
[24] PARK, T. and CASELLA, G. (2008). The Bayesian Lasso. *J. Amer. Statist. Assoc.* **103** 681–686.
[25] TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242
[26] TIBSHIRANI, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* **16** 385–395.
[27] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York. MR1385671
[28] WAHBA, G. (1990). *Spline Models for Observational Data*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. SIAM, Philadelphia. MR1045442
[29] WEINBERGER, H. F. (1974). *Variational Methods for Eigenvalue Approximation*. SIAM, Philadelphia. MR0400004
[30] YIN, G., LI, H. and ZENG, D. (2008). Partially linear additive hazards regression with varying coefficients. *J. Amer. Statist. Assoc.* **103** 1200–1213. MR2462893
[31] ZOU, H. (2006). The adaptive LASSO and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469
[32] ZOU, H. (2008). A note on path-based variable selection in the penalized proportional hazards model. *Biometrika* **95** 241–247. MR2409726
[33] ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2007). On the "degree of freedom" of the LASSO. *Ann. Statist.* **35** 2173–2192. MR2363967
[34] ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.* **36** 1509–1533. MR2435443
[35] ZUCKER, D. M. and KARR, A. F. (1990). Nonparametric survival analysis with time-dependent covariate effects: A penalized partial likelihood approach. *Ann. Statist.* **18** 329–353. MR1041396

P. DU
DEPARTMENT OF STATISTICS
VIRGINIA TECH
BLACKSBURG, VIRGINIA 24061
USA
E-MAIL: pangdu@vt.edu

S. MA
DEPARTMENT OF EPIDEMIOLOGY
AND PUBLIC HEALTH
YALE UNIVERSITY
SCHOOL OF MEDICINE
NEW HAVEN, CONNECTICUT 06520
USA
E-MAIL: shuangge.ma@yale.edu

H. LIANG
DEPARTMENT OF BIOSTATISTICS
AND COMPUTATIONAL BIOLOGY
UNIVERSITY OF ROCHESTER
ROCHESTER, NEW YORK 14642
USA
E-MAIL: hliang@bst.rochester.edu