The Canadian Journal of Statistics
Vol. xx, No. yy, 20??, Pages 1–25
La revue canadienne de statistique

Variable Selection and Structure Estimation for Ultrahigh-Dimensional Additive Hazards Models

Li Liu[1], Yanyan Liu[1], Feng Su[2] and Xingqiu Zhao[3]*

[1]*School of Mathematics and Statistics, Wuhan University, Wuhan, China*
[2]*Division of Mathematics, Guangzhou Maritime Institute, Guangzhou, China*
[3]*Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong*

Key words and phrases: Additive hazards regression ; Ultrahigh-dimensional censored data ; Penalized sieve least squares ; Model pursuit ; Variable selection.

MSC 2010: Primary 62N01 ; 62N02 secondary 62F12

Abstract: We develop a class of regularization methods based on the penalized sieve least squares for simultaneous model pursuit, variable selection, and estimation in high-dimensional additive hazards regression models. In the framework of sparse ultrahigh-dimensional models, the asymptotic properties of the estimators including the structure identification consistency and the variable selection oracle properties are established. The computational process can be efficiently implemented by applying the blockwise majorization descent algorithm. Simulation studies demonstrate the performance of the proposed methodology, and the primary biliary cirrhosis data analysis is provided for illustration. *The Canadian Journal of Statistics* xx: 1–25; 20?? © 20?? Statistical Society of Canada

1. INTRODUCTION

Survival models have been extensively used in the fields such as biomedicine, finance and economics with the goal of assessing the effectiveness of predictors on event times of interest. However, when the dimension of covariates p is large compared to sample size n, e.g., $p = O(n^{\alpha_1})$ or $p = O(\exp(n^{\alpha_2}))$ for some positive constants α_1 and α_2, the "curse of dimensionality" makes the traditional statistical analysis methods infeasible. As a tool of alleviating problems of high dimensional data, variable selection aims to select a subset composing of sparse important variables from a huge number of covariates, and then enhances the efficiency of estimates and improves the predictive power. Among various variable selection techniques, regularization methods have gained more popularity since this type of methods can identify important variables and estimate the parameters simultaneously. Examples include but not limited to the least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani (1996), adaptive LASSO studied by Zou (2006), the smoothly clipped absolute deviation (SCAD) penalty explored by Fan & Li (2001), and the minimum concave penalty (MCP) considered by Zhang (2010). These selectors were then extended to the Cox proportional hazards model by many researchers, such as Tibshirani (1997), Fan & Li (2002), and Zhang & Lu (2007). Different from the Cox model which concerns the relative risk, an additive hazards model describes the risk difference. Due to its easy interpretation, this model has been studied in many literatures, such as Lin & Ying (1994), Ma, Kosorok, & Fine (2006), and Xie, Strickler, & Xue (2013). Besides, for additive hazards regression models under a fixed dimensional setting, Leng & Ma (2007) used the weighted LASSO

** Author to whom correspondence may be addressed.*
E-mail: xingqiu.zhao@polyu.edu.hk

approach to obtain a path consistent model selector, Martinussen & Scheike (2009) considered several regularization methods such as the LASSO, adaptive LASSO and Dantzig selector; under a high-dimensional setting, Lin& Lv (2013) proposed regularization methods based on the non-concave penalized likelihood approach, Zhang, Cheng, & Liu (2017) studied the properties of the weighted LASSO, and Wang & Xiang (2017) presented the penalized empirical likelihood inference procedure.

The above referred work assumed that covariates have only linear effects on the hazard function of the survival time. This assumption may cause a seriously biased estimation problem if a nonparametric component is misspecified as a linear part. On the other hand, the model would be more complex and lose efficiency if parametric parts are treated as nonparametric. Lee et al. (2015), Kong et al. (2018), and Hao et al. (2020) studied a functional Cox model with a known structure. In practice, how to identify the model structure is critical in the process of the statistical inference, and this issue has drawn an extensive attention recently. Zhang, Cheng, & Liu (2011) and Huang, Wei, & Ma (2012) studied a model pursuit problem for a partial linear model. Lian, Lai, & Liang (2013) and Cao et al. (2016) considered the problem of simultaneous structure selection and estimation for the Cox model under a fixed dimensional setting. Yan & Huang (2012) and Bradic & Song (2015) proposed hierarchical group penalties to identify the structure of a varying coefficient Cox model. Furthermore, Honda & Yabe (2017) used the orthogonal penalty to study a problem of variable selection and structure identification for a varying coefficient Cox model.

The Cox model assumes that effects of covariates on the hazard function are multiplicative. When this assumption is violated, an additive hazards model may be more appropriate and feasible. However, to the best of our knowledge, the problem of simultaneous variable selection, model identification and estimation has not been studied for high-dimensional additive hazards models with censored data. On the other hand, it was assumed in Honda & Yabe (2017) that the number of important variables should be finite and $p \sim \exp(n^{\alpha})$ with $n^{\alpha} = o(n^{2/5})$, which may be restrictive in high-dimensional settings. These motivate us to provide an automatic procedure for variable selection, model structure identification and estimation simultaneously in high-dimensional additive hazards models and to establish the asymptotic properties of the proposed estimators under some weaker conditions.

The main contributions of our work are threefold. First, we give milder conditions such that additive hazards models with an unknown structure are identifiable in the process of statistical inference compared to Zhang, Cheng, & Liu (2011). Second, we develop a class of regularization approaches for simultaneous model pursuit, variable selection, and estimation in high-dimensional additive hazards models with censored data by adopting the orthogonal decomposition approach. Third, we establish the asymptotic properties of the estimators including both the consistency of structure identification and the oracle properties of variable selection using the modern empirical process theory. Our methods allow the number of important variables to be diverging and the model dimensionality to be exponentially increasing. In particular, the conditions required in the existing methods are relaxed, and thus the proposed approaches are more applicable to general situations in high-dimensional survival data analysis.

The remainder of the paper is organized as follows. In Section 2, we propose the inference method of orthogonal sieve estimation with penalties, and present the blockwise majorization descent algorithm. The asymptotic properties of the estimates are established in Section 3. We then present simulation studies in Section 4 to evaluate the finite-sample performance of some competing penalized estimates, and in Section 5, we apply the proposed procedure to the primary biliary cirrhosis data analysis. Some concluding remarks are made in Section 6. The proofs of the main results are relegated to the Appendix.

2. ESTIMATION PROCEDURE

2.1. Model setting

We denote $\boldsymbol{Z}(t) = (Z_1(t), \dots, Z_p(t))^T$, $\boldsymbol{\phi} = (\phi_1, \dots, \phi_p)^T$ and $\boldsymbol{g} = (g_1, \dots, g_p)^T$ as $p-$dimensional vectors. Suppose that T^U denotes a failure time satisfying the following additive hazards model:

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t) + \sum_{j=1}^{p} g_j(\tilde{\phi}_{0j}, t), \tag{1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, $\tilde{\phi}_{0j}, j = 1, \dots, p$ are unknown functions, and $g_j(\tilde{\phi}_{0j}, t) = Z_j(t)\tilde{\phi}_{0j}(Z_j(t))$ with $Z_j(t)$ being the jth uniformly bounded covariate. This kind of model setting takes a low cost to explain a nonparametric additive effect of covariate Z_j on one hand, and it ensures the selection process identifiable on the other hand. In practice, covariate Z_j is unimportant if and only if $\tilde{\phi}_{0j}(\cdot) = 0$; if not, the covariate has a linear effect on the hazard function only when $\tilde{\phi}_{0j}(\cdot)$ equals a non-zero constant, and it is a nonparametric component otherwise. Thus, the problem of model pursuit and variable selection turns into the discriminating problem on function $\tilde{\phi}_{0j}(\cdot)$.

Let C be a censoring time, $T = T^U \wedge C$ be a observed time, and $\Delta = I(T^U \leq C)$ where $I(\cdot)$ is an indicator function. We assume that the failure time T^U and the censoring time C are independent given covariate $\boldsymbol{Z}(\cdot)$. Then the observed data consist of $(T_i, \Delta_i, \boldsymbol{Z}_i(\cdot))$ for individual $i = 1, 2, \dots, n$. Define the observed failure counting process as $N_i(t) = I(T_i \leq t, \Delta_i = 1)$ and the at-risk indicator $Y_i(t) = I(T_i \geq t)$. In this paper, we suppose that the true model satisfies the following partially linear model with

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t) + \sum_{j=1}^{s_1} \beta_{0j} Z_j(t) + \sum_{j=s_1+1}^{s} Z_j(t)\phi_{0j}(Z_j(t)).$$

This true model implies that the first s variables are important, among which the first s_1 components are linear effect parts and the last s_2 ones are nonlinear parts with $s_2 = s - s_1$. Correspondingly, we set $\mathcal{A} = \{j : 1 \leq j \leq s\}$, $\mathcal{B} = \{j : 1 \leq j \leq s_1\}$, $\mathcal{C} = \{j : s_1 + 1 \leq j \leq s\}$, and $\mathcal{D} = \mathcal{A}^c$ as the index sets of important variables, important linear components, important nonlinear components, and unimportant components, respectively. We denote the true values of the parametric and nonparametric parts as $\boldsymbol{\beta}_0 = (\beta_{0j}, j \in \mathcal{B})$ and $\boldsymbol{\phi}_0 = (\phi_{0j}, j \in \mathcal{C})$. We use $\|\cdot\|$ and $\|\cdot\|_\infty$ to represent L_2-norm and $L_\infty-$norm, and assume that the follow-up time period is $[0, \tau]$ and $E\int_0^\tau g_j(\tilde{\phi}_{0j}, t)dt = 0$. Without loss of generality, we suppose that the covariate $\boldsymbol{Z}(t)$ takes values on $[a, b]^p$ with a density function $f_t(z_1, \dots, z_p)$, where $\int_0^\tau f_t(z_1, \dots, z_p)dt > 0$, and a and b are finite real numbers.

The following proposition shows that $\sum_{j=1}^{p} g_j(\tilde{\phi}_{0j}, t)$ with $\tilde{\phi}_{0j} \in L^2[a, b]$ can be identified into linear and nonlinear components.

Proposition 1 *(Identifiability)* *(i) Let $G(\cdot)$ be a Borel function on $[a, b]^p$. Then $G(Z_1(t), \dots, Z_p(t)) = 0$ for any $t \in [0, \tau]$ implies that $G(z_1, \dots, z_p) = 0$ for $(z_1, \dots, z_p) \in [a, b]^p$ when $\int_0^\tau f_t(z_1, \dots, z_p)dt > 0$;*

(ii) Suppose that g_j's $j = 1, \dots, p$ are Borel functions on $[a, b]$ satisfying $\sum_{j=1}^{p} g_j(z_j) = 0$ for $(z_1, \dots, z_p) \in [a, b]^p$. Then there exist constants C_j's such that $g_j = C_j$, $j = 1, \dots, p$;

(iii) For any function $\tilde{\phi}_j \in L^2[a,b]$*, there exists a unique* $\beta_j \in \mathbb{R}$ *and* $\phi_j \in \mathcal{H} = \{\phi : \int_a^b \phi(x)dx = 0, \phi \in L^2[a,b]\}$ *such that*

$$\sum_{j=1}^{p} Z_j(t)\tilde{\phi}_j(Z_j(t)) = \sum_{j=1}^{p} \beta_j Z_j(t) + \sum_{j=1}^{p} Z_j(t)\phi_j(Z_j(t)) \tag{2}$$

with $E\Big(\int_0^\tau Z_j(t)\tilde{\phi}_j(Z_j(t))dt\Big) = E\Big(\int_0^\tau [\beta_j Z_j(t) + Z_j(t)\phi_j(Z_j(t))]dt\Big)$ *for each* $j = 1, \ldots, p$.

2.2. Sieve estimation

Let k be a nonnegative integer such that $\varsigma = k + \alpha > 0.5$ for some $\alpha \in (0,1]$. We define $\phi^{(k)}$ as the kth derivative of function ϕ and let

$$\tilde{\mathcal{G}} = \{\phi : |\phi(x_1) - \phi(x_2)| \leq C|x_1 - x_2|^\alpha, x_1, x_2 \in [a,b], \phi^{(k)} \in L^2[a,b]\} \subset L^2[a,b].$$

Throughout the paper, we suppose that the kth derivative $\tilde{\phi}_{0j}^{(k)}$ exists and $\tilde{\phi}_{0j} \in \tilde{\mathcal{G}}$, $j = 1, \ldots, p$.

We use the sieve estimation to approximate the unknown function $\tilde{\phi}_{0j}$'s in (1). The interval $[a,b]$ is divided into $K_n = O(n^\nu)$ subintervals $I_{K_n j} = [\xi_j, \xi_{j+1}), j = 0, \ldots, K_n - 1$ and $I_{K_n K_n} = [\xi_{K_n}, \xi_{K_n+1}]$, where $\xi_0 = a$, $\xi_{K_n+1} = b$, and $0 < \nu < 0.5$ such that $\max_{1 \leq j \leq K_n+1} |\xi_k - \xi_{k-1}| = O(n^{-\nu})$. Denote $\mathcal{S}_n$ to be the space of polynomial splines of order $m \geq 1$ consisting of functions l, where l is a polynomial of order m on interval $I_{K_n j}$ for $1 \leq j \leq K_n$, and l is m' times continuously differentiable on $[a,b]$ for $m \geq 2$ and $0 \leq m' \leq m - 2$. According to Schumaker (1981), there exists a local basis $\bar{\boldsymbol{B}} = \{\bar{B}_k, 1 \leq k \leq q_n\}$ such that for any $\phi_{nj} \in \mathcal{S}_n$,

$$\phi_{nj}(Z_j(t)) = \sum_{k=1}^{q_n} \bar{\theta}_{jk}\bar{B}_k(Z_j(t)), \tag{3}$$

where $q_n = K_n + m$. To identify linear components from nonlinear parts, following the idea of Honda & Yabe (2017), we introduce a q_n-dimensional equispaced orthogonal B-spline basis $\boldsymbol{B} = A\bar{\boldsymbol{B}} = (B_k, 1 \leq k \leq q_n)$ such that

$$\int_0^\tau \boldsymbol{B}(Z_j(t))\boldsymbol{B}(Z_j(t))^T dt = q_n^{-1} I_{q_n},$$

where $B_1(\cdot) = q_n^{-1/2}$ reflects the base function of constant coefficients, I_{q_n} is the q_n-order identity matrix, and the matrix A can be constructed as in Honda & Yabe (2017). Thus, (3) can be rewritten as

$$\phi_{nj}(Z_j(t)) = \sum_{k=1}^{q_n} \theta_{jk}B_k(Z_j(t)) = \boldsymbol{\theta}_j^T \boldsymbol{B}(Z_j(t)),$$

where $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jq_n})^T$ with $\boldsymbol{\theta}_j = (\theta_{j1}, \boldsymbol{\theta}_{j1-})^T$. Then ϕ_{nj} is an appropriate approximation for $\tilde{\phi}_{0j}$ in that there exists some $\phi_{nj} \in \mathcal{S}_n$ close enough to $\tilde{\phi}_{0j}$ for any $\tilde{\phi}_{0j} \in \tilde{\mathcal{G}}$ according to Schumaker (1981), and the linear and nonlinear effects on the jth convariate can be decomposed into θ_{j1} and $\boldsymbol{\theta}_{j1-}$, respectively.

We define $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_p^T)^T$ and $\boldsymbol{X}_i(t) = \boldsymbol{B}(\boldsymbol{Z}_i(t))\boldsymbol{Z}_i(t)$, where $\boldsymbol{B}(\boldsymbol{Z}_i(t)) = \mathrm{d}iag\big(\boldsymbol{B}(Z_{i1}(t)), \ldots, \boldsymbol{B}(Z_{ip}(t))\big)$ is a $pq_n \times p$ block diagonal matrix. Then following Lin & Ying

(1994), the estimated regression coefficient can be obtained by solving the pseudoscore estimating equation

$$\boldsymbol{U}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}(\boldsymbol{X}_i(t) - \overline{\boldsymbol{X}}(t))\Big(dN_i(t) - Y_i(t)\boldsymbol{\theta}^T\boldsymbol{X}_i(t)dt\Big) = 0,$$

where $\overline{\boldsymbol{X}}(t) = \sum\limits_{i=1}^{n} Y_i(t)\boldsymbol{X}_i(t)/\sum\limits_{i=1}^{n} Y_i(t)$. Let

$$\boldsymbol{b} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}(\boldsymbol{X}_i(t) - \overline{\boldsymbol{X}}(t))dN_i(t) \text{ and } \boldsymbol{V} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} Y_i(t)(\boldsymbol{X}_i(t) - \overline{\boldsymbol{X}}(t))^{\otimes 2}dt.$$

Then the estimating equation is equivalent to the linear equation $\boldsymbol{b} - \boldsymbol{V}\boldsymbol{\theta} = 0$. Integrating $-U(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ yields the least squares type loss function

$$L(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{V}\boldsymbol{\theta} - \boldsymbol{b}^T\boldsymbol{\theta}.$$

Based on this loss function, we propose the objective function

$$Q(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + \sum_{j=1}^{p} P(\|\boldsymbol{\theta}_j\|;\lambda_n), \tag{4}$$

where $P(\|\boldsymbol{\theta}\|;\lambda)$ is a penalty function. The estimator $\hat{\boldsymbol{\theta}}$ is obtained by minimizing the objective function (4), and then we define $\hat{\phi}_{nj}(Z_j(t)) = \hat{\boldsymbol{\theta}}_j^T\boldsymbol{B}(Z_j(t))$. To identify the model structure and to select important variables simultaneously, we introduce the penalty function

$$P(\|\boldsymbol{\theta}_j\|;\lambda_n) = \lambda_{1n}\rho_1(|\theta_{j1}|;\lambda_{1n}) + \lambda_{2n}\rho_2(\|\boldsymbol{\theta}_{j1-}\|;\lambda_{2n}), \tag{5}$$

where the tuning parameter $\lambda_n = (\lambda_{1n}, \lambda_{2n})^T$, and the functions $\rho_i(\cdot),\ i = 1, 2$ are set to select important linear and nonlinear components, respectively. For the sake of simplicity, we omit the subscript n in tuning parameters subsequently.

2.3. Blockwise majorization descent algorithm

Let the current value of $\boldsymbol{\theta}$ be $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_p)$ and write $\boldsymbol{\theta}^* = (\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_{j-1}, \boldsymbol{\theta}_j, \tilde{\boldsymbol{\theta}}_{j+1}, \ldots, \tilde{\boldsymbol{\theta}}_p)$. Define $\boldsymbol{u}_j = \partial L(\tilde{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}_j$. Let $\boldsymbol{V}_j$ be the sub-matrix of $\boldsymbol{V}$ corresponding to $\boldsymbol{\theta}_j$ and h_j be the largest eigenvalue of $\boldsymbol{V}_j$. Then

$$\begin{aligned} L(\boldsymbol{\theta}^*) \leq & L(\tilde{\boldsymbol{\theta}}) + (\boldsymbol{\theta}_j - \tilde{\boldsymbol{\theta}}_j)^T\boldsymbol{u}_j + \frac{1}{2}(\boldsymbol{\theta}_j - \tilde{\boldsymbol{\theta}}_j)^T\boldsymbol{V}_j(\boldsymbol{\theta}_j - \tilde{\boldsymbol{\theta}}_j) \\ \leq & L(\tilde{\boldsymbol{\theta}}) + (\boldsymbol{\theta}_j - \tilde{\boldsymbol{\theta}}_j)^T\boldsymbol{u}_j + \frac{h_j}{2}\|\boldsymbol{\theta}_j - \tilde{\boldsymbol{\theta}}_j\|^2. \end{aligned}$$

Based on this relation, Yang & Zou (2015) proposed the blocked majorization descendent (BMD) algorithm to obtain the optimal point $\boldsymbol{\theta}_j$ in (4) by taking

$$\hat{\boldsymbol{\theta}}_j = \arg\min_{\boldsymbol{\theta}_j}\left\{\frac{1}{2}\|\boldsymbol{\theta}_j - (\tilde{\boldsymbol{\theta}}_j - \boldsymbol{u}_j/h_j)\|^2 + \frac{1}{h_j}P(\|\boldsymbol{\theta}_j\|;\lambda)\right\}. \tag{6}$$

Thus, the value of $\boldsymbol{\theta}$ is updated as $\boldsymbol{\theta} = (\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_{j-1}, \hat{\boldsymbol{\theta}}_j, \tilde{\boldsymbol{\theta}}_{j+1}, \ldots, \tilde{\boldsymbol{\theta}}_p)$. The solution of (6) has a closed form for the commonly used penalties. We list the solutions for the group LASSO, group SCAD and group MCP penalties as follows.

For the group LASSO penalty $P(\|\boldsymbol{\theta}_j\|; \lambda) = \lambda\|\boldsymbol{\theta}_j\|$, $\hat{\boldsymbol{\theta}}_j = S(\boldsymbol{c}_j; \lambda/h)$, where $S(\boldsymbol{c}_j; \lambda) = (1 - \lambda/\|\boldsymbol{c}_j\|)_+ \boldsymbol{c}_j$ with $\boldsymbol{c}_j = \tilde{\boldsymbol{\theta}}_j - \boldsymbol{u}_j/h_j$.

For the group SCAD penalty with parameter a, we have

$$P'(\|\boldsymbol{\theta}_j\|; \lambda) = \lambda I(\|\boldsymbol{\theta}_j\| \le \lambda) + \frac{(a\lambda - \|\boldsymbol{\theta}_j\|)_+}{a-1} I(\|\boldsymbol{\theta}_j\| > \lambda),$$

and

$$\hat{\boldsymbol{\theta}}_j = \begin{cases} S(\boldsymbol{c}_j; \lambda/h_j), & \|\boldsymbol{c}_j\| \le \lambda + \lambda/h_j, \\ \frac{[h_j(a-1) - \lambda a/\|\boldsymbol{c}_j\|]\boldsymbol{c}_j}{h_j a - h_j - 1}, & \lambda + \lambda/h_j < \|\boldsymbol{c}_j\| \le \lambda a, \\ \boldsymbol{c}_j, & \|\boldsymbol{c}_j\| > \lambda a. \end{cases}$$

For the group MCP penalty with parameter a, we have $P'(\|\boldsymbol{\theta}_j\|) = (a\lambda - \|\boldsymbol{\theta}_j\|)_+ a^{-1}$, and

$$\hat{\boldsymbol{\theta}}_j = \begin{cases} S\Big(\frac{h_j \boldsymbol{c}_j}{h_j - 1/a}; \frac{\lambda}{h_j - 1/a}\Big), & \|\boldsymbol{c}_j\| \le \lambda a, \\ \boldsymbol{c}_j, & \|\boldsymbol{c}_j\| > \lambda a. \end{cases}$$

Concretely, for the penalty form as in (5), we write h_{j1} as the first entry of $\boldsymbol{V}_j$, and h_{j1-} as the largest eigenvalue of the sub-matrix of $\boldsymbol{V}_j$ by deleting the first row and the first column. Using the BMD algorithm coordinately to θ_{j1} and $\boldsymbol{\theta}_{j1-}$ in (5), we have the following steps for computation:

Step 1. Calculate h_{j1} and h_{j1-} for $j = 1, \ldots, p$;
Step 2. Choose an initial estimate $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(0)}$;
Step 3. Repeat the following blockwise updates until convergence.
(i) Calculate $u_{j1} = \partial L(\tilde{\boldsymbol{\theta}})/\partial\tilde{\theta}_{j1}$;
(ii) Find the minimizer $\hat{\theta}_{j1}$ and set $\tilde{\boldsymbol{\theta}}_j = (\hat{\theta}_{j1}, \tilde{\boldsymbol{\theta}}_{j1-})$;
(iii) Calculate $\boldsymbol{u}_{j-} = \partial L(\tilde{\boldsymbol{\theta}})/\partial\tilde{\boldsymbol{\theta}}_{j-}$;
(iv) Find the minimizer $\hat{\boldsymbol{\theta}}_{j1-}$ and get $\hat{\boldsymbol{\theta}}_j = (\hat{\theta}_{j1}, \hat{\boldsymbol{\theta}}_{j1-})$;
(v) Set $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_{j-1}, \hat{\boldsymbol{\theta}}_j, \tilde{\boldsymbol{\theta}}_{j+1}, \ldots, \tilde{\boldsymbol{\theta}}_p)$.

Tseng & Yun (2009) established the global convergence and the linear convergence under a local Lipschitzian error bound assumption for the block coordinate gradient descent method with the L_1-regularization, which provides the theoretical assurance for the proposed algorithm.

3. ASYMPTOTIC PROPERTIES

For simplicity, we write $g_{0j}(t) = g_j(\tilde{\phi}_{0j}, t)$, and define

$$\hat{g}_{nj}(t) = g_j(\hat{\phi}_{nj}, t) - \frac{1}{\tau}\mathbb{P}_n \int_0^\tau g_j(\hat{\phi}_{nj}, t)dt.$$

Here we use bold-faced letters to represent functional vectors. For example, we denote $\boldsymbol{g} = (g_1, \ldots, g_p)^T$ with $\boldsymbol{g}_i$ being the ith observation of $\boldsymbol{g}$, $\boldsymbol{g}_0 = (g_{01}, \ldots, g_{0p})^T$ and $\hat{\boldsymbol{g}}_n =$

$(\hat{g}_{n1}, \ldots, \hat{g}_{np})^T$. We use the notation $\Lambda(\Gamma)$ and $\Lambda_{\min}(\Gamma)$ to denote the eigenvalue and the minimum eigenvalue of matrix Γ, respectively, and use the subscript $\mathcal{A}$ for a vector or a matrix to denote the corresponding sub-vector or sub-matrix. For example, $\boldsymbol{x}_{\mathcal{A}}$ means the $|\mathcal{A}|-$dimensional vector consisting of components $\{x_j, j \in \mathcal{A}\}$, and $\boldsymbol{V}_{\mathcal{A}\mathcal{A}}$ means the $|\mathcal{A}|-$dimensional squared matrix with entries $\{V_{ij}, i \in \mathcal{A}, j \in \mathcal{A}\}$.

To present our main results, for $k = 0, 1, 2$, we define

$$\boldsymbol{S}^{(k)}(\boldsymbol{g}, t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\boldsymbol{g}_i^{\otimes k}, \qquad \boldsymbol{s}^{(k)}(\boldsymbol{g}, t) = E[Y(t)\boldsymbol{g}^{\otimes k}],$$

$$\boldsymbol{e}(\boldsymbol{g}, t) = \boldsymbol{s}^{(1)}(\boldsymbol{g}, t)/\boldsymbol{s}^{(0)}(\boldsymbol{g}, t), \text{ and } \boldsymbol{D} = \mathrm{E}\Big[\int_0^\tau Y(t)\{\boldsymbol{g}_0(t) - \boldsymbol{e}(\boldsymbol{g}_0, t)\}^{\otimes 2} dt\Big].$$

Let $d = \frac{1}{2}\min_{j\in\mathcal{A}} \|g_{0j}\|_\infty$, $\kappa(\lambda, \boldsymbol{\theta}) = \max_{1\le j\le q} \dfrac{\partial^2 P(\|\boldsymbol{\theta}\|; \lambda)}{\partial|\theta_j|^2}$ for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^T$, $\kappa_0 = \sup\{\kappa(\lambda, \boldsymbol{\theta}) : \|\boldsymbol{g} - \boldsymbol{g}_0\|_\infty \le d\}$, and $\mu = \Lambda_{\min}(\boldsymbol{D}_{\mathcal{A}\mathcal{A}}) - \kappa_0$. Let $\rho_\lambda(\boldsymbol{\theta})$ represent $\rho_1(\boldsymbol{\theta}; \lambda_1)$ or $\rho_2(\boldsymbol{\theta}; \lambda_2)$. We need the following conditions.

Condition 1 The function $\rho_\lambda(\boldsymbol{\theta})$ is increasing and concave on each component θ_{jl} of $\boldsymbol{\theta}$, and has a continuous partial derivative $\partial\rho_\lambda(\boldsymbol{\theta})/\partial\theta_{jl}$ for $j = 1, \ldots, p$, $l = 1, \ldots, q_n$. In addition, $\rho'_\lambda(\boldsymbol{\theta})$ is increasing on tuning parameter λ, and $\rho'_\lambda(\boldsymbol{0}+) = \rho'(\boldsymbol{0}+) > 0$ is independent of λ.

Condition 2 $\int_0^\tau \lambda_0(t)dt < \infty$ and $P(Y(\tau) = 1) > 0$.

Condition 3 Let $\tilde{\boldsymbol{Z}}_{\mathcal{C}}(t)$ be the projection of $\boldsymbol{Z}_{\mathcal{B}}(t)$ on $\boldsymbol{Z}_{\mathcal{C}}(t)$, $\boldsymbol{\Gamma}_1 = E\Big[\int_0^\tau (\boldsymbol{Z}_{\mathcal{B}}(t) - \tilde{\boldsymbol{Z}}_{\mathcal{C}}(t))^{\otimes 2} dt\Big]$, $\boldsymbol{\Gamma}_2 = E\Big[\int_0^\tau \boldsymbol{Z}_{\mathcal{C}}(t)\boldsymbol{Z}_{\mathcal{C}}(t)^T dt\Big]$, and $\boldsymbol{\Gamma}_3 = E\Big[\int_0^\tau \tilde{\boldsymbol{Z}}_{\mathcal{C}}(t)\tilde{\boldsymbol{Z}}_{\mathcal{C}}(t)^T dt\Big]$. There exist finite positive constants ρ_1, ρ_2 and ρ_3^* such that $\Lambda(\boldsymbol{\Gamma}_j) \ge \rho_j$ for $j = 1, 2$ and $\Lambda(\boldsymbol{\Gamma}_3) \le \rho_3^*$.

Condition 1 is mild, which can be met by all the penalties listed in Section 2.3. Condition 2 is standard for survival models, and Condition 3 is needed technically to derive the L_2-loss convergence rate of the proposed estimators.

Let $\hat{\boldsymbol{\theta}}_{\mathcal{D}} = \{\hat{\boldsymbol{\theta}}_j, j \in \mathcal{D}\}$, $\hat{\boldsymbol{\theta}}_{\mathcal{B}}^{1-} = \{\hat{\boldsymbol{\theta}}_{j1-}, j \in \mathcal{B}\}$, $\hat{\boldsymbol{\beta}}_n = \{\hat{\phi}_{nj}, j \in \mathcal{B}\}$, $\hat{\boldsymbol{\phi}}_n = \{\hat{\phi}_{nj}, j \in \mathcal{C}\}$, and $\eta_n = n^{-\nu\varsigma} + n^{-(1-\nu)/2}$. We now present the results regarding the selection consistency of the proposed estimators.

Theorem 1. *(Consistency of the selection) Suppose that Conditions 1-3 hold. Also assume that*

$$\begin{gathered}\frac{\eta_n^{-2}}{\log p} \to \infty, \frac{n + \eta_n^{-2}}{\log s \vee \log n} \to \infty, \frac{(n + \eta_n^{-2})\mu^2}{s^2(\log s \vee \log n)} \to \infty, \\ \frac{\lambda^2 n}{\log p \vee \log n} \to \infty, \frac{\lambda^2 n\eta_n^{-2}}{p^2(\log p \vee \log n)} \to \infty,\end{gathered} \tag{7}$$

where $\lambda = \min\{\lambda_1, \lambda_2\}$, $\mu > 0$, *and* $0.25/\varsigma < \nu < 0.5$ *with* $\varsigma = k + \alpha > 0.5$ *for some nonnegative integer k. Then with probability tending to one, we have*
(i) (Variable selection Sparsity) $\hat{\boldsymbol{\theta}}_{\mathcal{D}} = 0$*;*
(ii) (Structure identification consistency) $\hat{\boldsymbol{\theta}}_{\mathcal{B}}^{1-} = 0$*;*
(iii) (L_2-loss) $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|^2 = O_p(s^2\eta_n^2)$ *and* $\|\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0\|^2 = O_p(s^2\eta_n^2)$*.*

Property (i) in Theorem 1 shows that the unimportant variables can be selected with high probability; property (ii) illustrates the ability identifying the model structure; property (iii) implies the convergence rate of the regression coefficient estimates of important variables in L_2-norm.

The first part in (7) is a condition on the relationship between the variable dimension p and the important variable number s. Especially, if μ is a constant, then the proposed estimators can handle the case with $\log p = o(n^{2\nu\varsigma} + n^{(1-\nu)})$ and $s = o(n^{\nu\varsigma}/\log n + n^{(1-\nu)/2}/\log n)$. This shows that the dimension of covariates is allowed to grow nonpolynomially and the dimension of the true sparse model can be divergent. The following oracle property shows that the estimators of important linear effects are asymptotically distributed normal.

Theorem 2. *(Asymptotical Normality) Suppose that the conditions of Theorem 1 hold and $\lambda_1 = o(s_1^{-1} n^{(1-\nu)/2})$. Then for every $\boldsymbol{u} \in \mathbb{R}^{s_1}$ with $\|\boldsymbol{u}\| = 1$, $\sqrt{n}\boldsymbol{u}\boldsymbol{\Sigma}^{-1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ is asymptotically standard normal, where the matrix $\boldsymbol{\Sigma}$ is defined as in the Appendix.*

4. SIMULATION STUDIES

Simulation studies were conducted to evaluate the finite sample properties of the proposed penalized estimators via group LASSO, adaptive group LASSO, group SCAD, and group MCP penalties. In the study, the failure time T^U was generated from the additive hazards regression model (1), and the censoring time C was generated form the Uniform$(\tau/2, \tau)$ such that the censoring rate reaches at 20%. For the adaptive selector, we adopt the linear coefficients selected by group lasso as the weights of the penalized linear part. The parameter a in group SCAD was taken as 3.7, and $a = 2/\{1 - \max_{i \neq j} x_i^T x_i/n\}$ in group MCP. The tuning parameters λ_1 and λ_2 were selected by using the EBIC criterion suggested by Chen & Chen (2008). The orthogonal cubic B-splines were adopted to select the important variables and to identify the structure, where q_n was taken 8 approaching to the order of $O(n^{1/3})$. The simulation results are based on 200 replications for $p = 15$ and 100 replications for $p = 500, 1000$ with sample size $n = 500$ using R software. To compare the performance of different penalties, we computed the L_1 prediction error (PE$= \|\hat{\boldsymbol{g}}_n - \boldsymbol{g}_0\|_1$), the true positive rate (TPR) representing the rate that an important variable is correctly selected, the false positive rate (FPR) representing the rate that an unimportant variable is wrongly selected, TPRN representing the rate that an important variable with a nonlinear effect are correctly selected, and FPRN representing the rate that an unimportant variable or an important variable with a linear effect are wrongly detected as having a nonlinear effect.

Example 1. Consider model (1) with

$$\lambda(t) = 2 + 2Z_1 - 2Z_2 + 2f_1(Z_3) + \sum_{j=1}^{15} 0 \cdot Z_j,$$

where $f_1(z) = \sin(z) + 2z\cos(2z)$. The covariates Z_j's were generated from an AR(1) model trimmed to $[-1, 1]$ with the initial standard normal distribution and $Cov(Z_{j_1}, Z_{j_2}) = 0.4^{|j_1 - j_2|}$ for $j_1, j_2 = 1, \ldots, 15$. Specifically, to generate the covariates Z_j's, we first generated i.i.d. random variables $W_j, j = 1, \ldots, 15$ from a normal distribution with mean 0 and variance $1 - 0.4^2$, and Z_1 independently from a standard normal distribution. Then we took $Z_j = 0.4Z_{j-1} + W_j$ for $j = 2, \ldots, 15$, and trimmed Z_j to $[-1, 1]$ for $j = 1, \ldots, 15$. There are three important variables with two linear effects and one nonlinear effect.

Example 2. Consider model (1) with

$$\lambda(t) = 2 + Z_1 - 1.5Z_2 + 0.8Z_3 + 2f_1(Z_4) - 0.5f_2(Z_5) + \sum_{j=6}^{p} 0 \cdot Z_j,$$

where $f_1(z) = \sin(z) + 2z\cos(2z)$ and $f_2(z) = z(\exp(2z^2) - 3\log(2 + z^2))$. The setup for the covariates Z_j's $(j = 1, \ldots, p)$ was the same as in Example 1. There are five important variables with three linear effects and two nonlinear effects. Different combinations of sample size $n = 500$ and model sizes $p = 15, 500, 1000$ were considered to evaluate the performance of the proposed method.

Example 3. Consider model (1) with

$$\lambda(t) = 2t^3 + Z_1(t) - 1.5Z_2(t) + 0.8Z_3(t) + 2f_3(Z_4(t)) + \sum_{j=5}^{15} 0 \cdot Z_j(t),$$

where $f_3(z) = z^3$, and $Z_j(t), j = 1, \ldots, 15$ are time-dependent covariates. To generate the covariates $Z_j(t), j = 1, \ldots, 15$, we first generated η_j's from an AR(1) model trimmed to $[-1, 1]$ with the initial standard normal distribution and $Cov(\eta_{j_1}, \eta_{j_2}) = 0.4^{|j_1 - j_2|}$ for $j_1, j_2 = 1, \ldots, 15$. Then we took $Z_j(t) = \eta_j t^3$ for $j \neq 4$, and $Z_4(t) = \eta_4 t$. There are four important variables with three linear effects and one nonlinear effect.

The simulation results are summarized in Tables 1-9. Tables 1 and 2 report the rate of each component being selected as an important variable and the rate of each component being identified as having a nonlinear effect over 200 replications in Example 1, respectively. Based on Example 2, Table 3 reports the rate of each component being selected as an important variable over 200 replications for $p = 15$; Table 4 reports the rate of each important variable being correctly selected, PE, TPR and FPR for $p = 500, 1000$; Table 5 reports the rate of each component being selected as having a nonlinear effect for $p = 15$; Table 6 reports the rates of each important variable being identified as having a nonlinear effect, TPRN and FPRN for $p = 500, 1000$. Based on Example 3, Tables 7 and 8 report the rate of each component being selected as an important variable and the rate of each component being selected as having a nonlinear effect over 200 replications for $p = 15$, respectively. All the four penalized detection methods perform well. They can correctly identify the important variables and the nonlinear effects on the hazard rate function with high probability for all situations considered. In addition, the results indicate that for different combinations of sample size and model size (e.g. $(n, p) = (500, 15), (500, 500), (500, 1000)$), the adaptive group LASSO, the group SCAD, and the group MCP perform better than the group LASSO in that they could select important variables and nonlinear components with much higher true positive rates and lower false positive rates. Table 9 reports the estimation results for the linear regression coefficients of the important covariates by the adaptive group LASSO for $p = 15$, including the bias, the sample standard deviation (SSE), the estimated standard error (ESE), and the mean squared error of the estimated regression coefficients for important covariates in three examples. The results show that the proposed estimates are nearly unbiased and the ESEs are reasonably well.

Moreover, Figure 1 displays the functional estimates for important linear components z_1, z_2 and z_3 in Example 1, while Figure 2 shows the functional estimates for important nonlinear components z_4 and z_5 in Example 2 when $p = 15$. As the figures are similar, we only show the figures by using the adaptive group LASSO selector in Examples 1 and 2. In these figures, the solid line is the pointwise mean estimate from 200 Monto Carlo repetitions, the dot and dash line is the true function, and the dotted lines are the 95% pointwise confidence bands. The 95% confident intervals are produced by calculating the pointwise standard error using 200 bootstrap replications. From Figures 1 and 2, we can see that the fitted functions are closed to the true functions and the pointwise confidence bands cover the true values of the functions completely. These results demonstrate that the proposed method can efficiently estimate linear and nonlinear effects of covariates.

5. AN APPLICATION

We applied the proposed methods to analyzing the primary biliary cirrhosis (PBC) data. The data set recorded 424 patients suffering from PBC, which was conducted by the Mayo Clinic between 1974 and 1984. The main purpose of this trial was to investigate the risk factors of PBC. There were two treatment groups (D-penicillamine and placebo) and 16 baseline covariates including clinical and laboratory measurements collected for 312 randomized participants, denoted by TRT, Age, Sex, Ascites, Hepato, Spiders, Edema, Stage, Bili, Chol, Albumin, Copper, Alk.phos, SGOT, Trig, Platelet, and Protime. We used 276 complete observations for the analysis, discarding the missing data. To analyze the hazard of the disease before the liver transplantation, we considered the death time as the failure time, and the earlier of the liver transplantation time or the follow-up time as the censoring time. The censoring rate of this data set is about 60%. More details about the PBC data can be found in Fleming and Harrington (1991). The data have been analyzed by many researchers, including Ma & Huang (2005), Zhang & Lu (2007), Leng & Ma (2007), Cao et al. (2016), and Wang & Xiang (2017) among others.

To fit the model, we scaled each covariate by subtracting its minimum value and then dividing the corresponding range such that all covariates took values in [0, 1] for convenience. We considered four competing selectors: the group LASSO, the adaptive group LASSO, the group SCAD, and the group MCP. The tuning parameter was determined by the EBIC criterion (Chen & Chen, 2008). The results of variable selection and structure identification are summarized in Table 10. Obviously, the models identified by the adaptive group LASSO, the group SCAD, and the group MCP selectors are more sparse than the group LASSO selector. Our methods identify covariates Age, Ascites, Spiders, and Edema as important variables having linear effects, and Bili, Albumin, Copper, Alk.phos, SGOT, Trig, Platlet, and Protime as important factors having nonlinear effects. As a comparison, Ma & Huang (2005) fitted the data to an additive hazards model and identified Age, Stage, $\log$(Bili) and $\log$(Copper) as important variables with linear effects using the Lasso regularization method, while Leng & Ma (2007) and Wang & Xiang (2017) detected nine covariates (Age, Ascites, Edema, Bili, Albumin, Copper, SGOT, Protime, and Stage) as important variables with linear effects using the modified Lasso and the penalized empirical likelihood method with the SCAD penalty for the additive hazards model, respectively. Our new finding is that covariates Alk.phos, SGOT, Trig, Platelet, and Protime have nonlinear effects on the hazard function, although Alk.phos, Trig and Platelet could not be detected as important predictors with linear effects on the hazard function by the existing methods. This provides new insights into the study of PBC data.

Furthermore, we took the selection results by using the adaptive group LASSO penalty as an example to observe the trend of nonlinear effects on the hazard function shown in Figure 3, where the solid lines display the estimates of the important nonlinear effects and the dotted lines show the corresponding 95% pointwise confidence intervals based on 200 bootstrap replications.

6. DISCUSSION

We have considered the problem of simultaneous model pursuit, variable selection and estimation for additive hazards regression models in the framework with ultrahigh-dimensional covariates and diverging dimension of important variables. We have constructed the sieve least squares loss function with the orthogonal penalty, and established the asymptotic properties of the resulting estimators under mild conditions, showing that the proposed method can identify the model structure and select important variables consistently as if we knew the true model in advance. The proposed procedure can be implemented through the blockwise majorization descent algorithm. The numerical results show that the proposed method performs well in identifying the model structure, selecting important variables, and estimating the effects of variables simultaneously.

Although we focus on additive hazards models in this paper, the proposed methods can be extended to dealing with the following varying coefficient additive hazards model

$$\lambda(t) = \lambda_0(t) + \sum_{j=1}^{p} Z_j(t) g_j(X_j(t)),$$

where $Z_j(t)$'s are covariates and $X_j(t)$ is an influential variable treated as the index variable. The equivalent problem becomes to decide whether the effect of variable $Z_j(t)$ is linear, coefficient-varying or unimportant depending on the form of function g_j. A critical issue in the selection process is to determine stable optimal tuning parameter when the dimensionality of variables is rarely high, which is still an open question now.

ACKNOWLEDGEMENTS

The authors would like to thank Editor-in-Chief, Professor Fang Yao, the Associate Editor and the two reviewers for their constructive and insightful comments and suggestions that greatly improved the paper. Li Liu's research is partly supported by the National Natural Science Foundation of China (No. 11971362). Yanyan Liu and Feng Su's research are partly supported by the National Natural Science Foundation of China (No. 115771263). Zhao's research is partly supported by the National Natural Science Foundation of China (No. 11771366), and the Research Grant Council of Hong Kong (15301218, 15303319).

APPENDIX

Proof of Proposition 1.

(i) Since $E(|G(Z_1(t), \ldots, G(Z_p(t))|) = 0$, we have

$$\int_{[a,b]^p} |G(z_1, \ldots, z_p)| f_t(z_1, \ldots, z_p) dz_1 \cdots dz_p = 0.$$

Define $f(z_1, \ldots, z_p) = \int_0^\tau f_t(z_1, \ldots, z_p)dt$. Integrating two sides of the above equation on t from 0 to τ, we get that

$$\int_{[a,b]^p} |G(z_1, \ldots, z_p)| f(z_1, \ldots, z_p) dz_1 \cdots dz_p = 0.$$

Noting that $f(z_1, \ldots, z_p) > 0$, we have $G(z_1, \ldots, z_p) = 0$.

(ii) For any $j = 1, \ldots, p$, $g_j(z_j) = -\sum_{k \neq j} g_k(z_k)$. By the Fubini's theorem, there exists some $(z_1^0, \ldots, z_{j-1}^0, z_{j+1}^0, \ldots, z_p^0) \in [a, b]^{p-1}$ such that $g_j(z_j) = -\sum_{k \neq j} g_k(z_k^0)$ holds for any $z_j \in [a, b]$.

(iii) As we have $\tilde{\phi}_j(x) = \beta_j + \phi_j(x)$ with $\beta_j = \int_a^b \tilde{\phi}_j(x)dx$ and $\phi_j(x) = \tilde{\phi}_j(x) - \beta_j$ for $j = 1, \ldots, p$, the decomposition (2) holds. To show the uniqueness of the decomposition, we assume that there exist $\beta_j^l \in \mathbb{R}$ and $\phi_j^l \in \mathcal{H}$, $j = 1, \ldots, p, l = 1, 2$ such that

$$\sum_{j=1}^{p} Z_j(t)[\beta_j^1 + \phi_j^1(Z_j(t))] = \sum_{j=1}^{p} Z_j(t)[\beta_j^2 + \phi_j^2(Z_j(t))]. \tag{A.1}$$

It suffices to prove that $\beta_j^1 = \beta_j^2$ and $\phi_j^1(x) = \phi_j^2(x)$ for each $j = 1, \ldots, p$. To this end, we note that (A.1) implies that

$$\sum_{j=1}^{p} z_j \Big([\beta_j^1 - \beta_j^2] + [\phi_j^1(z_j) - \phi_j^2(z_j)] \Big) = 0$$

by (i). Then by (ii), there exists some constant C_j such that

$$z_j \Big([\beta_j^1 - \beta_j^2] + [\phi_j^1(z_j) - \phi_j^2(z_j)] \Big) = C_j$$

for each $j = 1, \ldots, p$ and $z_j \in [a, b]$. This yields that

$$Z_j(t) \Big([\beta_j^1 - \beta_j^2] + [\phi_j^1(Z_j(t)) - \phi_j^2(Z_j(t))] \Big) = C_j.$$

Since $E\Big(\int_0^\tau [\beta_j^l Z_j(t) + Z_j(t)\phi_j^l(Z_j(t))]dt \Big) = E\Big(\int_0^\tau Z_j(t)\tilde{\phi}_j(Z_j(t))dt \Big)$, $l = 1, 2$, we have

$$(\beta_j^1 - \beta_j^2) + [\phi_j^1(z) - \phi_j^2(z)] = 0 \tag{A.2}$$

for each $j = 1, \ldots, p$. Integrating two sides of (A.2) on variable z from a to b and noting that $\phi_j^l(x) \in \mathcal{H}$, $l = 1, 2$, we obtain that $\beta_j^1 = \beta_j^2$. Thus, (A.2) gives that $\phi_j^1(z) = \phi_j^2(z)$. This completes the proof of the proposition. ■

For any probability measure $\mathbb{P}$, define $L_2(\mathbb{P}) = \{f : \int f^2 d\mathbb{P} < \infty\}$ and $\|f\| = \Big(\int f^2 d\mathbb{P} \Big)^{1/2}$. For any subclass $\mathcal{F}$ of $L_2(\mathbb{P})$, define the bracketing number as $N_{[]}(\epsilon, \mathcal{F}, L_2(\mathbb{P}))$ and $J_{[]}(\eta, \mathcal{F}, L_2(\mathbb{P})) = \int_0^\eta \sqrt{1 + \log N_{[]}(\epsilon, \mathcal{F}, L_2(\mathbb{P}))} d\epsilon$. For i.i.d. random variables $X_1, \ldots, X_n$ with distribution $\mathbb{P}$, let $\mathbb{P}_n$ be the empirical measure of these random variables. We write $\overline{f}(t) = \{\sum_{i=1}^{n} Y_i(t) f_i(t)\} / \{\sum_{i=1}^{n} Y_i(t)\}$ for a given function $f(t)$. Let $\Omega_\eta = \{\omega : \sup_{t \in [0,\tau]} |g_{nj}(t) - g_{0j}(t)| \le \eta_n, j = 1, \ldots, p\}$ for some given η_n and $g_{nj}(t)$, the latter of which can be defined as (A.6).

To prove the theorems, we need some lemmas.

Lemma A. 1 *(Concentration of $\boldsymbol{S}^{(k)}(\cdot)$, $k = 0, 1, 2$) Recalling that the notation $\boldsymbol{g}_n = (g_{n1}, \ldots, g_{np})^T$, there exist constants C_1, C_2 and L such that*

$$P\big(\sup_{t \in [0,\tau]} |S_j^{(0)}(\boldsymbol{g}_n, t) - s_j^{(0)}(\boldsymbol{g}_0, t)| \ge C_1 n^{-1/2}(1 + x) + C_2 \eta_n \big) \le \exp(-Lx^2), \tag{A.3}$$

$$P\big(\sup_{t \in [0,\tau]} |S_j^{(1)}(\boldsymbol{g}_n, t) - s_j^{(1)}(\boldsymbol{g}_0, t)| \ge C_1 n^{-1/2}(1 + x) + C_2 \eta_n \big| \Omega_\eta \big) \le \exp(-Lx^2), \tag{A.4}$$

$$P\big(\sup_{t \in [0,\tau]} |S_{ij}^{(2)}(\boldsymbol{g}_n, t) - s_{ij}^{(2)}(\boldsymbol{g}_0, t)| \ge C_1 n^{-1/2}(1 + x) + C_2 \eta_n \big| \Omega_\eta \big) \le \exp(-Lx^2) \tag{A.5}$$

hold for all $x > 0$ and $i, j = 1, \ldots, p$, where $S_j^{(1)}$ is the jth component of $\boldsymbol{S}^{(1)}(\cdot)$ and $S_{ij}^{(2)}(\cdot)$ is the $(i, j)-$th entry of the matrix $\boldsymbol{S}^{(2)}(\cdot)$.

Proof. We only show (A.4). (A.3) and (A.5) can be similarly proved and thus omitted. Since

$$\begin{aligned} R_j &= \sup_{t\in[0,\tau]} |S_j^{(1)}(\boldsymbol{g}_n, t) - s_j^{(1)}(\boldsymbol{g}_0, t)| \\ &\le \sup_{t\in[0,\tau]} |S_j^{(1)}(\boldsymbol{g}_n, t) - s_j^{(1)}(\boldsymbol{g}_n, t)| + \sup_{t\in[0,\tau]} |s_j^{(1)}(\boldsymbol{g}_n, t) - s_j^{(1)}(\boldsymbol{g}_0, t)| \\ &= I_1 + I_2. \end{aligned}$$

To apply a functional Hoeffding-type inequality, we need to control the term ER_j. Let $m_0(t) = Y(t)g_{nj}(t)$ and $\mathcal{M}(\eta_1) = \{m_0 : \|g_{nj} - g_{0j}\| \le \eta_1\}$. Then, similar to Corollary A1 of Huang (1999), we can get that

$$\log N_{[]}(\epsilon, \mathcal{M}(\eta_1), L_2(\mathbb{P})) \le C[q_n \log(\eta_1/\epsilon) + \log(\tau/\epsilon)],$$

$$\text{and} \quad J_{[]}(\eta_1, \mathcal{M}, L_2(\mathbb{P})) \le C[q_n^{1/2}\eta_1 + \eta_1 \log^{1/2}(1/\eta_1)].$$

Taking $\eta_1 = q_n^{-1/2}$ in Lemma 3.4.2 of van der Vaart & Wellner (1996), we have $EI_1 \le C_1 n^{-1/2}$. Noting that $\sup_{t\in[0,\tau]} |g_{nj}(t) - g_{0j}(t)| \le \eta_n$ for any $j = 1, \ldots, p$ conditional on Ω_η, we immediately obtain that $EI_2 \le C_2\eta_n$. Thus, $ER_j \le C_1 n^{-1/2} + C_2\eta_n$. It follows from Theorem 9 of Massart (2000) that

$$P(R_j \ge C_1 n^{-1/2}(1+x) + C_2\eta_n|\Omega_\eta) \le P(R_j \le ER_j + C_1 n^{-1/2}x|\Omega_\eta) \le \exp(-L^2x^2)$$

for some constant $L > 0$. ■

The following Lemmas A.2 and A.3 can be proved by using the similar arguments as used in the proofs of Lemma A4 and Lemma 1 in Lin & Lv (2013).

Lemma A. 2 *(Concentration of $\boldsymbol{V}$) We denote*

$$\boldsymbol{V}(\boldsymbol{g}_n) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau Y_i(t)(\boldsymbol{g}_{ni}(t) - \bar{\boldsymbol{g}}_n(t))^{\otimes 2}dt$$

as an empirical version of the matrix $\boldsymbol{D}$. Then there exist constants C_1, C_2, C and $L > 0$ such that

$$P(|V_{ij}(\boldsymbol{g}_n) - D_{ij}| \ge C_1 n^{-1/2}(1+x) + C_2\eta_n|\Omega_\eta) \le C\exp\big(-L(x^2 \wedge (n + \eta_n^{-2}))\big).$$

Lemma A. 3 *(Concentration of empirical matrices) If $\mu > 0$ and $\mu^{-1} = O((n^{1/2} + \eta_n^{-1})/s)$, there exist constants C and $L > 0$ such that*

$$P(\Lambda_{\min}(\boldsymbol{V}_{\mathcal{A}\mathcal{A}}) \le \lambda\kappa_0|\Omega_\eta) \le Cs^2q_n^2 \exp\Big(-L(n + \eta_n^{-2})\Big(\frac{\mu^2}{s^2} \wedge 1\Big)\Big).$$

Lemma A. 4 *(Concentration of $U_j(\boldsymbol{\theta}_n)$) Under Condition 2, there exist constants C_1, C_2, C and $L > 0$ such that for $j = 1, \ldots, pq_n$,*

$$P(|U_j(\boldsymbol{\theta}_n)| \ge C_1 n^{-1/2}(1+x) + C_2 p\eta_n(n^{-1/2}x + 1)|\Omega_\eta) \le C\exp\big(-L(x^2 \wedge (n + \eta_n^{-2}))\big),$$

where $U_j(\boldsymbol{\theta}_n)$ is the jth component of $\boldsymbol{U}(\boldsymbol{\theta}_n)$.

Proof. Let $X_j(t)$ and $\overline{X}_j(t)$ be the jth component of $\boldsymbol{X}(t)$ and $\overline{\boldsymbol{X}}(t)$, respectively. We write

$$U_j(\boldsymbol{\theta}_n) = \mathbb{P}_n \int_0^\tau (X_j(t) - \overline{X}_j(t)) dM(\boldsymbol{g}_n, t),$$

where

$$dM(\boldsymbol{g}_n, t) = dM(\boldsymbol{g}_0, t) - Y(t) \sum_{j=1}^{p} (g_{nj}(t) - g_{0j}(t)) dt$$

with $dM(\boldsymbol{g}_0, t) = dN(t) - Y(t) \sum\limits_{j=1}^{p} g_{0j}(t)dt$. Hence, we have

$$\begin{aligned} U_j(\boldsymbol{\theta}_n) =& \mathbb{P}_n \int_0^\tau \{X_j(t) - \overline{X}_j(t)\} dM_0(\boldsymbol{g}_0, t) - \mathbb{P}_n \int_0^\tau Y(t)\{X_j(t) - \overline{X}_j(t)\} \sum_{j=1}^{p} (g_{nj}(t) - g_{0j}(t)) dt \\ =& \mathbb{P}_n \int_0^\tau X_j(t) dM_0(\boldsymbol{g}_0, t) - \mathbb{P}_n \int_0^\tau Y(t) X_j(t) \sum_{j=1}^{p} (g_{nj}(t) - g_{0j}(t)) dt \\ & - \mathbb{P}_n \int_0^\tau \overline{X}_j(t) dM_0(\boldsymbol{g}_0, t) + \mathbb{P}_n \int_0^\tau Y(t) \overline{X}_j(t) \sum_{j=1}^{p} (g_{nj}(t) - g_{0j}(t)) dt \\ =& I_1 - I_2 - I_3 + I_4. \end{aligned}$$

Since term I_1 is a sum of i.i.d. zero-mean random variables, applying the Hoeffding's (1963) inequality to I_1 yields that $P(|I_1| \geq C_1 n^{-1/2} x | \Omega_\eta) \leq C \exp(-Lx^2)$.

To estimate term I_2, we note that $\sup\limits_{t \in [0,\tau]} |g_{nj} - g_{0j}| \leq \eta_n$ on set Ω_η. Thus, there exists a constant C_2 such that $E|I_2| \leq C_2 p \eta_n$. Using Hoeffding's inequality again yields that

$$P(|I_2| \geq C_2 p \eta_n (n^{-1/2} x + 1) | \Omega_\eta) \leq C \exp(-Lx^2).$$

Similar to Lemma A3 of Lin & Lv (2013), we have $E|I_3| \leq C_1 n^{-1/2}$, and then we can get that

$$P(|I_3| \geq C_1 n^{-1/2} (1 + x) | \Omega_\eta) \leq C \exp(-Lx^2 \wedge (n + \eta_n^{-2})).$$

For term I_4, by using the same clues as the term I_2, it follows that

$$P(|I_4| \geq C_2 p \eta_n (n^{-1/2} x + 1) | \Omega_\eta) \leq C \exp(-Lx^2 \wedge (n + \eta_n^{-2})).$$

Thus, the conclusion of the lemma holds. ■

Lemma A. 5 *Assume that Conditions 1 and 2 hold. Then for any $g_{0j} \in \mathcal{H}$, there exist functions $\phi_{nj} \in \mathcal{S}_n$ and $g_{nj}(t)$ defined by (A.6), and a constant L such that for any $x > 0$,*

$$P\big(\sup_{t \in [0,\tau]} |g_{nj}(t) - g_{0j}(t)| \geq C(n^{-\nu\varsigma} + n^{-(1-\nu)/2})(1 + x) \big) \leq \exp(-Lx^2)$$

with $\mathbb{P}_n \int_0^\tau g_{nj}(t)dt = 0$.

Proof. According to Corollary 6.21 of Schumaker (1981), for any $1 \le j \le p$, there exists $\phi_{nj} \in \mathcal{S}_n$ such that $\|\phi_{nj} - \tilde{\phi}_{0j}\|_\infty = O(n^{-\nu\varsigma})$. Define

$$g_{nj}(t) = g_j(\phi_{nj}, t) - \frac{1}{\tau}\mathbb{P}_n \int_0^\tau g_j(\phi_{nj}, t)dt. \tag{A.6}$$

Then it is easy to see that $\mathbb{P}_n \int_0^\tau g_{nj}(t)dt = 0$. Furthermore, we note that

$$\|g_{nj} - g_{0j}\|_\infty \le \frac{1}{\tau}\Big\|\mathbb{P}_n \int_0^\tau g_j(\phi_{nj}, t)dt\Big\|_\infty + \|g_j(\phi_{nj}, t) - g_{0j}(t)\|_\infty \triangleq I_{1n} + I_{2n}, \tag{A.7}$$

where

$$I_{1n} \le C\Big(\Big\|(\mathbb{P}_n - \mathbb{P})\int_0^\tau g_j(\phi_{nj}, t)dt\Big\|_\infty + \Big\|\mathbb{P}\int_0^\tau (g_j(\phi_{nj}, t) - g_{0j}(t))dt\Big\|_\infty\Big)$$

with C being a constant independent of n. By Lemma 3.4.2 in van der Vaart & Wellner (1996), we have $(\mathbb{P}_n - \mathbb{P})\int_0^\tau g_j(\phi_{nj}, t)dt = O_p(n^{-(1-\nu)/2})$. And the definition of ϕ_{nj} shows that $\Big\|\mathbb{P}\int_0^\tau (g_j(\phi_{nj}, t) - g_{0j}(t))dt\Big\|_\infty = O(n^{-\nu\varsigma})$. Hence we have

$$I_{1n} = O_p(n^{-\nu\varsigma} + n^{-(1-\nu)/2}). \tag{A.8}$$

In addition,

$$I_{2n} = \|Z_j(t)\phi_{nj}(Z_j(t)) - Z_j(t)\tilde{\phi}_{0j}(Z_j(t))\|_\infty = O_p(n^{-\nu\varsigma}). \tag{A.9}$$

Plugging (A.8) and (A.9) into (A.7), we have $E \sup_{t\in[0,\tau]} |g_{nj} - g_{0j}| \le C(n^{-\nu\varsigma} + n^{-(1-\nu)/2})$. Thus, applying Theorem 9 of Massart (2000) gives that

$$\begin{aligned}
&P\big(\sup_{t\in[0,\tau]} |g_{nj} - g_{0j}| \ge C(n^{-\nu\varsigma} + n^{-(1-\nu)/2})(1 + x)\big) \\
\le &P\big(\sup_{t\in[0,\tau]} |g_{nj} - g_{0j}| \ge E \sup_{t\in[0,\tau]} |g_{nj} - g_{0j}| + C(n^{-\nu\varsigma} + n^{-(1-\nu)/2})x\big) \le \exp(-Lx^2).
\end{aligned}$$

This completes the proof of the lemma. ■

Lemma A. 6 *Assume that Conditions 1-2 hold. If* $0.25/\varsigma < \nu < 0.5$ *and* $\lambda = o(sq_n^{-1})$, *then* $\|\hat{g}_{n\mathcal{A}} - g_{n\mathcal{A}}\|^2 = o_p(s^2q_n^{-1})$ *for* $\hat{g}_{n\mathcal{D}} = 0$ *with* $\hat{g}_{n\mathcal{A}} = (\hat{g}_{nj}, j \in \mathcal{A})$, $g_{n\mathcal{A}} = (g_{nj}, j \in \mathcal{A})$ *and* $\hat{g}_{n\mathcal{D}} = (\hat{g}_{nj}, j \in \mathcal{D})$.

Proof. Let $g_n^s(t) = \sum_{j=1}^s g_j(\phi_{nj}, t)$ and $h_n^s(t) = \sum_{j=1}^s g_j(\phi_{nj}^*, t)$ with $\phi_{nj}^*(t) = \boldsymbol{\theta}_{nj}^{*T}\boldsymbol{B}(Z_j(t))$ and $\|\phi_{nj}^*\|^2 = O(q_n^{-1})$. Then we have $\|\frac{1}{s}h_n^s\|^2 = O(q_n^{-1})$. Define $H_n(\alpha) = Q(\boldsymbol{\theta}_n + \alpha\boldsymbol{\theta}_n^*)$. To prove this lemma, it is sufficient to show that for any $\alpha_0 > 0$, with probability tending to one, $H_n'(\alpha_0) > 0$ and $H_n'(-\alpha_0) < 0$. Note that

$$\begin{aligned} H_n(\alpha_0) =& -\mathbb{P}\int_0^\tau (g_n^s(t) - \overline{g_n^s}(t))dN(t) - \alpha_0\mathbb{P}\int_0^\tau (h_n^s(t) - \overline{h_n^s}(t))dN(t) \\ &+ \frac{1}{2}\mathbb{P}\int_0^\tau Y(t)[(g_n^s(t) - \overline{g_n^s}(t)) + \alpha_0(h_n^s(t) - \overline{h_n^s}(t))]^2 dt \\ &+ \lambda\sum_{j=1}^s \rho(\|\boldsymbol{\theta}_{nj} + \alpha_0\boldsymbol{\theta}_{nj}^*\|). \end{aligned}$$

Then

$$\begin{aligned} H_n'(\alpha_0) =& -\mathbb{P}_n\Big[\int_0^\tau (h_n^s(t) - \overline{h_n^s}(t))dM_n^s(t)\Big] \\ &+ \alpha_0\mathbb{P}_n\Big[\int_0^\tau Y(t)(h_n^s(t) - \overline{h_n^s}(t))^2 dt\Big] \\ &+ \lambda\sum_{j=1}^s \rho'(\|\boldsymbol{\theta}_{nj} + \alpha_0\boldsymbol{\theta}_{nj}^*\|)\frac{\boldsymbol{\theta}_{nj}^{*T}(\boldsymbol{\theta}_{nj} + \alpha_0\boldsymbol{\theta}_{nj}^*)}{\|\boldsymbol{\theta}_{nj} + \alpha_0\boldsymbol{\theta}_{nj}^*\|} \\ \triangleq& H_1 + H_2 + H_3, \end{aligned}$$

where $dM_n^s(t) = dN(t) - Y(t)[g_n^s(t) - \overline{g_n^s}(t)]dt$. Let $dM_0(t) = dN(t) - Y(t)[g_0(t) - \overline{g}_0(t)]dt$. It can be seen that

$$\begin{aligned} H_1 =& -\mathbb{P}_n\int_0^\tau (h_n^s(t) - \overline{h_n^s}(t))dM_0(t) \\ &+ \mathbb{P}_n\int_0^\tau Y(t)h_n^s(t)[(g_n^s(t) - g_0(t)) - (\overline{g_n^s}(t) - \overline{g}_0(t))]dt \\ \triangleq& J_1 + J_2, \end{aligned}$$

with $J_1 = O_p(n^{-1/2}s^2)$,

$$\begin{aligned} J_2 \leq& \mathbb{P}_n\int_0^\tau h_n^s(t)Y(t)dt \cdot [\|g_n^s - g_0\|_\infty + \|\overline{g_n^s} - \overline{g}_0\|_\infty] \\ \leq& \eta_n O_p(n^{-1/2}s^2 + s^2 q_n^{-1/2}) \\ =& O_p(s^2(n^{-(1/2+\varsigma)\nu} + n^{-1/2})), \end{aligned}$$

and $\eta_n = n^{-\nu\varsigma} + n^{-(1-\nu)/2}$ by Lemma A.5. Under the condition that $0.25/\varsigma < \nu < 0.5$, it follows that $|H_1| = o_p(s^2 q_n^{-1})$. Next we focus on H_2. Note that

$$\begin{aligned} H_2 =& \alpha_0(\mathbb{P}_n - \mathbb{P})\int_0^\tau Y(t){h_n^s}^2(t)dt + \alpha_0\mathbb{P}\int_0^\tau Y(t){h_n^s}^2(t)dt - \alpha_0\mathbb{P}_n\int_0^\tau Y(t)\overline{h_n^s}^2(t)dt \\ =& O_p(n^{-1/2}s^2) + C\alpha_0 s^2 q_n^{-1} + J_3, \end{aligned}$$

where

$$\begin{aligned}J_3 =&\alpha_0(\mathbb{P}_n - \mathbb{P})\int_0^\tau Y(t)\overline{h_n^s}^2(t)dt - \alpha_0\mathbb{P}\int_0^\tau Y(t)[\overline{h_n^s}^2(t) - e_h^2(t)]dt - \alpha_0\mathbb{P}\int_0^\tau Y(t)e_h^2(t)dt\\ =&O_p(n^{-1/2}s^2) + C\alpha_0 s^2 q_n^{-1}\end{aligned}$$

with $e_h^2(t) = s^{(1)}(h_n^s, t)/s^{(0)}(h_n^s, t)$. Therefore, $H_2 \geq Cs^2 q_n^{-1}$ for some constant $C > 0$.

At last, we consider H_3. Since $\|\boldsymbol{\theta}_{nj}^{*T}\boldsymbol{B}(Z_j(t))\| = O(q_n^{-1/2})$ and $\int_0^\tau \boldsymbol{B}(Z_j(t))\boldsymbol{B}(Z_j(t))^T dt = q_n^{-1} I_{q_n}$, we can get that $\|\boldsymbol{\theta}_{nj}^*\| = O(1)$. Thus,

$$|H_3| \leq \lambda\sum_{j=1}^{s}\rho'(\|\boldsymbol{\theta}_{nj} + \alpha_0\boldsymbol{\theta}_{nj}^*\|)\|\boldsymbol{\theta}_{nj}^*\| \leq \lambda s\rho'(\mathbf{0}+)O(1) = o(s^2 q_n^{-1})$$

by using the assumption that $\lambda = o(sq_n^{-1})$ and $\rho'(\mathbf{0}+) = C$.

Consequently, $H_n'(\alpha_0) \geq Cs^2 n^{-\nu} + o_p(s^2 n^{-\nu}) > 0$ with probability tending to one. Similarly, we can prove that $H_n'(-\alpha_0) < 0$ with probability tending to one. ■

Proof of Theorem 1. We denote $\lambda\rho'(\cdot) = \lambda_1\rho_1'(\cdot;\lambda_1) + \lambda_2\rho_2'(\cdot;\lambda_2)$, and obtain several important inequalities at first. By Lemma A.4, we have

$$\begin{aligned}&P(\|\boldsymbol{U}_{\mathcal{B}}(\boldsymbol{\theta}_n)\|_\infty \geq \lambda_2\rho_2'(\mathbf{0}+)|\Omega_\eta)\\ \leq&\sum_{j\in\mathcal{B}} P(\|\boldsymbol{U}_j(\boldsymbol{\theta}_n)\|_\infty \geq \lambda_2\rho_2'(\mathbf{0}+)|\Omega_\eta)\\ \leq&Cs_1 q_n\exp\{-L[(\lambda_2^2 n(1\wedge p^{-2}\eta_n^{-2}))\wedge(n+\eta_n^{-2})]\},\end{aligned}$$

and

$$P(\|\boldsymbol{U}_{\mathcal{D}}(\boldsymbol{\theta}_n)\|_\infty \geq \lambda\rho'(\mathbf{0}+)|\Omega_\eta) \leq C(p-s)q_n\exp\{-L[(\lambda^2 n(1\wedge p^{-2}\eta_n^{-2}))\wedge(n+\eta_n^{-2})]\}.$$

Moreover, Lemma A.5 implies that

$$P(\Omega_\eta^c) \leq p\exp(-L\eta_n^{-2}).$$

Thus, the following inequalities hold

$$\begin{aligned}&\|\boldsymbol{U}_{\mathcal{B}}(\boldsymbol{\theta}_n)\|_\infty \leq \lambda_2\rho_2'(\mathbf{0}+),\\ &\|\boldsymbol{U}_{\mathcal{D}}(\boldsymbol{\theta}_n)\|_\infty < \lambda\rho'(\mathbf{0}+),\ \Lambda_{\min}(\boldsymbol{V}_{\mathcal{A}\mathcal{A}}) > \lambda\kappa_0\end{aligned}\tag{A.10}$$

with probability at least

$$\begin{aligned}1 &- Cs_1 q_n\exp\{-L[(\lambda_2^2 n(1\wedge p^{-2}\eta_n^{-2}))\wedge(n+\eta_n^{-2})]\}\\ &- Cs^2 q_n^2\exp\Big(-L(n+\eta_n^{-2})\Big(\frac{\mu^2}{s^2}\wedge 1\Big)\Big)\\ &- C(p-s)q_n\exp(-L[(\lambda^2 n(1\wedge p^{-2}\eta_n^{-2}))\wedge(n+\eta_n^{-2})]) - p\exp(-L\eta_n^{-2}).\end{aligned}$$

Note that there exists $\hat{\boldsymbol{\theta}}$ minimizing $Q(\boldsymbol{\theta})$ in the subspace $\mathbb{B} = \{\boldsymbol{v} \in R^{pq_n} : \boldsymbol{v}_{\mathcal{D}} = \mathbf{0}\}$ by Lemma A.6. To show that $P(\hat{\boldsymbol{\theta}}_{\mathcal{D}} = 0) = 1$, it remains to prove that $Q(\boldsymbol{\theta}_1) > Q(\hat{\boldsymbol{\theta}})$ for any $\boldsymbol{\theta}_1 \in R^{pq_n}\backslash\mathbb{B}$ that lies in a sufficiently small neighborhood of $\hat{\boldsymbol{\theta}}$. To the end, it suffices to show that $Q(\boldsymbol{\theta}_1) > Q(\boldsymbol{\theta}_2)$ with $\boldsymbol{\theta}_2$ being the projection of $\boldsymbol{\theta}_1$ on to $\mathbb{B}$ since $Q(\boldsymbol{\theta}_2) > Q(\hat{\boldsymbol{\theta}})$ by the definition of $\hat{\boldsymbol{\theta}}$. By the mean value theorem, we have

$$\begin{aligned} Q(\boldsymbol{\theta}_1) - Q(\boldsymbol{\theta}_2) &= \sum_{j\in\mathcal{D}:\|\boldsymbol{\theta}_{1j}\|\neq 0} \left(\frac{\partial Q(\boldsymbol{\theta}^*)}{\boldsymbol{\theta}_j}\right)^T \boldsymbol{\theta}_{1j} \\ &\geq \sum_{j\in\mathcal{D}:\|\boldsymbol{\theta}_{1j}\|\neq 0} \Big(- \boldsymbol{U}_j(\boldsymbol{\theta}^*) + \lambda\rho'(\|\boldsymbol{\theta}_j^*\|)\boldsymbol{\theta}_j^*/\|\boldsymbol{\theta}_j^*\|\Big)^T \boldsymbol{\theta}_{1j}, \end{aligned}$$

where $\boldsymbol{\theta}^*$ is a point on the line segment between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Thus, with high probability, for $\boldsymbol{\theta}_1$ closed enough to $\hat{\boldsymbol{\theta}}$, we can get that

$$\begin{aligned} Q(\boldsymbol{\theta}_1) - Q(\boldsymbol{\theta}_2) &\geq \sum_{j\in\mathcal{D}:\|\boldsymbol{\theta}_{1j}\|\neq 0} [-\|\boldsymbol{U}_{\mathcal{D}}(\boldsymbol{\theta}^*)\|_\infty + \lambda\rho'(\|\boldsymbol{\theta}_j^*\|)]\|\boldsymbol{\theta}_{1j}\| \\ &\geq \sum_{j\in\mathcal{D}:\|\boldsymbol{\theta}_{1j}\|\neq 0} [-\lambda\rho'(\mathbf{0}+) + \lambda\rho'(\mathbf{0}+)]\|\boldsymbol{\theta}_{1j}\| \geq 0 \end{aligned}$$

by using the concentration inequalities (A.10). Thus, result (i) is concluded.

We then show result (ii) using the proofs by contradiction. We assume that $\|\hat{\boldsymbol{\theta}}_{j1-}\| \neq 0$ for some $j \in \mathcal{B}$ and let

$$\tilde{\boldsymbol{\theta}}_j = \begin{cases} \hat{\boldsymbol{\theta}}_j, & j \notin \mathcal{B} \\ (\hat{\theta}_{j1}, \mathbf{0})^T, & j \in \mathcal{B}. \end{cases}$$

Then similar to the proof of result (i), for some point $\boldsymbol{\theta}^*$ between $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$, we have

$$\begin{aligned} Q(\hat{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}) &\geq \sum_{j\in\mathcal{B}:\|\hat{\boldsymbol{\theta}}_{j1-}\|\neq 0} [-\|U_{\mathcal{B}}(\boldsymbol{\theta}^*)\|_\infty + \lambda_2\rho_2'(\|\boldsymbol{\theta}_{j1-}^*\|)]\|\hat{\boldsymbol{\theta}}_{j1-}\| \\ &\geq \sum_{j\in\mathcal{B}:\|\hat{\boldsymbol{\theta}}_{j1-}\|\neq 0} [-\lambda_2\rho_2'(\mathbf{0}+) + \lambda_2\rho_2'(\mathbf{0}+)]\|\hat{\boldsymbol{\theta}}_{j1-}\| \geq 0 \end{aligned}$$

by using the concentration inequalities (A.10) again. This contradicts with the fact that $\hat{\boldsymbol{\theta}}$ minimizes $Q(\boldsymbol{\theta})$.

At last, to decide L_2 loss of the estimators, we define

$$m_0(\boldsymbol{g}) = \big(- \int_0^\tau (g^s(t) - \overline{g^s}(t))dN(t) + \frac{1}{2}\int_0^\tau Y(t)(g^s(t) - \overline{g^s}(t))^2 dt\big)/s$$

and denote $M_0 = \mathbb{P}m_0$ and $M_n = \mathbb{P}_n m_0$, where $g^s(t) = \sum_{j=1}^{s} g_j(\phi_j, t)$. Let

$$\begin{aligned} W &= M_n(\boldsymbol{g}) - M_n(\boldsymbol{g}_n) - (M_0(\boldsymbol{g}) - M_0(\boldsymbol{g}_n)) \\ &= \mathbb{P}_n m_0(\boldsymbol{g}) - \mathbb{P}_n m_0(\boldsymbol{g}_n) - (\mathbb{P}m_0(\boldsymbol{g}) - \mathbb{P}m_0(\boldsymbol{g}_n)) \\ &= (\mathbb{P}_n - \mathbb{P})(m_0(\boldsymbol{g}) - m_0(\boldsymbol{g}_n)). \end{aligned}$$

By Lemma 3.4.2 of van der Vaart & Wellner (1996), $E\{\sup_{\|\frac{1}{s}(\boldsymbol{g}-\boldsymbol{g}_n)\|\le\eta_n}|W|\} = n^{-1/2}\eta_n q_n^{1/2}$. Then by Theorem 3.4.1 of van der Vaart & Wellner (1996), taking the distance $d(\hat{\boldsymbol{g}}_n, \boldsymbol{g}_n) = -[\mathbb{P}m_0(\hat{\boldsymbol{g}}_n) - \mathbb{P}m_0(\boldsymbol{g}_n)]$, we have $-k_{1n}^2[\mathbb{P}m_0(\hat{\boldsymbol{g}}_n) - \mathbb{P}m_0(\boldsymbol{g}_n)] = O(1)$, where $k_{1n} = O(n^{1/2}q_n^{-1/2}) = O(n^{(1-\nu)/2})$. Therefore, $\mathbb{P}m_0(\hat{\boldsymbol{g}}_n) - \mathbb{P}m_0(\boldsymbol{g}_n) = O(n^{-(1-\nu)})$. Thus, similar to Lemma A6 in Huang (1999), we can get that $\|(\hat{\boldsymbol{g}}_n - \boldsymbol{g}_n)/s\|^2 = O_p(n^{-2\nu\varsigma} + n^{-(1-\nu)})$. Combining this with the result in Lemma A.5 that $\|\boldsymbol{g}_n - \boldsymbol{g}_0\|_\infty^2 = O_p(s^2(n^{-2\nu\varsigma} + n^{-(1-\nu)}))$, we have $\|\hat{\boldsymbol{g}}_n - \boldsymbol{g}_0\|^2 = O_p(s^2(n^{-2\nu\varsigma} + n^{-(1-\nu)}))$. By Condition 2, it follows that

$$E\big\|\boldsymbol{Z}_{\mathcal{C}}(t)(\hat{\boldsymbol{\phi}}_n(\boldsymbol{Z}_{\mathcal{C}}(t)) - \boldsymbol{\phi}_0(\boldsymbol{Z}_{\mathcal{C}}(t)) + \boldsymbol{Z}_{\mathcal{B}}(t)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\big\|^2 = O(s^2(n^{-2\nu\varsigma} + n^{-(1-\nu)})).$$

Denoting the projection of $\boldsymbol{Z}_{\mathcal{B}}$ on $\boldsymbol{Z}_{\mathcal{C}}$ as $\tilde{\boldsymbol{Z}}_{\mathcal{C}}$, we have

$$\begin{aligned}
&E\big\|(\boldsymbol{Z}_{\mathcal{B}}(t) - \tilde{\boldsymbol{Z}}_{\mathcal{C}}(t))(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + \tilde{\boldsymbol{Z}}_{\mathcal{C}}(t)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + \boldsymbol{Z}_{\mathcal{C}}(t)(\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0)\big\|^2\\
=&E\big\|(\boldsymbol{Z}_{\mathcal{B}}(t) - \tilde{\boldsymbol{Z}}_{\mathcal{C}}(t))(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\big\|^2 + E\big\|\tilde{\boldsymbol{Z}}_{\mathcal{C}}(t)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + \boldsymbol{Z}_{\mathcal{C}}(t)(\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0)\big\|^2\\
=&O(s^2(n^{-2\nu\varsigma} + n^{-(1-\nu)})).
\end{aligned}$$

By Condition 3, we obtain $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|^2 = O_p(s^2(n^{-(1-\nu)} + n^{-2\nu\varsigma}))$. This in turn implies that $E\big\|\boldsymbol{Z}_{\mathcal{C}}(t)(\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0)\big\|^2 = O(s^2(n^{-(1-\nu)} + n^{-2\nu\varsigma}))$. Therefore, $\|\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0\|^2 = O_p(s^2(n^{-(1-\nu)} + n^{-2\nu\varsigma}))$. This completes the proof of Theorem 1. ■

Proof of Theorem 2. Define

$$\begin{aligned}
U(\boldsymbol{Z},T;\boldsymbol{\beta},\hat{\boldsymbol{\phi}}) \triangleq& -\int_0^\tau (\boldsymbol{Z}_{\mathcal{B}}(t) - \overline{\boldsymbol{Z}}_{\mathcal{B}}(t))dN(t)\\
&+\int_0^\tau Y(t)(\boldsymbol{Z}_{\mathcal{B}}(t) - \overline{\boldsymbol{Z}}_{\mathcal{B}}(t))\Big(\sum_{j\in\mathcal{C}} Z_j(t)(\hat{\phi}_j(t) - \overline{\hat{\phi}_j}(t)) + \boldsymbol{\beta}^T(\boldsymbol{Z}_{\mathcal{B}}(t) - \overline{\boldsymbol{Z}}_{\mathcal{B}}(t))\Big)dt,
\end{aligned}$$

and

$$\hat{U}_n(\boldsymbol{\beta}) \triangleq \frac{1}{n}\sum_{i=1}^n U(\boldsymbol{Z}_i, T_i; \boldsymbol{\beta}, \hat{\boldsymbol{\phi}}_n).$$

Then we have $\hat{U}_n(\hat{\boldsymbol{\beta}}) = 0$ since $\hat{\boldsymbol{\beta}}$ is the root of $\partial Q(\boldsymbol{\beta}, \hat{\boldsymbol{\phi}}_n)/\partial\boldsymbol{\beta} = 0$.

Let $U_n(\boldsymbol{\beta}) \triangleq \frac{1}{n}\sum_{i=1}^n U(\boldsymbol{Z}_i, T_i; \boldsymbol{\beta}, \boldsymbol{\phi}_0)$ and $\tilde{\boldsymbol{\beta}}$ be the root of $U_n(\boldsymbol{\beta}) = 0$. We now show that $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ have the same distribution by using Lemma 5.1 of Newey (1994). To the end, we first note that $\|(\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0)/s\| = O_p(n^{-(1-\nu)/2} + n^{-\nu\varsigma}) = o_p(n^{-1/4})$, which ensures that assumption 5.1 in Newey (1994) is satisfied. Then the Fréchet derivative of $U(\boldsymbol{Z}, T; \boldsymbol{\beta}_0, \boldsymbol{\phi})$ at $\boldsymbol{\phi}_0$ in the direction $\boldsymbol{h}$ is

$$\begin{aligned}
D(\boldsymbol{Z},T;\boldsymbol{h}) &= \lim_{\alpha\to 0}\frac{U(\boldsymbol{Z},T;\boldsymbol{\beta}_0,\boldsymbol{\phi}_0 + \alpha\boldsymbol{h}) - U(\boldsymbol{Z},T;\boldsymbol{\beta}_0,\boldsymbol{\phi}_0)}{\alpha}\\
&= \int_0^\tau Y(t)(\boldsymbol{Z}_{\mathcal{B}}(t) - \overline{\boldsymbol{Z}}_{\mathcal{B}}(t))(\boldsymbol{Z}_{\mathcal{C}}(t) - \overline{\boldsymbol{Z}}_{\mathcal{C}}(t))^T\boldsymbol{h}(t)dt
\end{aligned}$$

with $\boldsymbol{h} \in \{h_1 + \ldots + h_{s_2}, h_j \in \mathcal{H}, \quad j = 1, \ldots, s_2\}$. Thus, $\sqrt{n}(\mathbb{P}_n - \mathbb{P})\{D(\boldsymbol{Z}, T; \hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0)\} \xrightarrow{p} 0$ by Lemma 3.4.2 of van der Vaart & Wellner (1996). It follows that the stochastic equicontinuity assumption 5.2 in Newey (1994) holds. At last, since a straightforward calculation yields that $ED(\boldsymbol{Z}, T; \boldsymbol{\phi} - \boldsymbol{\phi}_0) = 0$ for $\boldsymbol{\phi}$ close enough to $\boldsymbol{\phi}_0$, the mean square continuity assumption 5.3 holds with $\alpha(\boldsymbol{Z}, T) = 0$.

Next, we derive the asymptotic distribution of $\tilde{\boldsymbol{\beta}}$. Let $\iota_n = n^{-1/2}$ and $V_{1n}(\boldsymbol{a}) = Q(\boldsymbol{\beta}_0 + \iota_n(\boldsymbol{a}^T, \boldsymbol{0}^T)^T, \boldsymbol{\phi}_0) - Q(\boldsymbol{\beta}_0, \boldsymbol{\phi}_0)$, where $\boldsymbol{a} = (a_1, \ldots, a_{s_1})^T$ is a s_1-dimensional constant vector and $\boldsymbol{0}$ is a $(p - s)$-dimensional zero vector. Note that

$$\begin{aligned}
V_{1n}(\boldsymbol{a}) =& Q(\boldsymbol{\beta}_0 + \iota_n(\boldsymbol{a}^T, \boldsymbol{0}^T)^T, \boldsymbol{\phi}_0) - Q(\boldsymbol{\beta}_0, \boldsymbol{\phi}_0) \\
=& \Big(\iota_n \boldsymbol{a}^T U_n(\boldsymbol{\beta}_0) + \frac{\iota_n^2}{2} \boldsymbol{a}^T \nabla_{\boldsymbol{\beta}} U_n(\boldsymbol{\beta}_0) \boldsymbol{a}\Big) \\
&+ \sum_{j=1}^{s_1} \Big(\lambda_1 \rho_1(|\tilde{\theta}_{j1}|; \lambda_1) - \lambda_1 \rho_1(|\theta_{0j1}|; \lambda_1)\Big) \\
\triangleq& A_{1n}(\boldsymbol{a}) + A_{2n}(\boldsymbol{a}).
\end{aligned}$$

Since $\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 = \iota_n \boldsymbol{a} = q_n^{-1/2}(\tilde{\boldsymbol{\theta}}_{\mathcal{B}1} - \boldsymbol{\theta}_{0\mathcal{B}1})$, we have

$$\begin{aligned}
A_{2n}(\boldsymbol{a}) &= \sum_{j=1}^{s_1} \lambda_1 \left[\rho_1'(|\theta_{0j1}|; \lambda_1)\mathrm{sgn}(\theta_{0j1}) + o_p(1)\right] q_n^{1/2} \iota_n a_j \\
&\leq C \sum_{j=1}^{s_1} \lambda_1 \rho_1'(0+) q_n^{1/2} \iota_n a_j = O_p(s_1 \lambda_1 n^{-(1-\nu)/2}) = o_p(1).
\end{aligned}$$

In addition,

$$\begin{aligned}
nA_{1n}(\boldsymbol{a}) &= \boldsymbol{a}^T \left(\sqrt{n} U_n(\boldsymbol{\beta}_0)\right) + \frac{1}{2} \boldsymbol{a}^T \nabla_{\boldsymbol{\beta}} U_n(\boldsymbol{\beta}_0) \boldsymbol{a} \\
&\triangleq \boldsymbol{a}^T T_1 + \boldsymbol{a}^T T_2 \boldsymbol{a},
\end{aligned}$$

where $T_2 \xrightarrow{p} \boldsymbol{\Sigma}_1$ with $\boldsymbol{\Sigma}_1 = E\Big(\int_0^\tau Y(t)(\boldsymbol{Z}_{\mathcal{B}}(t) - \overline{\boldsymbol{Z}}_{\mathcal{B}}(t))^{\otimes 2} dt\Big)$, and $\boldsymbol{u}\boldsymbol{\Sigma}_2^{-1/2} T_1$ is asymptotically distributed by $N(0, 1)$ for any $\boldsymbol{u} \in \mathbb{R}^{s_1}$ with $\|\boldsymbol{u}\| = 1$ with $\boldsymbol{\Sigma}_2 = Var(\int_0^\tau (\boldsymbol{Z}_{\mathcal{B}}(t) - \overline{\boldsymbol{Z}}_{\mathcal{B}}(t)) dM(\boldsymbol{g}_0, t))$. Let $\hat{\boldsymbol{a}} = \text{argmin}\{V_1(\boldsymbol{a}) = \boldsymbol{a}^T T_1 + \frac{1}{2} \boldsymbol{a}^T \boldsymbol{\Sigma}_1 \boldsymbol{a} : \boldsymbol{a} \in \mathbb{R}^{s_1}\}$. According to the continuous mapping theorem of Kim & Pollard (1990), $\sqrt{n}\boldsymbol{u}\boldsymbol{\Sigma}^{-1/2}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ has the same standard normal distribution asymptotically as $\boldsymbol{u}\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{a}}$ for any $\boldsymbol{u} \in \mathbb{R}^{s_1}$ with $\|\boldsymbol{u}\| = 1$, where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}_2 \boldsymbol{\Sigma}_1^{-1}$. This completes the proof of Theorem 2. ■

BIBLIOGRAPHY

Bradic, J. & Song, R. (2015). Structured estimation for the nonparametric Cox model. *Electron. J. Statist.*, 9, 492–534.

Cao, Y., Huang, J., Liu, Y., & Zhao, X. (2016). Sieve estimation of Cox models with latent structures. *Biometrics*, 72, 1086–1097.

Chen, J. & Chen, Z. (2008). Extended bayesian information criterion for model selection with large model spaces. *Biometrika*, 95, 759–771.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96, 1348–1360.

Fan, J. & Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.*, 30, 74–99.

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

Hao, M., Liu, K.-Y., Xu, W., & Zhao, X. (2020). Semiparametric inference for the functional Cox model. *J. Amer. Statist. Assoc.*, DOI: 10.1080/01621459.2019.1710155.

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

Honda, T. & Yabe, R. (2017). Variable selection and structure identification for varying coefficient Cox models. *J. Multivar. Anal.*, 161, 103–122.

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

Huang, J., Wei, F., & Ma, S. (2012). Semiparametric regression pursuit. *Statist. Sinica*, 22, 1403–1426.

Kim, J. & Pollard, D. B. (1990). Cube root asymptotics. *Ann. Statist.*, 18, 191–219.

Kong, D., Ibrahim, J. G., Lee, E., & Zhu, H. (2018). FLCRM: Functional linear Cox regression model. *Biometrics*, 74, 109–117.

Lee, E., Zhu, H., Kong, D., & Wang, Y., Giovanello, K. S., & Ibrahim, J. G. (2015). BFLCRM: A Bayesian functional linear Cox regression model for predicting time to conversion to Alzheimer's disease. *Ann. Statist.*, 9, 2153–2178.

Leng, C. & Ma, S. (2007). Path consistent model selection in additive risk model via Lasso. *Statist. Med.*, 26, 3753–3770.

Lian, H., Lai, P., & Liang, H. (2013). Partially linear structure selection in Cox models with varying coefficients. *Biometrics*, 69, 348–357.

Lin, W. & Lv, J. (2013). High-dimensional sparse additive hazards regression. *J. Amer. Statist. Assoc.*, 108, 247–264.

Lin, D. & Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, 81, 61–71.

Ma, S. & Huang, J. (2005). Lasso method for additive risk models with high dimensional covariates. Technical report, Department of Statistics and Actuarial Science, University of Iowa, Iowa.

Ma, S., Kosorok, M. R., & Fine, J. P. (2006). Additive risk models for survival data with high-dimensional covariates. *Biometrics*, 62, 202–210.

Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.*, 28, 863–884.

Martinussen, T. & Scheike, T. H. (2009). Covariate selection for the semiparametric additive risk model. *Scand. J. Statist.*, 36, 602–619.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 62, 1349–1382.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. Series B*, 58, 267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statist. Med.*, 16, 385-395.

Tseng, P & Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Math. Program. Ser. B.*, 117, 387–423.

van der Vaart, A. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer: New York.

Wang, S. & Xiang, L. (2017). Penalized empirical likelihood inference for sparse additive hazards regression with a diverging number of covariates. *Statist. Comput.*, 27, 1347–1364.

Xie, X., Strickler, H. D., & Xue, X. (2013). Additive hazard regression models: an application to the natural history of human papillomavirus. *Comput. Math. Methods Med.*, 2013, Article ID 796270.

Yan, J. & Huang, J. (2012). Model selection for Cox models with time-varying coefficients. *Biometrics*, 68, 419-428.

Yang, Y. & Zou, H. (2015). A fast unified algorithm for solving group-lasso penalized learning problems. *Statist. Comput.* , 25, 1129–1141.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38, 894–942.

Zhang, H. H., Cheng, G., & Liu, Y. (2011). Linear or nonlinear? Automatic structure discovery for partially linear models. *J. Amer. Statist. Assoc.*, 106, 1099–1112.

Zhang, H. & Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model, *Biometrika*, 94, 691–703.

Zhang, H., Sun, L., Zhou, Y., & Huang, J. (2017). Oracle inequalities and selected consistency for weighted lasso in high-dimensional additive hazards model. *Statist. Sinica*, 27, 1903–1920.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101, 1418–1429.

TABLE 1: : Rates of each component being selected as an important variable by the penalized methods with the group LASSO (GL), adaptive group LASSO (AGL), group SCAD (GS), and group MCP (GM) over 200 replications, and the L_1 prediction error (PE) of $\hat{\boldsymbol{g}}_n$ with the sample standard deviation of PE for sample size $n = 500$ and model size $p = 15$ in Example 1.

Method	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	PE
GL	1	1	1	.235	.265	.240	.290	.270	.245	.250	.250	.240	.275	.270	.250	1.04(.32)
AGL	1	1	1	.015	.035	.025	.025	.015	.025	.035	.040	.035	.015	.040	.030	1.03(.33)
GS	1	1	.990	.015	.025	.025	.030	.010	.020	.040	.035	.030	.010	.045	.030	1.03(.32)
GM	1	1	1	.015	.030	.030	.030	.015	.025	.035	.040	.035	.015	.040	.035	1.03(.32)
true	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0(0)

TABLE 2: : Rates of each component being identified as having a nonlinear effect by the penalized methods with the group LASSO (GL), adaptive group LASSO (AGL), group SCAD (GS), and group MCP (GM) over 200 replications for sample size $n = 500$ and model size $p = 15$ in Example 1.

Method	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
GL	.025	.070	.995	.005	.005	0	0	0	0	.005	0	0	.005	.005	0
AGL	.015	.050	.995	.005	0	0	0	0	0	0	0	0	.005	0	0
GS	.015	.070	.980	.005	0	0	0	0	0	.005	0	0	0	.005	0
GM	.025	.050	.995	.005	0	0	0	0	0	0	0	0	.005	0	0
true	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Received 9 July 2009
Accepted 8 July 2010

TABLE 3: : Rates of each component being selected as an important variable by the penalized methods with the group LASSO (GL), adaptive group LASSO (AGL), group SCAD (GS), and group MCP (GM) over 200 replications, and the L_1 prediction error (PE) with the sample standard deviation of $\hat{\boldsymbol{g}}_n$ for sample size $n = 500$ and model size $p = 15$ in Example 2.

Method	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	PE
GL	.960	.995	.865	.995	1	.700	.455	.440	.420	.425	.420	.370	.400	.390	.360	1.05(.37)
AGL	.960	.995	.865	.995	1	.080	.035	.065	.045	.045	.045	.055	.050	.050	.050	1.05(.37)
GS	.960	.995	.850	.995	1	.085	.030	.070	.040	.040	.035	.055	.050	.040	.050	1.05(.38)
GM	.960	.995	.860	.995	1	.065	.030	.065	.050	.035	.035	.065	.055	.050	.060	1.05(.38)
true	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0(0)

TABLE 4: : Rates of each component being selected as an important variable by the penalized methods with the group LASSO (GL), adaptive group LASSO (AGL), group SCAD (GS), and group MCP (GM) over 100 replications; the L_1 prediction error (PE) for $\hat{\boldsymbol{g}}_n$; the true positive rate (TPR) representing the rate that the important variables are selected; and the false positive rate (FPR) representing the rate that the unimportant variables are selected for sample size $n = 500$ and model sizes $p = 500, 1000$ in Example 2.

Method	p	x_1	x_2	x_3	x_4	x_5	PE	TPR	FPR
GL	500	.94	.99	.93	1	1	1.51(.21)	.972	.092
	1000	.96	.99	.93	1	1	1.53(.24)	.976	.047
AGL	500	.94	.99	.92	1	1	1.49(.21)	.970	.064
	1000	.95	.99	.91	1	1	1.53(.25)	.970	.035
GS	500	.95	.99	.91	1	1	1.50(.21)	.970	.061
	1000	.95	.96	.90	1	1	1.54(.24)	.962	.031
GM	500	.94	.99	.92	1	1	1.51(.21)	.970	.063
	1000	.94	.98	.93	1	1	1.53(.26)	.970	.033
true		1	1	1	1	1	0(0)	1	0

TABLE 5: : Rates of each component being identified as having a nonlinear effect by the penalized methods with the group LASSO (GL), adaptive group LASSO (AGL), group SCAD (GS), and group MCP (GM) over 200 replications for sample size $n = 500$ and model size $p = 15$ in Example 2.

Method	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
GL	.025	.045	.050	.970	.940	.020	.010	.005	.005	.010	.005	.020	.010	.020	.005
AGL	.005	.025	.020	.980	.940	.015	.015	0	0	.005	.005	.010	.005	.010	.005
GS	.030	.050	.035	.970	.975	.010	0	.005	.005	0	0	.010	.010	.010	.005
GM	.035	.045	.035	.985	.955	.005	.005	0	0	0	0	.010	.015	.010	.010
true	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0

TABLE 6: : Rates of each important component being identified as having a nonlinear effect by the penalized methods with the group LASSO (GL), adaptive group LASSO (AGL), group SCAD (GS), and group MCP (GM); TPRN representing the rate that an important variable with a nonlinear effect is correctly selected; and FPRN representing the rate that an unimportant variable or an important variable with a linear effect is wrongly detected as having a nonlinear effect over 100 replications for sample size $n = 500$ and model size $p = 500, 1000$ in Example 2.

Method	p	x_1	x_2	x_3	x_4	x_5	TPRN	FPRN
GL	500	.12	.30	.16	1	1	1	.015
	1000	.18	.29	.14	.98	1	.990	.008
AGL	500	.09	.22	.13	1	1	1	.013
	1000	.16	.25	.12	.99	1	.995	.007
GS	500	.14	.35	.22	1	1	1	.014
	1000	.22	.35	.19	.99	1	.995	.008
GM	500	.19	.39	.18	.99	1	.995	.015
	1000	.21	.40	.20	.94	1	.985	.008
true		0	0	0	1	1	1	0

TABLE 7: : Rates of each component being selected as an important variable by the penalized methods with the group LASSO (GL), adaptive group LASSO (AGL), group SCAD (GS), and group MCP (GM) over 200 replications, and the L_1 prediction error (PE) with the sample standard deviation of $\hat{\boldsymbol{g}}_n$ for sample size $n = 500$ and model size $p = 15$ in Example 3.

Method	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	PE
GL	1	1	.995	1	.055	.070	.045	.050	.050	.065	.070	.030	.075	.050	.060	2.60(.14)
AGL	1	1	.995	1	.045	.060	.040	.065	.055	.065	.055	.035	.075	.050	.060	2.60(.13)
GS	1	1	.990	1	.015	.050	.040	.020	.025	.050	.025	.050	.060	.030	.045	2.62(.12)
GM	1	1	.950	1	.030	.055	.045	.035	.035	.055	.040	.060	.065	.050	.065	2.62(.13)
true	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0(0)

TABLE 8: : Rates of each component being identified as having a nonlinear effect by the penalized methods with the group LASSO (GL), adaptive group LASSO (AGL), group SCAD (GS), and group MCP (GM) over 200 replications for $\hat{\boldsymbol{g}}_n$ for sample size $n = 500$ and model size $p = 15$ in Example 3.

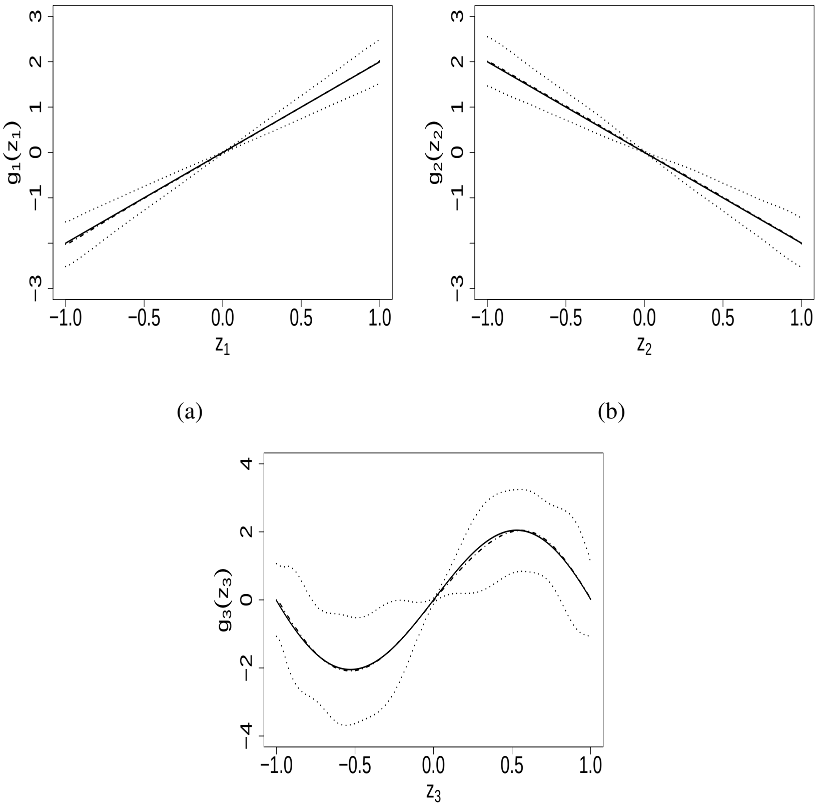
Method	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
GL	.035	.050	.040	.830	.025	.050	.020	.040	.035	.040	.040	.025	.045	.050	.045
AGL	.025	.045	.020	.815	.020	.050	.020	.055	.040	.050	.030	.030	.050	.050	.045
GS	.050	.050	.045	.820	.015	.045	.030	.020	.025	.040	.025	.050	.060	.030	.045
GM	.070	.070	.080	.830	.030	.055	.045	.035	.035	.055	.040	.060	.065	.050	.065
true	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

TABLE 9: : Estimation results for linear regression coefficients of important covariates by the adaptive group LASSO, including the estimated bias, the sample standard deviation (SSE), the estimated standard error (ESE), and the mean squared error (MSE) for sample size $n = 500$ and model size $p = 15$ in Examples 1-3.

	β_1			β_2			β_3		
	Bias(SSE)	ESE	MSE	Bias(SSE)	ESE	MSE	Bias(SSE)	ESE	MSE
Ex1	.022(.226)	.248	.249	$-.006(.260)$	.273	.273			
Ex2	$-.002(.189)$	.227	.227	.042(.263)	.256	.259	.039(.178)	.208	.212
Ex3	$-.033(.158)$	.158	.161	.065(.209)	.190	.201	-.026(.154)	.160	.162

TABLE 10: : Variable selection and structure identification results by the penalized methods with the group LASSO (GL), adaptive group LASSO (AGL), group SCAD (GS), and group MCP (GM) for the PBC data; 0/1/2 represent unimportant/linear/nonlinear effects for covariates.

Covariate	GL	AGL	GS	GM
TRT	1	0	0	0
Age	1	1	1	1
Sex	1	0	0	0
Ascites	1	1	1	1
Hepato	1	0	0	0
Spiders	1	1	1	1
Edema	1	1	1	1
Stage	1	0	0	0
Bili	2	2	2	2
Chol	2	0	2	2
Albumin	2	2	2	2
Copper	2	2	2	2
Alk.phos	2	2	2	2
SGOT	2	2	2	2
Trig	2	2	2	2
Platelet	2	2	2	2
Protime	2	2	2	2

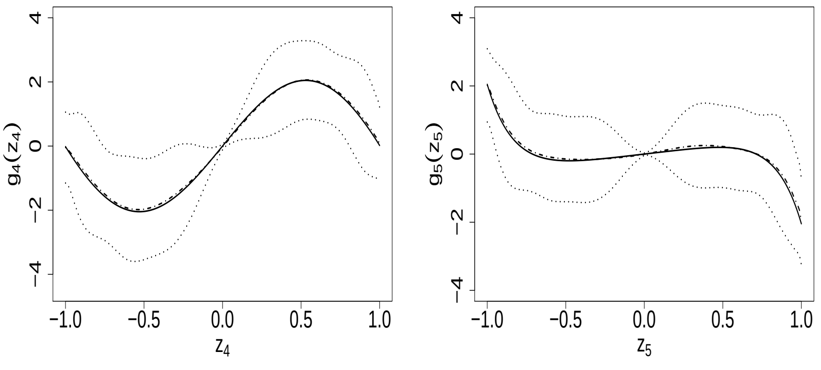


(a)

(b)

(c)

FIGURE 1: : Estimates of functions by using adaptive group LASSO for the important variables in Example 1; (a), (b) and (c) show the estimated functional plots of variables z_1, z_2 and z_3, respectively. The solid line is the pointwise mean estimate from 200 Monte Carlo repetitions, the dot and dash line is the true function, and the dotted lines are the 95% pointwise confidence intervals.



(a) (b)

FIGURE 2: : Estimates of important nonlinear effects identified by using adaptive group LASSO in Example 2; (a) and (b) show the estimated functional plots of variables z_4 and z_5, respectively. The solid line is the pointwise mean estimate from 200 Monte Carlo repetitions, the dot and dash line is the true function, and the dotted lines are the 95% pointwise confidence intervals.

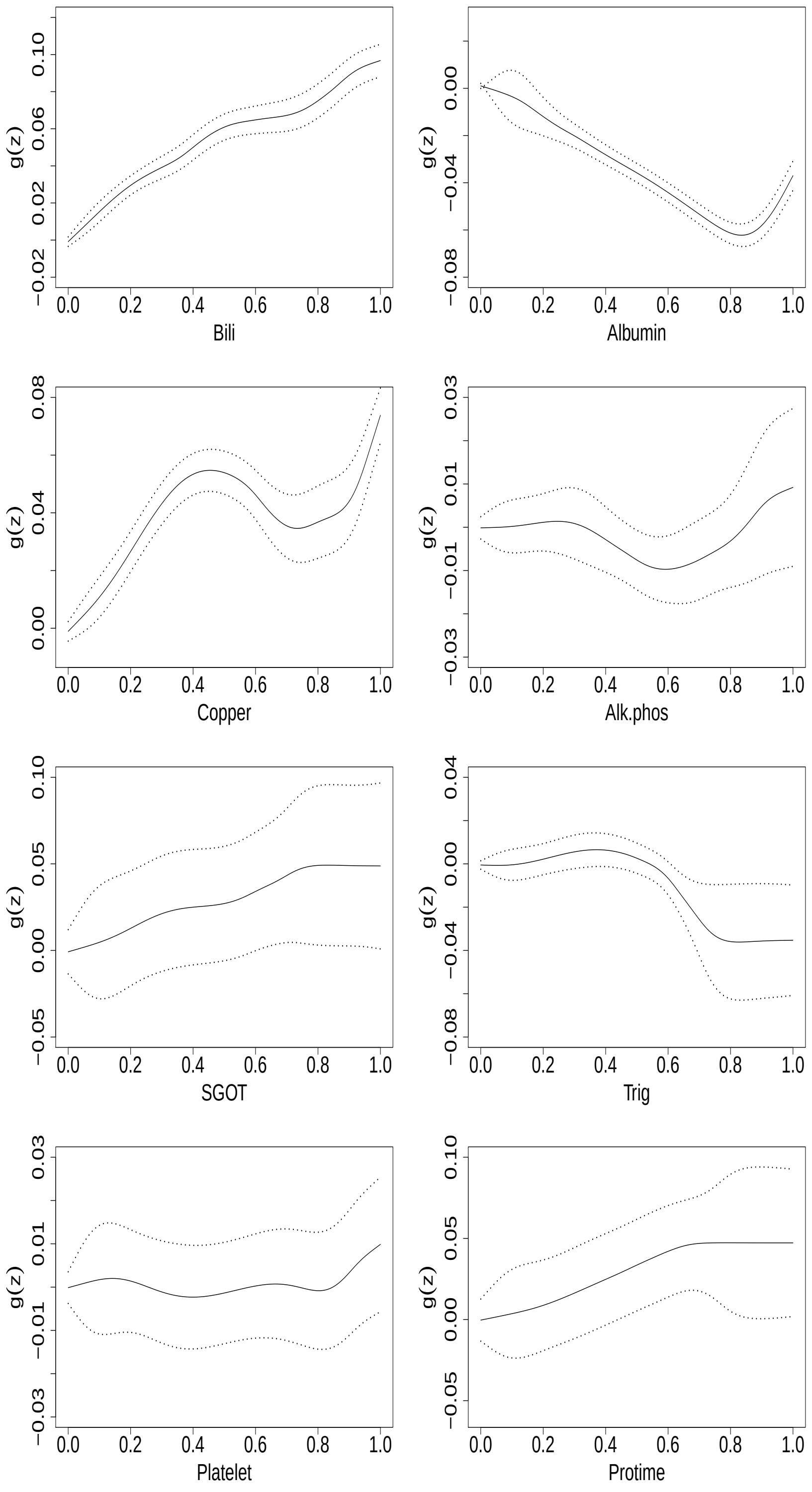


FIGURE 3: : Estimates of important nonlinear effects identified by using adaptive group LASSO for PBC data.