Biometrics 72, 1086–1097
December 2016

DOI: 10.1111/biom.12529

Sieve Estimation of Cox Models with Latent Structures

Yongxiu Cao,[1,2] Jian Huang,[3,4] Yanyan Liu,[1] and Xingqiu Zhao[5,*]

[1]School of Mathematics and Statistics, Wuhan University, Wuhan, China
[2]School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, China
[3]Department of Statistics and Actuarial Science, University of Iowa, Iowa City, Iowa, U.S.A.
[4]School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China
[5]Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong
**email:* xingqiu.zhao@polyu.edu.hk

Summary. This article considers sieve estimation in the Cox model with an unknown regression structure based on right-censored data. We propose a semiparametric pursuit method to simultaneously identify and estimate linear and nonparametric covariate effects based on B-spline expansions through a penalized group selection method with concave penalties. We show that the estimators of the linear effects and the nonparametric component are consistent. Furthermore, we establish the asymptotic normality of the estimator of the linear effects. To compute the proposed estimators, we develop a modified blockwise majorization descent algorithm that is efficient and easy to implement. Simulation studies demonstrate that the proposed method performs well in finite sample situations. We also use the primary biliary cirrhosis data to illustrate its application.

Key words: Group selection; Model-pursuit consistency; Modified blockwise majorization descent algorithm; Partially linear Cox model; Penalized partial likelihood.

1. Introduction

The proportional hazards model (Cox, 1972) is widely used in the analysis of censored survival data. This model specifies that covariates have log-linear effects on the hazard function of survival time. However, the true covariate effects may be more complex than a log-linear effect in practice. Such examples include the well-known primary biliary cirrhosis (PBC) data that motivated this research. This data set can be found in the Appendix D of Fleming and Harrington (1991). Fleming and Harrington (1991) and Grambsch, Therneau, and Fleming (1995) among others explored functional form of covariates in the Cox model using residual plots. Huang (1999) proposed a more parsimonious and interpretable partially linear Cox model which specifies the covariate effects through a partially linear structure, i.e., a combination of linear and nonparametric additive parts in the Cox model. Cai, Fan, Jiang, and Zhou (2007) extended the model to a multivariate case. Du, Ma, and Liang (2010) studied the variable selection with the nonparametric component being estimated using smoothing splines.

The aforementioned work assumed that it is known in advance which covariates have a linear effect and which have a nonlinear effect on the logarithm of the hazards function. However, this is rarely known in advance. If a nonlinear effect is misspecified to be linear, it will cause a bias in the estimation. On the other hand, if a linear effect is misspecified as nonlinear, it increases model complexity and leads to loss of efficiency. Therefore, it is desirable to correctly determine the linear and nonlinear components in the model.

In the context of partially linear regression models with complete data, several authors have considered the problem of identifying linear and nonlinear components. For example, Zhang, Cheng, and Liu (2011) proposed a two-step approach using smoothing splines for determining the zero, linear, and nonlinear components. They obtained the rate of convergence of their proposed estimator and showed that their method is selection consistent in the special case of tensor product design. Huang, Wei, and Ma (2012) proposed a concave group selection approach for determining linear and nonlinear components. They provided sufficient conditions under which their proposed approach can correctly determine which covariates have a linear effect and which do not with high probability.

To the best of our knowledge, this has not been studied in the context of the Cox model. Because of the wide applications of the Cox model and its central role in survival analysis, it is important to also have a systematic approach that can correctly determine the form of covariate effects in this model. In our proposed approach, we first take unknown nonlinear covariate effects as a nonparametric additive form, and then use B-splines to approximate unknown smooth functions in the model. With this approximation, we can transform the problem of model specification into a group selection problem. Using a penalized approach with concave penalties, we can detect the linear or nonlinear components of covariate effects and estimate both parametric and nonparametric components simultaneously. Furthermore, we show that, with probability tending to one, the proposed method can correctly specify the linear and nonlinear covariate effects on the log relative risk hazard function. In addition, it enjoys an asymptotic oracle property in the sense that it performs as well as the oracle estimator obtained by assuming the underlying model structure is known in advance.

An additional contribution of this article is that we develop a fast convergent algorithm for implementing the proposed approach. For a group variable selection problem, the calculation is challenging because the solution paths in selecting groups are not piecewise linear. To overcome this difficulty, Yuan and Lin (2006) proposed a blockwise descent (BD) algorithm for the group-LASSO (least absolute shrinkage and selection operator) penalized least squares by following the idea of Fu (1998). Meier, van de Geer, and Bühlmann (2008) developed a block coordinate gradient descent (BCGD) algorithm for solving the group-LASSO penalized logistic regression. Both the BD and BCGD algorithms need the groupwise orthogonality condition, which is usually violated in practice. The detailed discussions can be found in Yang and Zou (2015), who proposed a blockwise majorization decent (BMD) algorithm for solving the general group-LASSO learning problems under the condition that the loss function satisfies a quadratic majorization (QM) condition. However, the BMD depends on the maximum eigenvalue of a quadratic majorization matrix. If the maximum eigenvalue is large, the algorithm requires a large number of iterations to achieve convergence. In order to increase the computation speed, we introduce the backtracking line search approach to properly shrink the maximum eigenvalue to a reasonable value so that a more accurate quadratic majorization is used in the BMD. We refer to our method as modified BMD (MBMD). The MBMD is easy to implement since it uses a closed-form expression at each iteration. Our numerical studies also demonstrate that MBMD has good numerical performance.

The reminder of this article is organized as follows. In Section 2, we formulate the model structure estimation problem into a group selection problem and propose the penalized spline-based partial-likelihood method and describe the MBMD algorithm. The theoretical properties of the proposed approach are also presented in this section. In Section 3, we conduct simulation studies to evaluate the finite-sample performance of the penalized estimator. In Section 4, we analyze the PBC data primary biliary to illustrate the utility of the method. Some concluding remarks are made in Section 5. The technical proofs are given in the Supplementary Materials.

2. Estimation Procedures and Asymptotic Results

2.1. *Group-Penalized Spline-Based Partial-Likelihood Method*

Let T^u and T^c denote the potential survival time and censoring time, respectively. The observed random variable is $(T, \Delta, \boldsymbol{X}) \in \mathbb{R}^+ \times \{0, 1\} \times \mathbb{R}^d$, where $T = \min\{T^u, T^c\}$, $\Delta = I(T^u \leq T^c)$. Here, $I(\cdot)$ is the indicator function and $\boldsymbol{X} = (X_1, \dots, X_d)$ is a d-dimensional vector of covariates. We assume that T^u and T^c are conditionally independent given $\boldsymbol{X}$. Suppose we observe $\{(T_i, \Delta_i, \boldsymbol{X}_i) : i = 1, \dots, n\}$ that are independent and identically distributed as $(T, \Delta, \boldsymbol{X})$. We embed the partially linear additive Cox model into the nonparametric additive Cox model (Hastie and Tibshirani, 1986)

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t) \exp(g(\boldsymbol{X})), \tag{1}$$

with $g(\boldsymbol{x}) = \sum_{j=1}^d g_j(x_j)$, where $x_1, \dots, x_d$ are the elements of $\boldsymbol{x}$. The main goal is to determine which g_j's take a linear form and which do not. For this purpose, we decompose g_j into a sum of linear and nonlinear terms

$$g_j(x) = \beta_j x + \phi_j(x). \tag{2}$$

Thereby, $g(\boldsymbol{x})$ can be rewritten as

$$g(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta} + \boldsymbol{\phi}(\boldsymbol{x}),$$

with $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)^{\mathrm{T}}$ and $\boldsymbol{\phi}(\boldsymbol{x}) = \sum_{j=1}^d \phi_j(x_j)$. If some g_j's are linear, the corresponding nonparametric parts $\boldsymbol{\phi}_j$'s should be zero. For the identifiability in models (1) and (2), we assume that $\mathrm{E}[\Delta\phi_j(X_j)] = 0, 1 \leq j \leq d$ and $\mathrm{E}[\Delta X_j] = 0, 1 \leq j \leq d$ because the regression functions can only be identified up to a constant and centering can remove this ambiguity (Huang, 1999). For the nonlinear term, the smoothness assumption is also often used in nonparametric curve estimation. These assumptions are given in Supplemental Materials, and a detailed explanation about them can be also found in Huang (1999).

We use the B-splines to approximate the nonparametric components ϕ_j, $j = 1 \dots, d$. Assume that the covariate $\boldsymbol{X}$ takes values in $[a, b]^d$ where a and b are two finite real numbers. Let $a = \xi_0 \leq \xi_1 \dots \leq \xi_{K_n} \leq \xi_{K_n+1} = b$ be a partition of $[a, b]$ into K_n subintervals, where $K_n = O(n^\nu)$ with $0 < \nu < 0.5$ is a positive integer such that $\max\limits_{1 \leq j \leq K_n+1} |\xi_j - \xi_{j-1}| = O(n^{-\nu})$. Denote by $I_{K_n t} = [\xi_t, \xi_{t+1})$, $t = 0, \dots, K_n - 1$ and $I_{K_n K_n} = [\xi_{K_n}, \xi_{K_n+1}]$. Let $\mathcal{S}_n$ be the space of polynomial splines of order $m \geq 1$ which consists of functions h satisfying the following two conditions: (i) the restriction of h to $I_{K_n t}$ is a polynomial of order m for $t = 1, \dots, K_n$; (ii) for $m \geq 2$ and $0 \leq m' \leq m - 2$, h is m' times continuously differentiable on $[a, b]$.

Let Φ_n be the collection of functions $\boldsymbol{\phi}$ on $[a, b]^d$ with the additive form $\boldsymbol{\phi}(\boldsymbol{x}) = \sum_{j=1}^d \phi_j(x_j)$, where each component ϕ_j belongs to $\mathcal{S}_n$. By Schumaker (1981), there exists a local basis $\{\psi_k, k = 1, \dots, q_n\}$ for $\mathcal{S}_n$, where $q_n = K_n + m$ is the number of the basis. Thus for $\phi_j \in \mathcal{S}_n$, we can write

$$\boldsymbol{\phi}_j(x_j) = \sum_{k=1}^{q_n} \theta_{jk} \psi_k(x_j), \quad j = 1, \dots, d.$$

Under some suitable smoothness assumptions, the true nonparametric parts ϕ_{0j}'s can be well approximated by the functions in $\mathcal{S}_n$ (see, for example, Lemma A5 of Huang (1999)).

Let $\|\cdot\|$ and $\|\cdot\|_2$ denote the Euclidean norm and the L_2-norm with respect to a probability measure, respectively. Furthermore, $\|\cdot\|_\infty$ denotes the supremum norm. If $\theta_{jk} = 0$ for all $1 \leq k \leq q_n$, then the function g_j takes a linear form. Therefore, the problem now becomes determining which groups of $\boldsymbol{\theta}_j = (\theta_{j1}, \dots, \theta_{jq_n})^{\mathrm{T}}$ $(j = 1, \dots, d)$ are zeros. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \dots, \boldsymbol{\theta}_d^{\mathrm{T}})^{\mathrm{T}}$

and $Y_i(t) = I(T_i \geq t)$. To identify the linear and nonlinear structure in model (1), we estimate $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ via the group-penalized partial-likelihood function defined as

$$\widetilde{Q}_n(\boldsymbol{\beta}, \boldsymbol{\theta}) = \widetilde{\ell}_n(\boldsymbol{\beta}, \boldsymbol{\theta}) + \sum_{j=1}^{d} P(\|\boldsymbol{\theta}_j\|; \lambda), \tag{3}$$

where

$$\begin{aligned}\widetilde{\ell}_n(\boldsymbol{\beta}, \boldsymbol{\theta}) = & -n^{-1} \sum_{i=1}^{n} \Delta_i \sum_{j=1}^{d} \left(X_{ij}\beta_j + \sum_{k=1}^{q_n} \theta_{jk} \psi_k(X_{ij}) \right) \\ & + n^{-1} \sum_{i=1}^{n} \Delta_i \log \left\{ \sum_{k=1}^{n} Y_k(T_i) \exp \left[\sum_{j=1}^{d} \left(X_{kj}\beta_j \right.\right.\right. \\ & \left.\left.\left. + \sum_{t=1}^{q_n} \theta_{jt} \psi_t(X_{kj}) \right) \right] \right\}\end{aligned}$$

is the negative logarithm of partial-likelihood function for the observed right-censored data and $P(\|\boldsymbol{\theta}_j\|; \lambda)$ is the penalty function for vector $\boldsymbol{\theta}_j$ with a tuning parameter $\lambda \geq 0$. We employ the concave penalties to increase flexibility of our method. These concave penalties are applied to the Euclidean norm $||\boldsymbol{\theta}_j||$ such that the coefficients in $\boldsymbol{\theta}_j$ are selected as a group. Fan and Lv (2011) defined the concave penalties through the following Condition 1.

Condition 1. *Let $\rho(\nu; \lambda) = \lambda^{-1} P(\nu; \lambda)$. The $\rho(\nu; \lambda)$ is increasing and concave in $\nu \in [0, \infty)$ and has a continuous derivative $\rho'(\nu; \lambda)$ with $\rho'(0_+; \lambda) > 0$. In addition, $\rho'(\nu; \lambda)$ is increasing in $\lambda \in (0, \infty)$, and $\rho'(0_+; \lambda)$ is independent of λ.*

Note that most commonly used penalties such as the LASSO (Tibshirani, 1996, 1997), SCAD penalty (Fan and Li, 2001), and MCP (Zhang, 2010) satisfy Condition 1.

For a given λ, the group-penalized partial-likelihood solution is defined by

$$(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\theta}}_n) = \arg\min_{(\boldsymbol{\beta}, \boldsymbol{\theta})} \widetilde{Q}_n(\boldsymbol{\beta}, \boldsymbol{\theta}).$$

Then, the penalized estimator of ϕ_{0j} is given by $\widehat{\phi}_{nj}(x_j) = \sum_{j=1}^{d} b(x_j)^T \widehat{\boldsymbol{\theta}}_{nj}$ for $j = 1, \ldots, d$, where $b(\cdot) = (\psi_1(\cdot), \ldots, \psi_{q_n}(\cdot))^T$ is the spline basis vector.

2.2. *Modified BMD Algorithm*

We propose a modified blockwise majorization descent algorithm (MBMD). To describe our method, we first recall the QM condition required for the BMD algorithm. For simplicity of notation, the loss function is written as $L(\boldsymbol{\theta}|D)$ in this subsection with D representing the data, then the objective function is

$$G(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|D) + \sum_{j=1}^{d} P(\|\boldsymbol{\theta}_j\|; \lambda),$$

where $P(\|\boldsymbol{\theta}_j\|; \lambda)$ is the penalty function as defined before, and the penalized estimator $\widehat{\boldsymbol{\theta}}_n$ is the minimizer of $G(\boldsymbol{\theta})$.

Definition 1. *The loss function is said to satisfy the QM condition, if and only if the following two assumptions hold:*

(i) $L(\boldsymbol{\theta}|D)$ is differentiable as a function of $\boldsymbol{\theta}$, i.e., $\nabla L(\boldsymbol{\theta}|D)$ exists everywhere;

(ii) There exists a $\dim(\boldsymbol{\theta}) \times \dim(\boldsymbol{\theta})$ matrix H, which does not depend on $\boldsymbol{\theta}$, such that for all $\boldsymbol{\theta}$ and $\widetilde{\boldsymbol{\theta}}$

$$\begin{aligned}L(\boldsymbol{\theta}|D) \leq & \, L(\widetilde{\boldsymbol{\theta}}|D) + (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})^{\mathrm{T}} \nabla L(\widetilde{\boldsymbol{\theta}}|D) \\ & + \frac{1}{2} (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})^{\mathrm{T}} H (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}),\end{aligned} \tag{4}$$

where $\dim(\boldsymbol{\theta})$ denotes the dimension of vector $\boldsymbol{\theta}$.

We next give a rough outline about the BMD algorithm. Suppose the current value of $\boldsymbol{\theta}$ is $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\theta}}_1, \ldots, \widetilde{\boldsymbol{\theta}}_{j-1}, \widetilde{\boldsymbol{\theta}}_j, \widetilde{\boldsymbol{\theta}}_{j+1}, \ldots, \widetilde{\boldsymbol{\theta}}_d)$. Write $\boldsymbol{\theta}^*$ as $\boldsymbol{\theta}^* = (\widetilde{\boldsymbol{\theta}}_1, \ldots, \widetilde{\boldsymbol{\theta}}_{j-1}, \boldsymbol{\theta}_j, \widetilde{\boldsymbol{\theta}}_{j+1}, \ldots, \widetilde{\boldsymbol{\theta}}_d)$. Let $H^{(j)}$ be the corresponding sub-matrix of H corresponding to the jth group and h_j be the largest eigenvalue of $H^{(j)}$. Define $\boldsymbol{u}_j = \partial L(\widetilde{\boldsymbol{\theta}}|D)/\partial \boldsymbol{\theta}_j$. By (4), we have

$$\begin{aligned}L(\boldsymbol{\theta}^*|D) & \leq L(\widetilde{\boldsymbol{\theta}}|D) + (\boldsymbol{\theta}_j - \widetilde{\boldsymbol{\theta}}_j)^{\mathrm{T}} \boldsymbol{u}_j + \frac{1}{2} (\boldsymbol{\theta}_j - \widetilde{\boldsymbol{\theta}}_j)^{\mathrm{T}} H^{(j)} (\boldsymbol{\theta}_j - \widetilde{\boldsymbol{\theta}}_j) \\ & \leq L(\widetilde{\boldsymbol{\theta}}|D) + (\boldsymbol{\theta}_j - \widetilde{\boldsymbol{\theta}}_j)^{\mathrm{T}} \boldsymbol{u}_j + \frac{h_j}{2} \|\boldsymbol{\theta}_j - \widetilde{\boldsymbol{\theta}}_j\|^2.\end{aligned} \tag{5}$$

Generally, there is no closed form for the solution of objective function $G(\boldsymbol{\theta})$. Yang and Zou (2015) proposed a BMD algorithm, that is, instead of minimizing the objective function $G(\boldsymbol{\theta})$, the optimal $\boldsymbol{\theta}_j$ is defined as

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{nj}(h_j) = \arg\min_{\boldsymbol{\theta}_j} & \left\{ L(\widetilde{\boldsymbol{\theta}}|D) + (\boldsymbol{\theta}_j - \widetilde{\boldsymbol{\theta}}_j)^{\mathrm{T}} \boldsymbol{u}_j + \frac{h_j}{2} \|\boldsymbol{\theta}_j - \widetilde{\boldsymbol{\theta}}_j\|^2 \right. \\ & \left. + P(\|\boldsymbol{\theta}_j\|; \lambda) \right\}.\end{aligned}$$

By a simple calculation, we obtain that $\widehat{\boldsymbol{\theta}}_{nj}(h_j)$ is the minimizer of $m_j(\boldsymbol{\theta}_j; h_j)$, where

$$\begin{aligned}m_j(\boldsymbol{\theta}_j; h_j) & = \frac{1}{2} \|\boldsymbol{\theta}_j - \widetilde{\boldsymbol{\theta}}_j\|^2 + \frac{1}{h_j} \boldsymbol{u}_j^{\mathrm{T}} (\boldsymbol{\theta}_j - \widetilde{\boldsymbol{\theta}}_j) \\ & \quad + \frac{1}{h_j} P(\|\boldsymbol{\theta}_j\|; \lambda) + \frac{\|\boldsymbol{u}_j\|^2}{2h_j^2} \\ & = \frac{1}{2} \|\boldsymbol{\theta}_j - (\widetilde{\boldsymbol{\theta}}_j - \boldsymbol{u}_j/h_j)\|^2 + \frac{1}{h_j} P(\|\boldsymbol{\theta}_j\|; \lambda).\end{aligned}$$

Thereby, the updated value of $\boldsymbol{\theta}$ can be written as

$$\boldsymbol{\theta}(h_j) = (\widetilde{\boldsymbol{\theta}}_1, \ldots, \widetilde{\boldsymbol{\theta}}_{j-1}, \widehat{\boldsymbol{\theta}}_{nj}(h_j), \widetilde{\boldsymbol{\theta}}_{j+1}, \ldots, \widetilde{\boldsymbol{\theta}}_d).$$

Generally, $\widehat{\boldsymbol{\theta}}_{nj}(h_j)$ has a closed form for the commonly used penalties such as LASSO, SCAD, and MCP.

For the group-LASSO penalty with $P_{\mathrm{LASSO}}(\|\boldsymbol{\theta}_j\|;\lambda) = \lambda\|\boldsymbol{\theta}_j\|$, $\widehat{\boldsymbol{\theta}}_{nj}(h_j) = S(\boldsymbol{c}_j;\lambda/h_j)$, where $S(\boldsymbol{c}_j;\lambda) = (1-\lambda/\|\boldsymbol{c}_j\|)_+\boldsymbol{c}_j$ with $\boldsymbol{c}_j = \widetilde{\boldsymbol{\theta}}_j - \boldsymbol{u}_j/h_j$. For the group-SCAD penalty with

$$P_{\mathrm{SCAD}}(\|\boldsymbol{\theta}_j\|;\lambda) = \begin{cases} \lambda\|\boldsymbol{\theta}_j\|, & \|\boldsymbol{\theta}_j\| \le \lambda, \\ \dfrac{2\gamma\lambda\|\boldsymbol{\theta}_j\| - \|\boldsymbol{\theta}_j\|^2 - \lambda^2}{2(\gamma-1)}, & \lambda < \|\boldsymbol{\theta}_j\| \le \gamma\lambda, \\ (\gamma^2-1)\lambda^2/(2(\gamma-1)), & \|\boldsymbol{\theta}_j\| > \gamma\lambda, \end{cases}$$

$$\widehat{\boldsymbol{\theta}}_{nj}(h_j) = \begin{cases} S(\boldsymbol{c}_j;\lambda/h_j), & \|\boldsymbol{c}_j\| \le \lambda + \lambda/h_j, \\ \dfrac{\left[h_j(\gamma-1) - \frac{\gamma\lambda}{\|\boldsymbol{c}_j\|}\right]\boldsymbol{c}_j}{(h_j\gamma - h_j - 1)}, & \lambda + \lambda/h_j < \|\boldsymbol{c}_j\| \le \gamma\lambda, \\ \boldsymbol{c}_j, & \|\boldsymbol{c}_j\| > \gamma\lambda. \end{cases}$$

For the group-MCP penalty with

$$P_{\mathrm{MCP}}(\|\boldsymbol{\theta}_j\|;\lambda) = \begin{cases} \lambda\|\boldsymbol{\theta}_j\| - \dfrac{\|\boldsymbol{\theta}_j\|^2}{2\gamma}, & \|\boldsymbol{\theta}_j\| \le \lambda\gamma, \\ \dfrac{\lambda^2\gamma}{2}, & \|\boldsymbol{\theta}_j\| > \lambda\gamma, \end{cases}$$

$$\widehat{\boldsymbol{\theta}}_{nj}(h_j) = \begin{cases} S\left(\dfrac{h_j\boldsymbol{c}_j}{h_j - 1/\gamma}; \dfrac{\lambda}{h_j - 1/\gamma}\right), & \|\boldsymbol{c}_j\| \le \lambda\gamma, \\ \boldsymbol{c}_j, & \|\boldsymbol{c}_j\| > \lambda\gamma. \end{cases}$$

Obviously, different values of h_j can make the updated values $\widehat{\boldsymbol{\theta}}_{nj}(h_j)$ different. A large value of h_j will increase the number of iterations to convergence. Note that the only role of h_j is to guarantee that the inequality (5) holds for $\boldsymbol{\theta}(h_j)$ and $\widetilde{\boldsymbol{\theta}}$. If we can find a smaller $\widetilde{h}_j$ such that the inequality also holds for $\boldsymbol{\theta}(\widetilde{h}_j)$ and $\widetilde{\boldsymbol{\theta}}$, then the number of iterations should be reduced. Inspired by this idea, we propose to use the backtracking line search approach to find smaller $\widetilde{h}_j$. Since the largest eigenvalue h_j of $H^{(j)}$ guarantees inequality (5), the initial value of $\widetilde{h}_j$ can be taken as h_j. Our modified BMD (MBMD) algorithm is summarized as follows.

MBMD Algorithm:

```
Step 1. For a known matrix $H$, compute $h_j$ for $j = 1, \ldots, d$.
Step 2. Initialize $\widetilde{\boldsymbol{\theta}}$ and choose $\eta$ with $0 < \eta < 1$.
Step 3. Repeat the following cyclic blockwise updates until convergence:
    (i) Set $\widetilde{h}_j = h_j$, calculate $\boldsymbol{u}_j = \partial L(\widetilde{\boldsymbol{\theta}}|D)/\partial\widetilde{\boldsymbol{\theta}}_j$ for $\widetilde{h}_j$, and find the minimizer $\widehat{\boldsymbol{\theta}}_{nj}(\widetilde{h}_j)$ of $m_j(\boldsymbol{\theta}_j;\widetilde{h}_j)$.
    (ii) Update $h_j = \widetilde{h}_j * \eta$, and find the minimizer $\widehat{\boldsymbol{\theta}}_{nj}(h_j)$ of $m_j(\boldsymbol{\theta}_j;h_j)$.
    (iii) If inequality (5) holds for $\widetilde{\boldsymbol{\theta}}$, $h_j$ and $\boldsymbol{\theta} = \boldsymbol{\theta}(h_j)$, set $\widetilde{h}_j = h_j$ and return to (ii); otherwise, go to (iv).
    (iv) Set
        $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\theta}}_1^T, \ldots, \widetilde{\boldsymbol{\theta}}_{j-1}^T, \widehat{\boldsymbol{\theta}}_{nj}^T(\widetilde{h}_j), \widetilde{\boldsymbol{\theta}}_{j+1}^T, \ldots, \widetilde{\boldsymbol{\theta}}_d^T)^T.$
```

Obviously, this algorithm depends on the choice of η. We suggest $\eta = 0.6$ via simulation study, and this value of η works well in the following numerical studies. The descent property of MBMD algorithm is proved in Supplemental Materials. Clearly, this algorithm can be applied to general group selection problems.

2.3. *Implementation*

In the objective function, there are two parts of parameters $(\boldsymbol{\beta}, \boldsymbol{\theta})$ where only parameter $\boldsymbol{\theta}$ is penalized, and the nonpenalized parameter can be treated as a pseudo-penalized one based on the same penalty with the tuning parameter $\lambda_0 = 0$ and the group size of one. Denote the ith observation of covariates as $\tilde{x}_i = (x_{i1}, \ldots, x_{id}, \psi_1(x_{i1}), \ldots, \psi_{q_n}(x_{i1}), \ldots, \psi_1(x_{id}), \ldots, \psi_{q_n}(x_{id}))^{\mathrm{T}}$. By the work of Böhning and Lindsay (1988), we conclude that $\widetilde{\ell}_n(\boldsymbol{\beta}, \boldsymbol{\theta})$ defined in Section 2.1 satisfies the QM condition with the square matrix H taking the following form

$$H = \frac{1}{2n}\sum_{i=1}^{n}\sum_{l=1}^{n}\Delta_i Y_l(T_i)\tilde{x}_i\tilde{x}_l^{\mathrm{T}}.$$

Write this square matrix H as

$$H = \begin{pmatrix} H^{(11)} & H^{(12)} \\ H^{(21)} & H^{(22)} \end{pmatrix},$$

where $H^{(11)}$ is a $d \times d$ matrix with respect to $\boldsymbol{\beta}$, and $H^{(22)}$ is the $(dq_n) \times (dq_n)$ matrix with respect to $\boldsymbol{\theta}$. Let $\widetilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_d)^{\mathrm{T}}$ and $\widetilde{\boldsymbol{\theta}}$ be the current values of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, respectively. Let s_j be the jth diagonal element of matrix $H^{(11)}$. Then, $\tilde{\beta}_j$ can be updated by

$$\tilde{\beta}_j(s_j) = \tilde{\beta}_j - \frac{1}{s_j}\frac{\partial\widetilde{\ell}_n(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\theta}})}{\partial\tilde{\beta}_j}. \tag{6}$$

The implementation of the proposed MBMD algorithm for computing $(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\theta}}_n)$ is summarized as follows.

```
Step 1. Compute $H$, and obtain $s_j$ and $h_j$ for $j = 1, \ldots, d$.
Step 2. Initialize $(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\theta}})$.
Step 3. Repeat the following cyclic blockwise updates until convergence
    (3.1) For the current value of $\widetilde{\boldsymbol{\theta}}$, update $\tilde{\beta}_j$ by formula (6) for $j = 1, \ldots, d$.
    (3.2) For the current value of $\widetilde{\boldsymbol{\beta}}$, update $\widetilde{\boldsymbol{\theta}}$ by Step 3 of the MBMD algorithm described in Section 2.2.
```

Remark 1. *Certainly, one can use a similar technique to find smaller s_j's to accelerate the calculation. Since the main purpose is to select significant groups, we fix the value of s_j for computational simplicity.*

2.4. *Tuning Parameter Selection*

To select tuning parameter λ, we can use the generalized cross-validation (GCV) (Craven and Wahba, 1979) criterion. Since the main purpose here is to identify the linear or nonlinear form of covariate effects, the GCV criterion is approximated by

$$\mathrm{GCV}(\lambda) = \frac{\widetilde{\ell}_n(\boldsymbol{\beta}, \boldsymbol{\theta})}{\{1 - d(\lambda)/n\}^2} \qquad (7)$$

with $d(\lambda) \approx \sum_{j=1}^{d} \|\widehat{\boldsymbol{\theta}}_{nj}\|^0$ being the number of estimated active groups. The optimal value of $\hat{\lambda}$ is the minimizer of GCV(λ) over a grid of values for λ.

For the folded concave penalties with two tuning parameters, the value of another tuning parameter γ can be chosen as commonly used. For example, we choose $\gamma = 3.7$ for the SCAD penalty (Fan and Li, 2001, 2002), and determine $\gamma = \frac{2}{1-\max_{i \neq j} x_i^{\mathrm{T}} x_j / n}$ for the MCP penalty following Zhang (2010).

2.5. *Theoretical Properties*

We present the results on the rate of convergence, the model-pursuit consistency, and asymptotic normality of the proposed estimator for the parametric component, and relegate the proofs to the Supplementary Materials.

Denote by $\boldsymbol{\beta}_0$ and $\boldsymbol{\phi}_0$ the true value of $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$. For $m = 0, 1, 2$, define

$$S^{(m)}(t; \boldsymbol{\beta}, \boldsymbol{\phi}) = \mathrm{E}\left[Y(t)\boldsymbol{X}^{\bigotimes m} \exp(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\phi}(\boldsymbol{X}))\right],$$

with $\boldsymbol{X}^{\bigotimes 0} = 1$, $\boldsymbol{X}^{\bigotimes 1} = \boldsymbol{X}$ and $\boldsymbol{X}^{\bigotimes 2} = \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$, and

$$\Sigma = \int_0^{\tau} \left[\frac{S^{(2)}(t; \boldsymbol{\beta}_0, \boldsymbol{\phi}_0)}{S^{(0)}(t; \boldsymbol{\beta}_0, \boldsymbol{\phi}_0)} - \left\{ \frac{S^{(1)}(t; \boldsymbol{\beta}_0, \boldsymbol{\phi}_0)}{S^{(0)}(t; \boldsymbol{\beta}_0, \boldsymbol{\phi}_0)} \right\}^{\bigotimes 2} \right] \times S^{(0)}(t; \boldsymbol{\beta}_0, \boldsymbol{\phi}_0) d\Lambda_0(t).$$

In this section, we use λ_n rather than λ to emphasize its dependence on n.

THEOREM 1. *Suppose assumptions (A1)–(A6) stated in the Supplementary Materials hold. If the tuning parameter* $\lambda_n = o(n^{-\nu})$ *for* $0 < \nu < 0.5$, *then* $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_p(n^{-\nu p} + n^{-(1-\nu)/2})$, *and* $\|\widehat{\phi}_{nj} - \phi_{0j}\|_2 = O_p(n^{-\nu p} + n^{-(1-\nu)/2}), 1 \leq j \leq d$.

Specifically, if $\nu = 1/(2p+1)$, the rate convergence of $\widehat{\phi}_{nj}$ is $n^{-p/(2p+1)}$ which achieves the optimal rate in nonparametric regression. The following theorem states that the convergence rate of $\widehat{\boldsymbol{\beta}}_n$ achieves $n^{-1/2}$ under some regularity conditions.

THEOREM 2. *Under assumptions (A1)–(A8) stated in the Supplementary Materials, if*

$$0.25/p < \nu < 0.5 \qquad \textit{and} \qquad \nu(p+q) > 0.5$$

where p *is the measure of smoothness of* ϕ_{0j} *defined in (A1) and* q *is defined in (A7), then*

(i) (Group Sparsity) $\lim_{n\to\infty} P(\widehat{\boldsymbol{\theta}}_{nj} = 0 : j = s+1, \ldots, d) = 1;$

(ii) (Asymptotic Normality) $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \to_d N(0, \Sigma^{-1})$.

The group sparsity property shows that our proposed method can identify the linear or nonlinear component with a high probability. Although the convergence rate of $\widehat{\phi}_{nj}$ is slower than $n^{-1/2}$, $\widehat{\boldsymbol{\beta}}_n$ achieves $n^{1/2}$–consistency and is asymptotically normal. For the estimator $\widehat{\phi}_{nj}$, the optimal value of ν is $\frac{1}{2p+1}$. Clearly, Theorem 2 holds for $\nu = \frac{1}{2p+1}$. Therefore, for this choice of ν, both the estimators of parametric and nonparametric components achieve the optimal rates of convergence.

3. Simulation Studies

Simulation studies were conducted to evaluate the finite-sample properties of the proposed penalized estimators via the group-LASSO, the group-SCAD, and the group-MCP penalties. In the study, the survival time was generated from model (1) with $\lambda_0(t) = 1$ or $\lambda_0(t) = 2t$. Following Huang, Wei, and Ma (2012), the d dimensional covariates were taken as $x_i = (w_i + u)/2$ for $i = 1, \ldots, d$, where $w_1, \ldots, w_d$ and u were generated from the uniformly distribution $U[0, 1]$. The censoring time was generated from the uniformly distribution on $[0, c]$, where c was chosen to yield approximately 20 and 40% censoring, respectively. Following Fan and Li (2001, 2002), the tuning parameter γ for the group-SCAD penalty was set as $\gamma = 3.7$. Using the idea of Zhang (2010), the tuning parameter γ for the group-MCP penalty was determined by $\gamma = \frac{2}{1-\max_{i \neq j} x_i^{\mathrm{T}} x_j / n}$. The GCV criterion was applied to select the tuning parameter λ, wherein the degrees of freedom $d(\lambda)$ in (7) was approximated by the number of effective groups selected by our proposed method. Note that the proposed MBMD algorithm in Section 2 depends on selection of η. From the extensive simulation studies, we found that the MBMD algorithm with $\eta = 0.6$ always works well for different situations. To speed up the convergence, we used the KKT condition to discard unrelated predictors by utilizing the idea of Tibshirani et al. (2012)and Yang and Zou (2012). In the following, we present two simulations. The purpose of the first study is to assess the performance of proposed model structure estimation methods and compare the MBMD and the BMD algorithms, while the goal of the second study is to compare the estimation results for the parametric component by our method and the standard method. All the simulation results are based on 1000 replications with sample sizes $n = 100$ and $n = 200$, and the final estimates were reached at convergence.

3.1. *Evaluation of Model Structure Estimation*

In this subsection, we focus on checking whether the proposed method can correctly identify linear and nonlinear effects on the log-risk of the Cox model and comparing the MBMD and the BMD algorithms. Assume that g in model (1) took the form

$$g(\boldsymbol{x}) = f_1(x_1) + 1.5f_1(x_2) - 0.8f_1(x_3) + 2f_2(x_4) + 3f_3(x_5) + 3f_4(x_6),$$

where the four functions were defined on $[0, 1]$ with

$$f_1(x) = x, \;\; f_2(x) = \sin(2\pi x), f_3(x) = 9x^2 - 6x,$$
$$f_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin{(2\pi x)}^2 + 0.4\cos{(2\pi x)}^3 + 0.5\sin{(2\pi x)}^3.$$

In this model, the first three variables have linear effect and the remaining variables have nonlinear effect on the logarithm of hazard function. To fit the simulated data, we considered model (1) with $g(\boldsymbol{x}) = \sum_{j=1}^{6} \beta_j x_j + \sum_{j=1}^{6} \phi_j(x_j)$ as defined in (2), and used cubic-splines with seven B-spline basis functions to approximate each ϕ_j. Then for each data set, the penalized estimates $\widehat{\beta}_j$'s and $\widehat{\phi}_j$'s can be obtained by using the procedure described in Section 2.

The model error is one of the appropriate measures for the goodness-of-fit of the model. For a general regression model with $E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x})$, the model error of a predictor $\widehat{\mu}(\boldsymbol{x})$ is defined as $\mathrm{ME}(\widehat{\mu}) = E\{\widehat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x})\}^2$. For the Cox model (1),

$$\mu(\boldsymbol{x}) = E(T^u|\boldsymbol{x}) = \int_0^\infty t\lambda_0(t)\exp\{g(\boldsymbol{x})\}\exp \times \left\{-\int_0^t \lambda_0(u)\exp\{g(\boldsymbol{x})\}du\right\} dt.$$

By some straightforward calculations, we have $\mu(\boldsymbol{x}) = \exp\{-g(\boldsymbol{x})\}$ for $\lambda_0(t) = 1$ and $\mu(\boldsymbol{x}) = \frac{\sqrt{\pi}}{2}\exp\{-g(\boldsymbol{x})/2\}$ for $\lambda_0(t) = 2t$. Thus, the model error for the two cases in the simulation can be defined as

$$\mathrm{ME}(\widehat{\mu}) = \mathrm{E}[\widehat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x})]^2,$$

where $\widehat{\mu}(\boldsymbol{x})$ is the corresponding estimate of μ with $g(\boldsymbol{x})$ being replaced by the proposed estimate $\widehat{g}(\boldsymbol{x})$. On the other hand, we used the Cox proportional hazards model with $g(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x}$ to fit the simulated data and obtained $\widetilde{g}(\boldsymbol{x}) = \exp(\widetilde{\boldsymbol{\beta}}^T\boldsymbol{x})$ with $\widetilde{\boldsymbol{\beta}}$ being the maximum partial-likelihood estimate of $\boldsymbol{\beta}$. The model error from this fit is denoted by $\mathrm{ME}(\widetilde{\mu})$. The relative model error (RME) of the fitted model with $\widehat{g}$ to the fitted one with $\tilde{g}$ is defined as

$$\mathrm{RME}(\widehat{\mu}, \widetilde{\mu}) = \frac{\mathrm{ME}(\widehat{\mu})}{\mathrm{ME}(\widetilde{\mu})}.$$

Following Fan and Li (2002), we employ the median of the relative model error (MRME) over 1000 simulations rather than the mean of it as a compared measurement due to the consideration of stability.

Let $\mathcal{A} = \{j : \phi_{0j}(x_j) \neq 0, j = 1, \ldots, d\}$ be the true index set of covariates with nonlinear effect to the log-risk function, and $\widehat{\mathcal{A}} = \{j : \hat{\phi}_j(x_j) \neq 0, j = 1, \ldots, d\}$ be the corresponding estimated index set. Let $|\mathcal{A}|$ denote the cardinality of set $\mathcal{A}$. Define $R_+ = |\widehat{\mathcal{A}} - \mathcal{A}|/d$ representing the ratio of the number of covariates with linear effects erroneously selected as having nonlinear effects to d, and $R_- = |\mathcal{A} - \widehat{\mathcal{A}}|/d$ representing the ratio of the number of the covariates with nonlinear effects being wrongly detected as having linear effects to d.

The results with the group-LASSO (GLASSO), the group-SCAD (GSCAD), and the group-MCP (GMCP) by the MBMD and the BMD algorithms with $n = 200$ are summarized in Table 1, while those with $n = 100$ are omitted for the sake of space. In addition to the median of RME and the average values of R_+, R_-, and $|\hat{\mathcal{A}}|$ over 1000 runs, the table includes the percentage of occasions on which the exactly nonlinear components are selected (Correct%), and the average number of iterations needed for the convergence and the average running time for computing the penalized estimator at a given λ over 1000 replications.

It can be seen from Table 1 that the proposed detection procedures with three different penalties are comparable to each other. All the three penalized detection methods perform well. They can correctly identify the log-linear and log-nonlinear effects on the hazard rate function with high probability. The MRME's are less than 1; this result suggests that the proposed methods perform better than the classical Cox's proportional hazards regression method which omits the nonlinear effect. For all the three methods, with the sample size increasing from 100 to 200, the percentages of selecting the exactly correct model (Correct%) increase, as expected. The number of times of each component being identified as a nonlinear function is reported in Table 2, which shows the good performance of the proposed method in detecting the nonlinear components.

Clearly, the simulation results in the two tables demonstrated that the proposed MBDM algorithm outperforms the BMD algorithm for the group LASSO, SCAD, and MCP penalized procedures. In addition, both group SCAD and MCP penalized methods have similar performance and perform better than the group LASSO penalized method. In the following simulations and applications, we only present the obtained results by the MBDM algorithm.

Note that the number of basis functions, q_n, needs to be prespecified in B-spline approximation. In our simulation studies, we took different values for q_n and obtained similar results as shown in Tables 1 and 2.

3.2. *Comparison between the Group-Penalized Estimator and the Standard Partial-Likelihood Estimator*

In this subsection, we present some simulations to evaluate how well the proposed methods estimate the parametric component compared to the classical partial-likelihood method. The functional form of $g(\boldsymbol{x})$ in model (1) is set to be linear or partially linear with two covariates x_1 and x_2 for simplicity. We considered two scenarios for the true model: (i) $g(\boldsymbol{x}) = \beta_1 x_1 - 2x_2$ and (ii) $g(\boldsymbol{x}) = \beta_1 x_1 + f_3(x_2)$. To examine the performance of the proposed procedures in estimating parametric component $g_1(x_1) = \beta_1 x_1$, we used the estimated mean square error (MSE) over 1000 repetitions, i.e.,

$$\mathrm{MSE}(\widehat{g}_1) \approx \frac{1}{Mn}\sum_{k=1}^{M}\sum_{i=1}^{n}\left(\widehat{g}_1^{(k)}(x_{i1}) - \beta_1 x_{i1}\right)^2,$$

where $\widehat{g}_1^{(k)}$ is the penalized spline-based partial-likelihood estimate of g_1 with $\widehat{g}_1^{(k)}(x_{i1}) = \widehat{\beta}_1^{(k)}x_1 + \widehat{\phi}_1^{(k)}(x_{i1})$ based on the kth simulated data set, n is the sample size, and $M = 1000$ is the number of repetition. For comparison, we also computed $\widetilde{g}_1^{(k)}(x_{i1}) = \widetilde{\beta}_1^{(k)}x_{i1}$ with $\widetilde{\beta}_1^{(k)}$ being the partial-likelihood estimate for β_1 under the classical Cox model, and the estimated $\mathrm{MSE}(\widetilde{g}_1)$.

The estimated MSE results with $\beta_1 = 0.5$ are summarized in Table 3. It shows that the proposed group-penalized estimates are comparable with the standard partial-likelihood estimate when $g(\boldsymbol{x})$ is a linear function and outperforms the

Table 1

Simulation results for model structure estimation with sample size $n = 200$

$\lambda_0(t)$	CR	Algorithm	Method	Time	Iter	Corr%	AR$_+$	AR$_-$	ANN	MRME
1	20%	MBMD	GLASSO	0.3212	95.6430	0.972	0.018	0.010	3.008	0.7767
			GSCAD	0.2009	100.0000	0.983	0.017	0.017	3.017	0.3913
			GMCP	0.1724	100.0000	0.980	0.020	0.020	3.020	0.4701
		BMD	GLASSO	0.2429	99.9976	0.921	0.066	0.014	3.052	0.9512
			GSCAD	0.2392	100.0000	0.963	0.037	0.037	3.037	0.7593
			GMCP	0.2260	100.0000	0.963	0.037	0.037	3.035	0.7427
	40%	MBMD	GLASSO	0.2480	95.2522	0.688	0.005	0.309	2.696	0.9065
			GSCAD	0.1872	100.0000	0.973	0.027	0.027	2.979	0.5984
			GMCP	0.1675	100.0000	0.961	0.041	0.041	2.977	0.5854
		BMD	GLASSO	0.2061	99.9807	0.748	0.022	0.235	2.787	0.9752
			GSCAD	0.2053	99.9960	0.942	0.060	0.060	2.962	0.8389
			GMCP	0.1955	100.0000	0.949	0.052	0.052	2.964	0.8144
$2t$	20%	MBMD	GLASSO	0.3076	95.1210	0.958	0.021	0.021	3.000	0.0030
			GSCAD	0.1986	100.0000	0.981	0.019	0.019	3.019	0.0014
			GMCP	0.1696	100.0000	0.984	0.017	0.017	3.017	0.0016
		BMD	GLASSO	0.2365	99.9948	0.906	0.073	0.023	3.050	0.0047
			GSCAD	0.2335	100.0000	0.969	0.031	0.031	3.031	0.0028
			GMCP	0.2200	100.0000	0.969	0.031	0.031	3.031	0.0026
	40%	MBMD	GLASSO	0.2361	94.9567	0.623	0.005	0.373	2.632	0.0041
			GSCAD	0.1825	100.0000	0.956	0.044	0.044	2.962	0.0019
			GMCP	0.1634	100.0000	0.952	0.048	0.048	2.966	0.0020
		BMD	GLASSO	0.2004	99.9193	0.722	0.016	0.278	2.738	0.0051
			GSCAD	0.1989	100.0000	0.922	0.078	0.078	2.940	0.0032
			GMCP	0.1884	100.0000	0.929	0.071	0.071	2.947	0.0031

CR: the censoring rate;
GLASSO, GSCAD, GMCP: the penalized methods with the group LASSO, group SCAD, and group MCP penalties, respectively;
Time: the average running time (seconds) for computing the penalized estimator for a given λ;
Iter: the average number of iterations for the method achieve convergence;
Corr%: the average value of the ratio of exactly selecting the true structure of the model;
AR$_+$: the average value of R_+;
AR$_-$: the average value of R_-;
ANN: the average value of $|\widehat{\mathcal{A}}|$;
MRME: the median of the relative model error (RME).

standard partial-likelihood estimate when $g(\boldsymbol{x})$ is a partially linear function.

In addition, we computed the pointwise estimates of g_1 for $x_1 \in (0, 1)$ through the four methods (partial likelihood [PL], group-LASSO [GLASSO], group-SCAD [GSCAD], group-MCP [GMCP]) for each simulated data set. The mean of 1000 estimates, the 0.025 and the 0.975 quantiles of the 1000 estimates, and the 95% pointwise confidence intervals by the four approaches are displayed in Figures 1 and 2, which yield the same conclusion as those obtained from the comparison of their MSE estimates.

4. Application

We applied the proposed methods to analyze the PBC data from a study conducted by Mayo Clinic between 1974 and 1984, as mentioned in Section 1. In the study, there were 424 patients with PBC, a fatal chronic liver disease, and the 312 randomized participants were eligible for the analysis. The purpose of the study was to identify the risk factors related to the survival time of patients with PBC. During the study period, 125 patients of the 312 had died. Those patients who received transplantation were treated as censoring at the date of transplantation. The censoring rate is about 60%. Here, we study the dependence of the survival time on the following five covariates: presence of edema (x_1, coded as 1 for yes and 0 for no), age (x_2), prothrombin time (x_3), albumin (x_4) in gm/dl, and serum bilirubin (x_5) in mg/dl.

The PBC data have been analyzed by many authors in history (e.g., Fleming and Harrington, 1991; Grambsch, Therneau, and Fleming, 1995), and it is commonly known that the bilirubin predictor has a nonlinear effect on the log-risk function when albumin and prothrombin are taken as a logarithm scale in the Cox model. We used the proposed method to explore possible nonlinear effects among the five risk factors considered here. Let $\boldsymbol{X}_2 = (x_2, x_3, x_4, x_5)^{\mathrm{T}}$ and $\boldsymbol{X} = (x_1, \boldsymbol{X}_2^{\mathrm{T}})^{\mathrm{T}}$. Since x_1 is categorical, the survival time was assumed to follow model (1) with $g(\boldsymbol{X}) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\phi}(\boldsymbol{X}_2)$. Furthermore, we explored the same model with covariates x_1, x_2, $\log(x_3)$, $\log(x_4)$, and $\log(x_5)$.

To evaluate the goodness-of-fit of each method for the Cox model, we used a mean-square-type distance between the ob-

Table 2

Number of times of each component being selected as nonlinear effect over 1000 repetitions with sample size $n = 200$

$\lambda_0(t)$	CR	Algorithm	Method	g_1	g_2	g_3	g_4	g_5	g_6
1	20%	MBMD	GLASSO	10	5	3	990	1000	1000
			GSCAD	5	7	5	1000	1000	1000
			GMCP	8	6	6	1000	1000	1000
		BMD	GLASSO	21	26	19	986	1000	1000
			GSCAD	14	15	8	1000	1000	1000
			GMCP	14	15	7	999	1000	1000
	40%	MBMD	GLASSO	1	1	3	692	1000	999
			GSCAD	2	0	1	976	1000	1000
			GMCP	4	2	3	968	1000	1000
		BMD	GLASSO	7	11	4	768	998	999
			GSCAD	3	4	4	951	1000	1000
			GMCP	3	3	2	956	1000	1000
$2t$	20%	MBMD	GLASSO	9	10	2	979	1000	1000
			GSCAD	6	7	6	1000	1000	1000
			GMCP	8	4	5	1000	1000	1000
		BMD	GLASSO	24	29	20	977	1000	1000
			GSCAD	11	13	7	1000	1000	1000
			GMCP	11	14	6	1000	1000	1000
	40%	MBMD	GLASSO	2	1	2	628	999	1000
			GSCAD	2	0	1	959	1000	1000
			GMCP	2	4	1	959	1000	1000
		BMD	GLASSO	5	7	4	733	994	995
			GSCAD	2	4	3	931	1000	1000
			GMCP	2	5	2	938	1000	1000

CR: the censoring rate;
GLASSO, GSCAD, GMCP: the penalized methods with the group LASSO, group SCAD, and group MCP penalties, respectively.

Table 3

Estimation results of MSE for estimators of $g_1(x_1)$

CR	n	$\lambda_0(t)$	$g(\boldsymbol{x})$	PL	GLASSO	GSCAD	GMCP
20%	100	1	$0.5x_1 - 2x_2$	0.1515	0.1342	0.1349	0.1338
			$0.5x_1 + f_3(x_2)$	0.1495	0.1178	0.1297	0.1314
		$2t$	$0.5x_1 - 2x_2$	0.1575	0.1411	0.1415	0.1410
			$0.5x_1 + f_3(x_2)$	0.1481	0.1214	0.1294	0.1314
	200	1	$0.5x_1 - 2x_2$	0.0870	0.0929	0.0927	0.0892
			$0.5x_1 + f_3(x_2)$	0.0862	0.0562	0.0549	0.0541
		$2t$	$0.5x_1 - 2x_2$	0.0885	0.0928	0.0926	0.0902
			$0.5x_1 + f_3(x_2)$	0.0847	0.0565	0.0543	0.0545
40%	100	1	$0.5x_1 - 2x_2$	0.1945	0.1685	0.1696	0.1703
			$0.5x_1 + f_3(x_2)$	0.1907	0.1554	0.1644	0.1681
		$2t$	$0.5x_1 - 2x_2$	0.2026	0.1813	0.1820	0.1826
			$0.5x_1 + f_3(x_2)$	0.1989	0.1706	0.1800	0.1825
	200	1	$0.5x_1 - 2x_2$	0.1105	0.1028	0.1028	0.1024
			$0.5x_1 + f_3(x_2)$	0.1083	0.0721	0.0681	0.0660
		$2t$	$0.5x_1 - 2x_2$	0.1110	0.1002	0.1001	0.1001
			$0.5x_1 + f_3(x_2)$	0.1103	0.0766	0.0724	0.0702

CR: the censoring rate;
PL: the partial-likelihood method;
GLASSO, GSCAD, GMCP : the penalized methods with the group LASSO, group SCAD, and group MCP penalties, respectively.

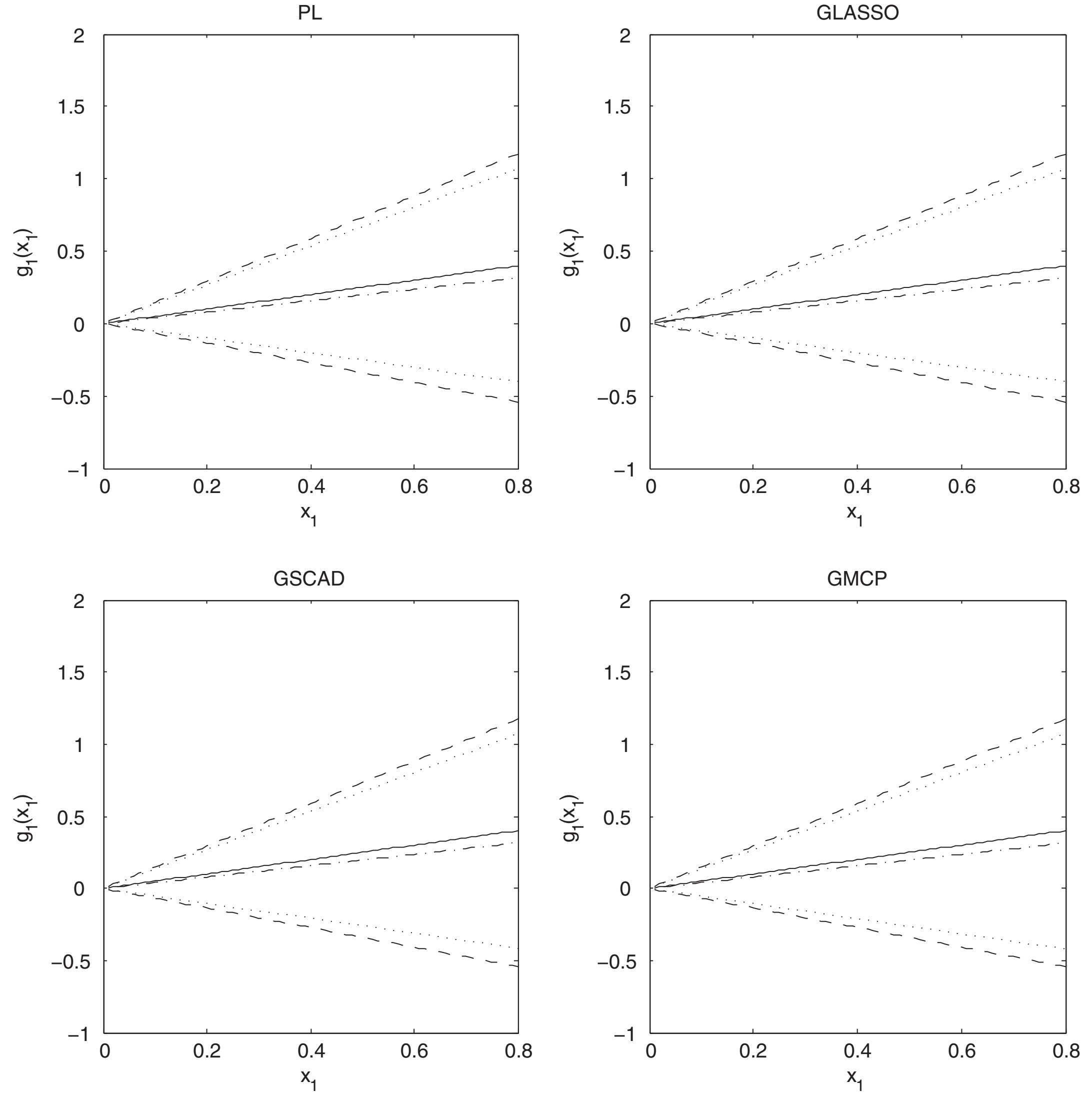

Figure 1. Estimates of $g_1(x_1) = 0.5x_1$ with $\lambda_0(t) = 2t$, $g(\boldsymbol{x}) = 0.5x_1 - 2x_2$, the sample size $n = 200$, and the censoring rate of 20%. The solid line is the true function of $g_1(x_1)$, the dot and dash line is the pointwise mean estimate, the dotted lines are the 0.025 and 0.975 quantiles of the pointwise estimates, and the dashed lines are the 95% pointwise confidence intervals.

served and expected numbers of events as follows:

$$\mathrm{D}^* = \sum_{l=1}^{n}\sum_{i=1}^{n} \Delta_l[\widehat{M}_i(T_l)]^2 / \sum_{l=1}^{n}\sum_{i=1}^{n} \Delta_l Y_i(T_l),$$

where $\widehat{\Lambda}_0(s)$ is the Breslow's estimator for the baseline cumulative hazard function and $\widehat{M}_i(t) = N_i(t) - \int_0^t Y_i(s)\exp\{\widehat{g}(\boldsymbol{X}_i)\}d\widehat{\Lambda}_0(s)$ is the estimated martingale residual which measures the distance between the observed and expected numbers of events for the i-th subject.

Using the PL method, we also fitted the data to the following two classical Cox models: Cox's regression model with the five covariates edema, age, prothrombin, albumin, and bilirubin; Cox's regression model with the five covariates edema, age, log(prothrombin), log(albumin), and log(bilirubin) (Fleming and Harrington, 1991).

Table 4 displays the covariates identified as having linear or nonlinear effects by the GLASSO, GSCAD, and GMCP methods, indicated by 0/1 (1, nonlinear; 0, linear). The first part in the table is based on the original five risk factors in the model, while the second part is based on the covariates x_3, x_4, and x_5 being log-transformed. From the first part, both group SCAD and group MCP methods detect that albumin (x_3), prothrombin time (x_4) in gm/dl, and serum bilirubin (x_5) have nonlinear effects. This result agrees with the analysis by Fleming and Harrington (1991) and Grambsch, Therneau, and Fleming (1995). However, the group LASSO method identifies age (x_2) and bilirubin (x_5) having nonlinear effects. From the second part, three penalized methods detect only $\log(x_5)$ having a nonlinear effect. Therefore, we can conclude that the three variables albumin, prothrombin, and bilirubin have nonlinear functional effects on the logarithm of hazard function, nonlinear effects of albumin, and prothrombin can be characterized

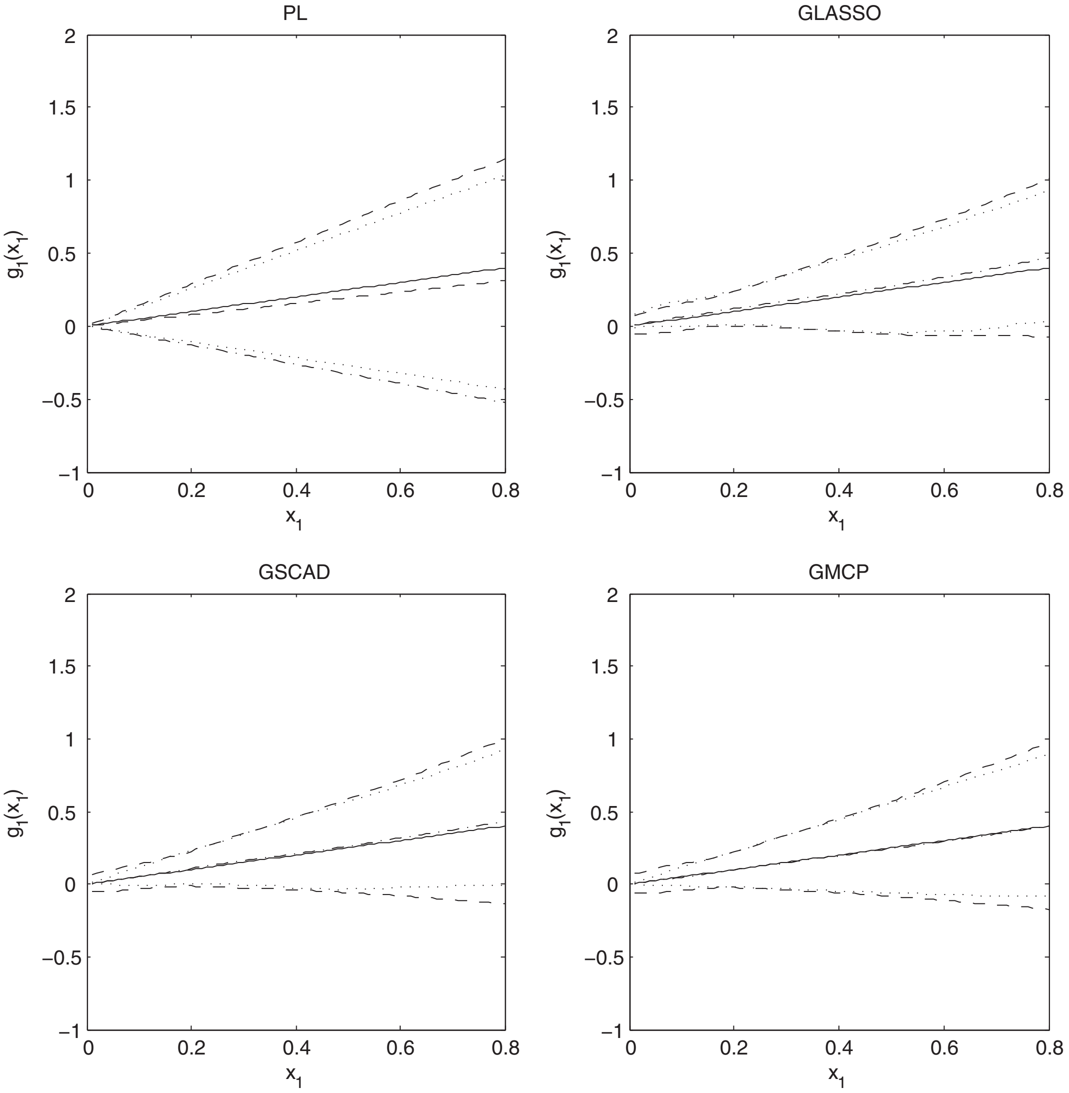

Figure 2. Estimates of $g_1(x_1) = 0.5x_1$ with $\lambda_0(t) = 2t$, $g(\boldsymbol{x}) = 0.5x_1 + f_3(x_2)$, the sample size $n = 200$, and the censoring rate of 20%. The solid line is the true function of $g_1(x_1)$, the dot and dash line is the pointwise mean estimate, the dotted lines are the 0.025 and 0.975 quantiles of the pointwise estimates, and the dashed lines are the 95% pointwise confidence intervals.

by a log transformation, but the nonlinear effect of bilirubin cannot be captured through a log transformation. Table 4 also shows the values of D^* for the eight fitted models. It can be seen from the table that the group-penalized methods yield smaller values of D^* than those from the fitted classical Cox models, and the fitted models by the group SCAD and the group MCP methods are the best in terms of D^*.

5. Concluding Remarks

In this article, we have studied the problem of which covariates have a linear effect and which have a nonlinear effect on the log-hazard rate function for the partially linear Cox model. This problem has been translated into a group selection problem through a semiparametric additive Cox model and B-spline approximation to each nonparametric component, and the penalized partial-likelihood approach has been applied to handle this group selection issue. The resulting estimators are consistent, the nonlinear effect can be successfully identified with high probability, and the estimated linear regression parameters are asymptotically normal.

To efficiently compute the penalized estimators, we have developed the modified blockwise majorization descent (MBMD) algorithm through the backtracking line search approach. The numerical studies have demonstrated that the proposed group-penalized approaches and MBMD algorithm work well.

For simplicity, we have assumed the covariates are time-independent in the Cox model. The proposed method can be easily extended to the case of time-dependent covariates.

Note that we have only considered identifying which covariates have linear or nonlinear effects. For a further study, one can combine the variable selection problem with the identification of linear or nonlinear covariate effects together. On the

Table 4
Analysis results for PBC data

Method	Edema	Age	Albumin	Prothrombin	Bilirubin	D^*
GLASSO	0	1	0	0	1	0.4553
GSCAD	0	0	1	1	1	0.4261
GMCP	0	0	1	1	1	0.4261
COX						0.4605
	Edema	Age	log(Albumin)	log(Prothrombin)	log(Bilirubin)	D^*
GLASSO	0	0	0	0	1	0.3753
GSCAD	0	0	0	0	1	0.3808
GMCP	0	0	0	0	1	0.3808
COX						0.4597

GLASSO, GSCAD, GMCP: the penalized methods with the group LASSO, group SCAD, and group MCP penalties, respectively;
COX: the Cox's regression method;
0/1: the covariate effect on the logarithm of hazard function is identified to be linear/nonlinear by the group-penalized methods.

other hand, it is expected to extend our proposed approach to high-dimensional settings. Also, we can consider the same problem for other useful models such as accelerate failure time models and additive hazards models.

6. Supplementary Materials

The Web Appendices referenced in Section 2 are available with this article at the *Biometrics* website on Wiley Online Library.

Acknowledgements

The authors would like to thank the Editor, Professor Yi-Hau Chen, the Associate Editor, and the three referees for their constructive comments and suggestions that improved the article substantially. Huang's research is partly supported by the Program for Changjiang Scholars and Innovative Research Team in SUFE (IRT13077). Zhao's research is partly supported by the Research Grant Council of Hong Kong (503513), the National Natural Science Foundation of China (No. 11371299), and The Hong Kong Polytechnic University.

References

Böhning, D. and Lindsay, B. G. (1988). Monotonicity of quadratic-approximation algorithms. *Annals of the Institute of Statistical Mathematics* **40**, 641–663.

Cai, J., Fan, J., Jiang, J., and Zhou, H. (2007). Partially linear hazard regression for multivariate survival data. *Journal of the American Statistical Association* **102**, 538–551.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B,* **74**, 187–220.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerical Mathematics* **31**, 377–403.

Du, P., Ma, S., and Liang, H. (2010). Penalized variable selection procedure for Cox models with semiparametric relative risk. *The Annals of Statistics* **38**, 2092–2117.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics* **30**, 74–99.

Fan, J. and Lv, J. (2011). Non-concave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory* **57**, 5467–5484.

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

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics* **7**, 397–416.

Grambsch, P. M., Therneau, T. M., and Fleming, T. R. (1995). Diagnostic plots to reveal functional form for covariates in multiplicative intensity models. *Biometrics* **51**, 1469–1482.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science* **1**, 297–318.

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

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

Meier, L., van de Geer, S., and Bühlmann, P. (2008). The group LASSO for logistic regression. *Journal of the Royal Statistical Society, Series B* **70**, 53–71.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

Tibshirani, R. (1997). The Lasso method for variable selection in the Cox model. *Statistics in Medicine* **16**, 385–395.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., et al. (2012). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society, Series B* **74**, 245–266.

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

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* **68**, 49–67.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**, 894–942.

Zhang, H. H., Cheng, G., and Liu, Y. (2011). Linear or nonlinear? Automatic structure discovery for partially linear models. *Journal of the American Statistical Association* **106**, 1099–1112.

Received July 2015. *Revised March* 2016. *Accepted March* 2016.