Electronic Journal of Statistics
Vol. 15 (2021) 748–772
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1799

Bi-selection in the high-dimensional additive hazards regression model

Li Liu*

Wuhan University
School of Mathematics and Statistics
Wuhan, Hubei, China
e-mail: lliu.math@whu.edu.cn

Wen Su

The University of Hong Kong
Department of Statistics and Actuarial Science
Hong Kong
e-mail: jenna.wen.su@connect.hku.hk

Xingqiu Zhao†

The Hong Kong Polytechnic University
Department of Applied Mathematics
Hong Kong
e-mail: xingqiu.zhao@polyu.edu.hk

Abstract: In this article, we consider a class of regularized regression under the additive hazards model with censored survival data and propose a novel approach to achieve simultaneous group selection, variable selection, and parameter estimation for high-dimensional censored data, by combining the composite penalty and the pseudoscore. We develop a local coordinate descent (LCD) algorithm for efficient computation and subsequently establish the theoretical properties for the proposed selection methods. As a result, the selectors possess both group selection oracle property and variable selection oracle property, and thus enable us to simultaneously identify important groups and important variables within selected groups with high probability. Simulation studies demonstrate that the proposed method and LCD algorithm perform well. A real data example is provided for illustration.

MSC2020 subject classifications: Primary 62N01, 62N02; secondary 62F12.

Keywords and phrases: Additive hazards model, high dimension, composite penalty, local coordinate descent algorithm, oracle property.

Received July 2019.

*The work of Li Liu was supported in part by the Natural Science Foundation of China (No. 11971362).

†The work of Xingqiu Zhao was supported in part by the Natural Science Foundation of China (No. 11771366), the Research Grant Council of Hong Kong (15301218, 15303319), and The Hong Kong Polytechnic University.

Contents

1. Introduction

Recent advancements in experimental technology have enabled us to achieve numerous destinations in clinical studies that seemed impossible before. Most notably, availability of high-dimensional medical data, where the number of variables p is substantially greater than sample size n, introduces new opportunities in modern healthcare research but its high dimensionality characteristic poses tremendous challenges to classical statistical analysis methodologies. Luckily, the true regression model generally possesses the sparsity property, which means only a small number of nonzero components are attributable. Taking gene expression data as an example, which often involves tens of thousands of potential covariates, however, merely a handful of them are indeed related to the development of a particular disease. Variable selection techniques are effective tools in reducing dimensionality, cherry-picking the few covariates with significant contribution to the outcome at a certain threshold, resulting a simpler model for interpretation and more efficient parameter estimates. One particular type of variable selection methods frequently adopted to handle high-dimensional data is the regularization approach, achieving dimension reduction by adding a penalty function to the loss function. Moreover, the advantage associated with these methods includes its capability to simultaneously identify important variables and provide parameter estimates. Commonly used penalties include the least absolute shrinkage and selection operator (LASSO) [22], the smoothly clipped absolute deviation (SCAD) penalty [7], the adaptive LASSO [32], and the minimum concave penalty (MCP) [28], among others.

Furthermore, variable selection becomes especially challenging when the dataset exhibits group structure. Some examples include: multilevel categorical covariates in a regression model expressed by a group of dummy variables; a continuous covariate represented by a set of basis functions; genetic markers from the same gene considered as a group in genetic association studies; and in gene expression analysis, genes with the same biological pathway forming a natural group. Among others, [26, 13, 27, 21] considered penalty-based group selection methods. [12] implemented a group bridge penalty to achieve bi-selection, simultaneously selecting important groups and important variables within selected groups. To complement this methodology, [3] subsequently proposed the local coordinate descent (LCD) algorithm to calculate the bi-level selection estimates in generalized linear models.

While most survival data involves censorship casting additional complexity to data structure and difficulty in regression modeling, there has been a large class of literature proposing various approaches to specifically address variable selection at the individual and group level based on the Cox models. For instance, [23, 29, 8] extended the LASSO, the adaptive LASSO, and nonconcave penalized likelihood approach to the Cox model for picking statistically significant individual variables. Building on that, [19] applied the supervised group LASSO penalization to the Cox model to select important variable groups. Subsequently, [11] further discussed the capability of group bridge approach to simultaneously selecting important variables at both the individual and group levels in the framework of the Cox model. Furthermore, [4] studied the issue of identifying regression structure under the Cox model by a penalized group selection method with concave penalties.

Alternative to the popular Cox model, [5, 15] considered an additive hazards model, which relaxes the proportional hazard assumption by regressing the risk difference. Variable selection in the additive hazards model has drawn much attention recently. Under a fixed dimensional setting, [14] introduced a weighted LASSO approach; [20] discussed several regularization schemes including the LASSO, adaptive LASSO and Dantzig selector. On the contrary, under a high-dimensional setting, [31] developed tests for coefficients; [30] studied the properties of the weighted LASSO; [16, 25] explored implications of implementing regularized least squares and penalized empirical likelihood for sparse models, respectively. Under the framework of additive hazards model for high-dimensional data, we propose a novel approach that caputures group structure while retaining sparsity of covariates, such that it simultaneously selects important variables at the individual and group levels, at the same time providing parameter estimates. This is achieved by combining a composite penalty and the pseudoscore method, where the number of covariates p is allowed to grow nonpolynomially with a sample size n. The asymptotic properties of the proposed estimators include both group selection oracle property and variable selection oracle property, which means important groups and important variables within selected groups are consistently identified, and the resulting estimators are asymptotically normal under some regularity conditions. Furthermore, we incorporate the local coordinate descent algorithm first proposed by [3], and

demonstrate its effectiveness through simulation studies and real data analysis.

The remainder of the paper is organized as follows. In Section 2, we describe the penalized pseudoscore inference procedure, explore suitable penalty functions, and introduce the local coordinate descent algorithm. Theoretical properties of the estimators are studied in Section 3. We conduct simulation studies to evaluate the performance of the proposed method in Section 4, and in Section 5, we show an application of the proposed method to the breast cancer dataset. We draw some concluding remarks in Section 6 and relegate proofs of the key results to the Appendix.

2. Estimation procedure

2.1. Model setting and penalized procedure

Suppose that the failure time T^U satisfies the following additive hazards model:

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t) + \boldsymbol{\beta}_0^T \boldsymbol{Z}(t), \tag{2.1}$$

where $\lambda_0(t)$ is the unspecified baseline function, $\boldsymbol{Z}(t) = (Z_1(t), \ldots, Z_p(t))^T$ is a p-dimensional vector of covariates which is split into K groups, and $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})^T$ is the true value of regression coefficient of covariate $\boldsymbol{Z}(t)$. Let $A_k = \{k_1, \ldots, k_{J_k}\}$ be the subset of $\{1, \ldots, p\}$ representing the k-th known group, and $(\beta_{k_1}, \ldots, \beta_{k_{J_k}})^T$ be a J_k-dimensional vector of regression coefficients in the kth group. Let C be a censoring time, $T = T^U \wedge C$ be the observed survival 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 subject $i = 1, 2, \ldots, n$.

Define the observed failure counting process as $N_i(t) = I(T_i \leq t, \Delta_i = 1)$ and the at-risk indicator $Y_i(t) = I(T_i \geq t)$. Following Lin and Ying (1994), the regression coefficients can be estimated by solving the following pseudoscore estimating equation

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

where $\overline{\boldsymbol{Z}}(t) = \sum_{i=1}^{n} Y_i(t) \boldsymbol{Z}_i(t) / \sum_{i=1}^{n} Y_i(t)$, and τ is the maximum follow-up time. This equation can be rewritten as

$$\boldsymbol{b} - \boldsymbol{V}\boldsymbol{\beta} = 0,$$

where

$$\boldsymbol{b} = \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau (\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}(t)) dN_i(t),$$

$$\boldsymbol{V} = \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau Y_i(t) (\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}(t))^{\otimes 2} dt$$

with $\boldsymbol{v}^{\otimes 2}$ meaning $\boldsymbol{v}\boldsymbol{v}^T$ for a vector $\boldsymbol{v}$. While $\boldsymbol{V}$ is positive semidefinite, integrating $-U(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ produces the least-squares-type loss function below

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

In order to simultaneously select important groups and individual variables, we propose to obtain the estimator $\hat{\boldsymbol{\beta}}$ by minimizing the following objective function with composite penalty

$$Q(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) + \lambda_n \sum_{k=1}^{K} f_O^{(k)}\Big(\sum_{j=1}^{J_k} f_I^{(j)}(|\beta_{k_j}|;\lambda_n);\lambda_n\Big), \tag{2.2}$$

where the functions $f_O^{(k)}$ and $f_I^{(j)}$ are penalty functions, and λ_n is a tuning parameter. Variables can enter the model either by having a strong individual signal or by being a member of a group with strong collective signal. Intuitively, the outer penalty function $f_O^{(k)}$ shrinks the kth unimportant group effect to zero, and the inner penalty function $f_I^{(j)}$ excludes the unimportant variable effect within the groups at the same time. Hence the composite penalty function could achieve the bi-selective goal, simultaneously selecting important groups and important variables within the selected groups.

For simplicity, we omit the dependence of the penalty functions $f_O^{(k)}$ and $f_I^{(j)}$ on the tuning parameter λ_n, and assume that they are independent of k and j, which means that we can apply the same penalty functions across different variables and different groups. Subsequently, the subscript n in λ_n is also omitted and we rewrite the penalty part as

$$\sum_{k=1}^{K} f_O\Big(\sum_{j=1}^{J_k} f_I(|\beta_{k_j}|)\Big) = \rho_\lambda(|\boldsymbol{\beta}|),$$

where $|\boldsymbol{\beta}| = (|\beta_1|,\ldots,|\beta_p|)^T$.

2.2. *Penalty functions*

[18] studied a variety of penalty functions under the framework of generalized linear models. To determine the functional form of f_O and f_I, we mainly consider the following types of penalty $f_\lambda(\cdot)$.

(i) The smoothly clipped absolute deviation (SCAD) penalty [6, 7] given by the derivative

$$f'_\lambda(\theta) = I(\theta \le \lambda) + \frac{(a\lambda-\theta)_+}{(a-1)\lambda} I(\theta > \lambda), \quad \theta \ge 0,$$

where $a > 2$ is a shape parameter.

(ii) The minimax concave penalty (MCP) proposed by [28] with the derivative

$$f'_\lambda(\theta) = \frac{(a\lambda - \theta)_+}{a\lambda}, \quad \theta \geq 0,$$

where $a > 1$ is a shape parameter.

(iii) The smooth integration of counting and absolute deviation (SICA) penalty [18] with

$$f_\lambda(\theta) = \frac{(a+1)\theta}{a+\theta}, \quad \theta \geq 0,$$

where $a > 0$ is a shape parameter.

[3] set both f_O and f_I as the MCP penalty, and suggested the shape parameters in f_I and f_O to be $a = 3$ and $J_k a\lambda/2$, respectively. The authors named this penalty the "composite MCP (CMCP) penalty". In this paper, we construct several new composite penalties, including composite SCAD (CSCAD) and composite SICA (CSICA) with both f_O and f_I as SCAD penalty and SICA penalty, respectively. Additionally, we construct another composite MSICA penalty by taking f_O as the MCP penalty and f_I as the SICA penalty. In subsequent simulation studies and real data analysis, we compare the performance of these composite penalties paired with three variable selectors, MCP, SCAD and SICA.

2.3. Local coordinate descent algorithm

[22] proposed to accomplish parameter shrinkage and selection for a linear regression model by minimizing the following squared loss function with LASSO penalty

$$\frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

where $\boldsymbol{X} = (X_1, \cdots, X_p)^T$, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$, and $\|\boldsymbol{v}\|_2$ represents the L_2-norm for a vector $\boldsymbol{v}$. At the jth iteration step in the coordinate decent algorithm, the solution of β_j can be updated as

$$\tilde{\beta}_j = \frac{S(\frac{1}{n}X_j^T\boldsymbol{r} + \frac{1}{n}X_j^T X_j\beta_j, \lambda)}{\frac{1}{n}X_j^T X_j},$$

where $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$ and

$$S(z, c) = \begin{cases} z - c, & \text{if } z > c, \\ 0, & \text{if } |z| \leq c, \\ z + c, & \text{if } z < -c. \end{cases}$$

Motivated by the above results, [3] developed a fast and stable local coordinate decent (LCD) algorithm for bi-level variable selectors that approximates the

penalty proportional to $\tilde{\lambda}_{k_j}|\beta_{k_j}|$ by taking its first order Taylor series about β_{k_j}, where

$$\tilde{\lambda}_{k_j} = \lambda f_O'\Big(\sum_{j=1}^{J_k} f_I(|\beta_{k_j}|)\Big) f_I'(|\beta_{k_j}|) \tag{2.3}$$

for each $k_j \in A_k$. Subsequently, the coefficient β_{kj} is updated as

$$\tilde{\beta}_{k_j} = \frac{S(\frac{1}{n}X_{k_j}^T \boldsymbol{r} + \frac{1}{n}X_{k_j}^T X_{k_j}\beta_{k_j}, \tilde{\lambda}_{k_j})}{\frac{1}{n}X_{k_j}^T X_{k_j}}, \tag{2.4}$$

where $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$. To minimize $Q(\boldsymbol{\beta})$ in (2.2), we apply the LCD algorithm to calculate the estimates through the following steps:

Step 1. Choose an initial estimate $\boldsymbol{\beta}^{(0)}$;
Step 2. Let $\boldsymbol{X} = \boldsymbol{V}^{1/2}$ and $\boldsymbol{y} = \boldsymbol{X}^{-1}\boldsymbol{b}$. Update $\tilde{\lambda}_{k_j}$ and $\tilde{\beta}_{k_j}$ cyclically according to (2.3) and (2.4) for each $k_j \in A_k$, $k = 1, \cdots, K$;
Step 3. Repeat Step 2 until convergence.

The choice of the initial estimate is critical to the proposed algorithm. Under the high-dimensional case, the ridge solution of the least-squares-type loss function $L(\boldsymbol{\beta})$ is an ideal choice of $\boldsymbol{\beta}^{(0)}$ since it is easy to calculate and it could be close enough to the true parameter for some suitable tuning parameter. In our simulations, we set $\boldsymbol{\beta}^{(0)} = (\boldsymbol{V} + \lambda^* I)\boldsymbol{b}$ with the tuning parameter $\lambda^* = 0.1$, where I is the identity matrix.

The local coordinate descent (LCD) algorithm adopted here can be considered as an application of the algorithm developed in [17], and thus achieves optimal statistical rates.

3. Asymptotic results

According to the sparsity of the parameter, we split the true parameter as $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0\mathcal{B}}^T, \boldsymbol{\beta}_{0\mathcal{C}}^T)^T$, where $\mathcal{B} = \{j|\beta_{0j} = 0, \boldsymbol{\beta}_{0A_k} \neq \boldsymbol{0}, j \in A_k \text{ for some } k = 1, \ldots, K\}$ and $\mathcal{C} = \{j|\boldsymbol{\beta}_{0A_k} = \boldsymbol{0}, j \in A_k, k = 1, \ldots, K\}$. Let $\mathcal{A} = \{j|\beta_{0j} \neq 0\}$, $\mathcal{D} = \mathcal{A}^c$, $s = |\mathcal{A}|$, with $|\mathcal{A}|$ denoting the cardinality of $\mathcal{A}$ and s as the number of important variables. Note that both the number of covariates p and the number of important variables s are allowed to depend on sample size n throughout the paper, thus we omit the subscript n to simplify notations. We use $\Lambda_{\min}(\cdot)$ to denote the minimum eigenvalue of a matrix, and use the subscript $\mathcal{A}$ for a vector or a matrix to denote the sub-vector or sub-matrix containing them. For example, $\boldsymbol{x}_{\mathcal{A}}$ means the $|\mathcal{A}|$-dimensional vector consisting of components $\{x_j, j \in \mathcal{A}\}$ for the vector $\boldsymbol{x}$, and $\boldsymbol{V}_{\mathcal{A}\mathcal{A}}$ means the $|\mathcal{A}|$-dimensional squared matrix with entries $v_{ij}, i \in \mathcal{A}, j \in \mathcal{A}$ for the matrix $\boldsymbol{V} = (v_{ij})$. In addition, we denote that $|\beta_j|_{j\in\mathcal{A}} = (|\beta_j| : j \in \mathcal{A})^T$.

We first state the following lemma, which plays an important role in establishing the selection consistency of the estimators.

Lemma 3.1. *$\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ is a strict local minimizer of $Q(\boldsymbol{\beta})$ if the following conditions hold*

$$\boldsymbol{U}_{\hat{\mathcal{A}}}(\hat{\boldsymbol{\beta}}) - \lambda \frac{\partial \rho_\lambda(|\hat{\boldsymbol{\beta}}|)}{\partial |\beta_j|_{j\in\hat{\mathcal{A}}}} \circ sgn(\hat{\boldsymbol{\beta}}_{\hat{\mathcal{A}}}) = 0, \tag{3.1}$$

$$\|\boldsymbol{U}_{\hat{\mathcal{D}}}(\hat{\boldsymbol{\beta}})\|_\infty \le \lambda \min_{j\in\hat{\mathcal{D}}} \frac{\partial \rho_\lambda(|\hat{\boldsymbol{\beta}}|)}{\partial |\beta_j|}, \tag{3.2}$$

$$\Lambda_{\min}(\boldsymbol{V}_{\hat{\mathcal{A}}\hat{\mathcal{A}}}) > \lambda \kappa(\rho_\lambda, \hat{\boldsymbol{\beta}}_{\hat{\mathcal{A}}}), \tag{3.3}$$

where $\circ$ is the Hadamard (entrymise) product, $\|\boldsymbol{v}\|_\infty$ represents the L_∞-norm of the vector $\boldsymbol{v}$, and

$$\kappa(\rho_\lambda, \boldsymbol{\beta}_{\mathcal{A}}) = \max_{j\in\mathcal{A}} \left\{ \frac{\partial^2 \rho_\lambda(|\boldsymbol{\beta}|)}{\partial |\beta_j|^2} \right\}.$$

(3.1) and (3.3) in Lemma 3.1 imply that $\hat{\boldsymbol{\beta}}$ is a strict minimizer of $Q(\boldsymbol{\beta})$ in the subspace $\mathbb{B} = \{\boldsymbol{\beta} \in \mathbb{R}^p | \hat{\boldsymbol{\beta}}_{\hat{\mathcal{D}}} = 0\}$. Condition (3.2) ensures that $Q(\boldsymbol{\beta}_1) \ge Q(\boldsymbol{\beta}_2)$ for any $\boldsymbol{\beta}_1 \in \mathbb{R}^p / \mathbb{B}$ in a sufficiently small neighborhood of $\hat{\boldsymbol{\beta}}$, and for any $\boldsymbol{\beta}_2$ which is the projection of $\boldsymbol{\beta}_1$ onto the subspace $\mathbb{B}$. Thus, $\hat{\boldsymbol{\beta}}$ satisfying the conditions in Lemma 3.1 is indeed a strict local minimizer of $Q(\boldsymbol{\beta})$ on the whole space $\mathbb{R}^p$.

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

$$\begin{aligned}
&\boldsymbol{s}^{(k)}(t) = \mathrm{E}[Y(t)\boldsymbol{Z}(t)^{\otimes k}],\\
&\boldsymbol{e}(t) = \boldsymbol{s}^{(1)}(t)/\boldsymbol{s}^{(0)}(t),\\
&\boldsymbol{D} = \mathrm{E}\Big[\int_0^\tau Y(t)\{\boldsymbol{Z}(t) - \boldsymbol{e}(t)\}^{\otimes 2} dt\Big],\\
&\boldsymbol{\Sigma} = \mathrm{E}\Big[\int_0^\tau \{\boldsymbol{Z}(t) - \boldsymbol{e}(t)\}^{\otimes 2} dN(t)\Big]
\end{aligned}$$

where $\boldsymbol{v}^{\otimes 0} = 1$, $\boldsymbol{v}^{\otimes 1} = \boldsymbol{v}$ and $\boldsymbol{v}^{\otimes 2} = \boldsymbol{v}\boldsymbol{v}^T$ for a vector $\boldsymbol{v}$. Let $\phi = \|\boldsymbol{D}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty$, $d = \frac{1}{2}\min_{j\in\mathcal{A}} |\beta_{0j}|$, $\kappa_0 = \sup\{\kappa(\rho, \boldsymbol{\theta}) : \|\boldsymbol{\theta} - \boldsymbol{\beta}_{0\mathcal{A}}\|_\infty \le d\}$, and $\mu = \Lambda_{\min}(\boldsymbol{D}_{\mathcal{A}\mathcal{A}}) - \lambda\kappa_0$. Assume that $c_n^* = \max_j \sum_{k: A_k \ni j} I(j \in \mathcal{A})$ is bounded by a constant c_1. We suppose that $f_I(0) = 0$ and write $\rho'_\lambda(\mathbf{0}+) = f'_O(0+)f'_I(0+)$ and $\rho'_\lambda(\boldsymbol{d}) = f'_O(f_I(d))f'_I(d)$.

To establish the asymptotic properties of the proposed estimators, we need the following regularity conditions.

Condition 1. The function $\rho_\lambda(\boldsymbol{\theta})$ is increasing and concave on each component θ_j of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in [0, \infty]^p$, and has continuous partial derivatives $\partial\rho_\lambda(\boldsymbol{\theta})/\partial\theta_j$ on $\theta_j \in (0, \infty)$ for $j = 1, \ldots, p$. In addition, $\rho'_\lambda(\boldsymbol{\theta})$ is increasing on tuning parameter λ, and $\rho'_\lambda(\mathbf{0}+) = \rho'(\mathbf{0}+) > 0$ independent of λ.

Condition 2. (i) $\int_0^\tau \lambda_0(t)dt < \infty$; (ii) $P(Y(\tau) = 1) > 0$; (iii) There exist constants $M, K, r > 0$ such that

$$P\left(\sup_{t\in[0,\tau]} |Z_j(t)| > x\right) \le M \exp(-Kx^r)$$

for all $x > 0$ and $j = 1, \ldots, p$; (iv) The sample paths of $Z_j(\cdot), j = 1, \ldots, p$ are of uniformly bounded variation.

Condition 3. There exist constants $\alpha \in (0, 1]$, $\gamma \in [0, 1/2]$, and $c > 0$ such that

$$\|\boldsymbol{D}_{\mathcal{D}\mathcal{A}}\boldsymbol{D}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \leq \left\{(1-\alpha)\frac{\rho'(\mathbf{0}+)}{c_n^*\rho'_\lambda(\boldsymbol{d})}\right\} \wedge (cn^\gamma).$$

Condition 1 is a very mild requirement that can be easily met with some commonly used penalties. Indeed, many penalties, such as the LASSO penalty, SCAD penalty, MCP penalty, and SICA penalty, satisfy the relation $f_I(0) = 0$ if we take them as the function f_I. Moreover, it is easy to see that the composite penalties listed in Section 2.2 satisfy Condition 1. Conditions 2 and 3 are similar to those in [16], where Condition 2 is a commonly used condition for survival models and Condition 3 is the key condition for verifying the selection consistency, such that the empirical counterparts of the matrices, such as $\boldsymbol{D}_{\mathcal{D}\mathcal{A}}\boldsymbol{D}_{\mathcal{A}\mathcal{A}}^{-1}$, $\boldsymbol{D}_{\mathcal{A}\mathcal{A}}^{-1}$ and $\boldsymbol{D}_{\mathcal{A}\mathcal{A}}$, are close to them in some sense.

We now present the conclusions regarding the group selection consistency of the proposed estimators.

Theorem 3.2 (Consistency of group selection)**.** *Suppose that Conditions 1–3 hold. Also assume that*

$$\begin{aligned} &\frac{n(\rho'_\lambda(\boldsymbol{d})^{-1} \wedge n^\gamma)^2}{\phi^2 s^2 (\log p)^{r_1}} \to \infty, \quad \frac{n(\phi^{-1} \wedge \mu)^2}{s^2(\log s)^{r_1}} \to \infty, \\ &\frac{n\lambda^2}{(\log p)^{r_1}} \to \infty, \quad \frac{n^{1-2\gamma}\lambda^2}{(\log s)^{r_1}} \to \infty, \quad d \geq c_1\phi\lambda\rho'(\mathbf{0}+), \end{aligned} \tag{3.4}$$

where $\mu > 0$, $r_1 = (r+4)/r$, and $c_1 = 2 + 1/(4c)$. Then for some constant $M, K > 0$, with probability at least

$$\begin{aligned} 1 - M\exp\Big[-Kn^{1/r_1}\Big\{\frac{(\phi^{-1} \wedge \mu)^2}{s^2} \wedge 1\Big\}^{1/r_1}\Big] \\ - M\exp\Big[-Kn^{1/r_1}\Big\{\frac{\lambda^2}{n^{2\gamma}} \wedge 1\Big\}^{1/r_1}\Big] \to 1, \end{aligned}$$

we have

(a) (Sparisty) $\hat{\boldsymbol{\beta}}_{\mathcal{C}} = 0$;

(b) (L_∞-loss) $\|\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}\|_\infty \leq c_1\phi\lambda\rho'(\mathbf{0}+)$.

Part (a) in Theorem 3.2 shows that the unimportant groups can be excluded with high probability; part (b) provides the convergence rate of the estimated regression coefficients of important variables in L_∞-norm.

To explain the intuition for the conditions in Theorem 3.2, we consider some simplified cases. For the concave and composite penalties listed in Section 2.2, we have $\rho'_\lambda(\boldsymbol{d}) \leq \rho'(\mathbf{0}+)$. Thus, the first two conditions in (3.4) are satisfied if

$$\frac{n}{\phi^2 s^2 (\log p)^{r_1}} \to \infty \tag{3.5}$$

when $\Lambda_{\min}(\boldsymbol{D}_{\mathcal{A}\mathcal{A}})$ is bounded away from zero. In particular, if ϕ is a constant, then (3.5) is ensured by the condition that $n \gg s^2(\log p)^{r_1}$. This means that the dimension of the covariates is allowed to increase nonpolynomially with the sample size as large as $\log p = o(n^{1/r_1})$, where the dimension of the true sparse model $s = o(n^{1/2})$. Furthermore, for the bounded covariates, the third and the forth conditions in (3.4) reflect the requirement for the order of the regularization parameter λ as

$$\lambda \gg \sqrt{\frac{\log p}{n}} \vee \sqrt{\frac{\log s}{n^{1-2\gamma}}}.$$

The last inequality in (3.4) implies that the minimum signal d must satisfy

$$d \gg \phi\left(\sqrt{\frac{\log p}{n}} \vee \sqrt{\frac{\log s}{n^{1-2\gamma}}}\right).$$

The conditions in Theorem 3.2 are different from those in the existing literature. For example, [9, 2] demanded that $s = O(n^{\alpha})$ and $\log p = O(n^{\delta})$ with $\alpha, \delta \in (0, 1)$. As pointed out by [16], besides the difference in model assumptions, the critical difference is that they imposed a condition on a large empirical covariance matrix [see e.g., Condition 2 in [9] and Condition 8 in [2]]. As the empirical covariance matrix involves the outcome variables in survival models, the more nature idea is to provide a nonrandom condition on the population covariance matrix, as shown in Condition 3 of this paper. This population assumption can be viewed as high-dimensional extensions of the classical asymptotic regularity conditions in the low-dimensional setting.

To state the asymptotic normality of $\hat{\boldsymbol{\beta}}_{\mathcal{A}}$, we define $\Lambda_1 = \Lambda_{\min}(\boldsymbol{D}_{\mathcal{A}\mathcal{A}})$, $\Lambda_2 = \Lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}})$, and $\Lambda_3 = \Lambda_{\min}(\boldsymbol{D}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}\boldsymbol{D}_{\mathcal{A}\mathcal{A}}^{-1})$.

First, we present the oracle properties of the proposed estimator in the ideal case where the variables within each group are either all important or all unimportant.

Theorem 3.3. *Suppose that the conditions of Theorem 3.2 hold. Also assume that*

$$\frac{n\Lambda_1^2}{s^2(\log s)^{r_1}} \to \infty, \frac{n\Lambda_2^2}{s^2} \to \infty, \frac{n\Lambda_1^4\Lambda_3}{s^3} \to \infty, \frac{ns\lambda^2}{\Lambda_1^2\Lambda_3}\rho'_\lambda(\boldsymbol{d}) \to 0, \tag{3.6}$$

where $r_1 = (r+4)/r$, *and*

$$\rho'(\mathbf{0}+) \le \min_{j\in\mathcal{D}} \frac{\partial \rho_\lambda(|\hat{\boldsymbol{\beta}}|)}{\partial|\beta_j|}. \tag{3.7}$$

Then for some constant $M, K > 0$, *with probability at least*

$$1 - M\exp\Big[-Kn^{1/r_1}\Big(\frac{(\phi^{-1}\wedge\mu\wedge\Lambda_1)^2}{s^2}\wedge 1\Big)^{1/r_1}\Big] - M\exp\Big[-Kn^{1/r_1}\Big(\frac{\lambda^2}{n^{2\gamma}}\wedge 1\Big)^{1/r_1}\Big] \to 1,$$

we have

(a) (Sparsity) $\hat{\boldsymbol{\beta}}_{\mathcal{D}} = 0$*;*

(b) (Asymptotic normality) For every $\boldsymbol{u} \in \mathbb{R}^s$ *with* $\|\boldsymbol{u}\|_2 = 1$, $\sqrt{n}\boldsymbol{u}^T \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1/2} \boldsymbol{D}_{\mathcal{A}\mathcal{A}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}})$ *is asymptotically distributed as standard normal.*

The first three conditions in (3.6) require the true model dimension s bounded by both sample size n and eigenvalues of the matrices $\boldsymbol{D}_{\mathcal{A}\mathcal{A}}$, $\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}$, and $\boldsymbol{D}_{\mathcal{A}\mathcal{A}}^{-1}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}\boldsymbol{D}_{\mathcal{A}\mathcal{A}}^{-1}$. A special case of these conditions is $s = o(n^{1/3})$ when the eigenvalues of these matrices are bounded away from zero. The last condition in (3.6) is satisfied for all penalties listed in Section 2.2. It is also worth to mention that the condition in (3.7) holds if the variables within each group are either all important or all unimportant, which may not be true in practice.

To remove such strong condition and identify the important variables within selected groups, we propose to estimate $\boldsymbol{\beta}$ by minimizing $Q^{\omega}(\boldsymbol{\beta})$ defined as

$$Q^{\omega}(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) + \lambda\rho_{\lambda}(|\boldsymbol{\beta}^{\omega}|),$$

where $\boldsymbol{\beta}^{\omega} = (\omega_1\beta_1, \ldots, \omega_p\beta_p)$ and ω_j is a weight of β_j. Then the weighted estimator $\hat{\boldsymbol{\beta}}^{\omega}$ obtained by minimizing $Q^{\omega}(\boldsymbol{\beta})$ satisfies the variable selection oracle property.

Theorem 3.4 (Variable selection oracle property)**.** *Let* $\omega_{min}^{\mathcal{D}} = \min\{|\omega_j| : j \in \mathcal{D}\}$*. Suppose that Conditions 1–3, (3.4) and (3.6) hold. If* $f_O'(\theta)$ *is upper bounded by a constant* c *for all* θ *and* λ *and* $\omega_{min}^{\mathcal{D}} \to \infty$*, then there exists a root-n consistent local minimizer* $\hat{\boldsymbol{\beta}}^{\omega}$ *of* $Q^{\omega}(\boldsymbol{\beta})$ *such that*

(a) $P(\hat{\boldsymbol{\beta}}_{\mathcal{D}}^{\omega} = 0) \to 1$*;*

(b) $\sqrt{n}\boldsymbol{u}^T\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1/2}\boldsymbol{D}_{\mathcal{A}\mathcal{A}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\omega} - \boldsymbol{\beta}_{0\mathcal{A}})$ *is asymptotically distributed as standard normal for every* $\boldsymbol{u} \in \mathbb{R}^s$ *with* $\|\boldsymbol{u}\|_2 = 1$*.*

Corollary 3.5. *Let* $\tilde{\boldsymbol{\beta}}$ *be an* n^{α}*-consistent estimator, i.e.,* $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2 = O_p(n^{-\alpha})$ *with* $0 < \alpha \leq 1/2$ *and* $\omega_{k_j} = 1/|\tilde{\beta}_{k_j}|^r$ *where* $r > 0$*. Then under the conditions of Theorem 3.4, there exists the local minimizer* $\hat{\boldsymbol{\beta}}_{\omega}$ *of* $Q^{\omega}(\boldsymbol{\beta})$ *such that* $P(\hat{\boldsymbol{\beta}}_{\mathcal{D}}^{\omega} = 0) \to 1$ *and* $\sqrt{n}\boldsymbol{u}^T\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1/2}\boldsymbol{D}_{\mathcal{A}\mathcal{A}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\omega} - \boldsymbol{\beta}_{0\mathcal{A}}) \to N(0,1)$ *in distribution for every* $\boldsymbol{u} \in \mathbb{R}^s$ *with* $\|\boldsymbol{u}\|_2 = 1$*.*

The boundedness condition for f_O' is trivial, as it is satisfied for most penalties, such as LASSO, SCAD, MCP and SICA. Theorems 3.2 and 3.4 indicate that the weighted estimators possess both the group selection oracle property and the variable selection oracle property, i.e., both of the important groups and important variables within selected groups can be identified with sufficiently high probability. Moreover, Corollary 3.5 provides a possible way for choosing weights that meet the requirements in Theorem 3.4. For example, we could use the LASSO estimator as the weight since the LASSO penalty is convex and its solution is globally optimal.

4. Simulation studies

In this section, we conducted simulation studies to evaluate the finite-sample properties of the proposed method. To this end, we generated survival data from the following additive hazards regression model

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t) + \boldsymbol{\beta}_0^T \boldsymbol{Z}(t),$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, and $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^T$ were subjected to $\lambda_0(t) + \boldsymbol{\beta}_0^T \boldsymbol{Z}(t) > 0$. To generate the covariate $\boldsymbol{Z}(t)$, we first simulated $R_1, \ldots, R_p$ independently from the standard normal distribution, and $M_1, \ldots, M_K$ from an AR(1) model with the initial standard normal distribution and $Cov(M_{j_1}, M_{j_2}) = 0.5^{|j_1 - j_2|}$ for $j_1, j_2 = 1, \ldots, K$. In the following examples, we considered two cases for generating covariates Z_j's.

Case 1: Covariates Z_j's are independent of time t and they were generated by

$$Z_j = (M_{g_j} + R_j)/4 \qquad (j = 1, \ldots, p),$$

where g_j is the group number that Z_j belonged to. In this case, we set $\lambda_0(t) = 1$.

Case 2: Covariates Z_j's are dependent on time t. We first generated

$$\eta_j = (M_{g_j} + R_j)/4 \qquad (j = 1, \ldots, p),$$

where g_j is the same as that in Case 1. Then we set $Z_j(t) = \eta_j t$ and let $\lambda_0(t) = 2t$.

We assumed Uniform$(\tau/2, \tau)$ for censoring time C, where the value of τ was chosen such that the censoring rate reached at 25%. We set sampling size $n = 250$ in Case 1 and $n = 600$ in Case 2.

Example 1. We considered $p = 50$ such that the dimensionality of the covariates is comparable to sampling size but smaller. The true coefficient $\boldsymbol{\beta}_0$ had values $(\boldsymbol{v}^T, \boldsymbol{v}^T, \boldsymbol{v}^T, \boldsymbol{0}^T)^T$ with $\boldsymbol{v} = (1, 0, -1, 0, 0,)^T$. The covariates were divided into 10 groups with equal size of five in each group. Thus, the sparsity dimension was $s = 6$ and there were 3 important groups.

Example 2. We considered $p = 500$ to compare the performance of various methods when p is larger than n. Similar to Example 1, the true coefficient $\boldsymbol{\beta}_0$ had values $(\boldsymbol{v}^T, \boldsymbol{v}^T, \boldsymbol{v}^T, \boldsymbol{0}^T)^T$, and the variables were divided into 100 groups with equal size of five in each group. In this example, there were 6 important variables and 3 important groups.

Example 3. We use this example to further demonstrate robustness of the proposed method with respect to the sparsity assumption. Similar to the setup in Example 2, we considered $p = 500$ and divided the variables into 100 groups with equal size of five in each group. Next, we set the first 18 elements of $\boldsymbol{\beta}_0$ to $(\boldsymbol{v}^T, \boldsymbol{v}^T, \boldsymbol{v}^T)^T$ and randomly chose 24 elements from positions 19 to 400 to have values $\{1, -1, \ldots, 1, -1\}$, such that the number of important groups reached 12 with a total of 30 important variables.

TABLE 1

Simulation results for Example 1: CMCP: composite MCP penalty; CSICA: composite SICA penalty; CSCAD: composite SCAD penalty; MSICA: MCP SCIA penalty; GTPR: group selection true positive rate; GFPR: group selection false positive rate; TPR: true positive rate for variable selection; FPR: false positive rate for variable selection; L_2-loss: $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2$; estimated standard errors are summarized in parentheses.

Penalty	GTPR	GFPR	TPR	FPR	L_2-loss
			Case 1		
CMCP	0.977(0.092)	0.046(0.096)	0.963(0.104)	0.013(0.022)	6.557
CSICA	0.967(0.106)	0.014(0.045)	0.953(0.112)	0.012(0.019)	6.003
CSCAD	0.967(0.111)	0.048(0.096)	0.950(0.120)	0.014(0.024)	6.499
MSICA	0.982(0.083)	0.064(0.113)	0.963(0.102)	0.016(0.024)	6.569
MCP	0.980(0.086)	0.056(0.103)	0.960(0.105)	0.014(0.023)	6.535
SICA	0.978(0.082)	0.081(0.109)	0.947(0.112)	0.017(0.020)	5.720
SCAD	0.980(0.086)	0.055(0.103)	0.960(0.105)	0.014(0.022)	6.534
			Case 2		
CMCP	0.958(0.115)	0.218(0.163)	0.924(0.138)	0.068(0.052)	1.375
CSICA	0.927(0.157)	0.151(0.154)	0.895(0.168)	0.084(0.061)	1.293
CSCAD	0.963(0.110)	0.298(0.192)	0.914(0.136)	0.068(0.048)	1.361
MSICA	0.972(0.099)	0.275(0.173)	0.927(0.133)	0.069(0.047)	1.396
MCP	0.972(0.099)	0.289(0.201)	0.898(0.140)	0.067(0.050)	1.340
SICA	0.962(0.111)	0.320(0.192)	0.912(0.139)	0.071(0.047)	1.213
SCAD	0.973(0.097)	0.276(0.195)	0.908(0.135)	0.066(0.049)	1.346
True model	1	0	1	0	0

TABLE 2

Simulation results for Example 2: CMCP: composite MCP penalty; CSICA: composite SICA penalty; CSCAD: composite SCAD penalty; MSICA: MCP SCIA penalty; GTPR: group selection true positive rate; GFPR: group selection false positive rate; TPR: true positive rate for variable selection; FPR: false positive rate for variable selection; L_2-loss: $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2$; estimated standard errors are summarized in parentheses.

Penalty	GTPR	GFPR	TPR	FPR	L_2-loss
			Case 1		
CMCP	0.965(0.102)	0.165(0.080)	0.945(0.121)	0.042(0.022)	7.366
CSICA	0.935(0.137)	0.053(0.039)	0.919(0.146)	0.018(0.013)	3.108
CSCAD	0.913(0.165)	0.189(0.120)	0.855(0.204)	0.041(0.028)	5.440
MSICA	0.970(0.096)	0.219(0.102)	0.934(0.134)	0.049(0.025)	8.083
MCP	0.935(0.144)	0.067(0.053)	0.806(0.185)	0.014(0.011)	2.923
SICA	0.947(0.134)	0.207(0.115)	0.903(0.169)	0.044(0.026)	5.109
SCAD	0.930(0.148)	0.072(0.058)	0.801(0.190)	0.015(0.012)	3.000
			Case 2		
CMCP	0.945(0.128)	0.198(0.051)	0.902(0.148)	0.045(0.012)	3.537
CSICA	0.900(0.164)	0.106(0.043)	0.881(0.171)	0.031(0.014)	2.347
CSCAD	0.922(0.149)	0.231(0.063)	0.830(0.181)	0.048(0.013)	3.423
MSICA	0.943(0.134)	0.171(0.053)	0.841(0.178)	0.039(0.012)	3.102
MCP	0.947(0.123)	0.189(0.058)	0.800(0.184)	0.040(0.013)	3.111
SICA	0.930(0.144)	0.215(0.061)	0.812(0.186)	0.044(0.013)	2.358
SCAD	0.950(0.119)	0.190(0.060)	0.798(0.184)	0.041(0.013)	3.109
True model	1	0	1	0	0

TABLE 3

Simulation results for Example 3: CMCP: composite MCP penalty; CSICA: composite SICA penalty; CSCAD: composite SCAD penalty; MSICA: MCP SCIA penalty; GTPR: group selection true positive rate; GFPR: group selection false positive rate; TPR: true positive rate for variable selection; FPR: false positive rate for variable selection; L_2-loss: $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2$; estimated standard errors are summarized in parentheses.

Penalty	GTPR	GFPR	TPR	FPR	L_2-loss
			Case 1		
CMCP	0.539(0.131)	0.136(0.094)	0.271(0.021)	0.021(0.013)	8.422
CSICA	0.446(0.093)	0.061(0.048)	0.240(0.015)	0.015(0.008)	5.924
CSCAD	0.524(0.162)	0.162(0.134)	0.231(0.022)	0.022(0.014)	6.813
MSICA	0.558(0.112)	0.146(0.112)	0.262(0.024)	0.024(0.013)	8.719
MCP	0.505(0.142)	0.078(0.066)	0.284(0.091)	0.013(0.008)	5.807
SICA	0.547(0.132)	0.141(0.102)	0.257(0.085)	0.019(0.012)	6.399
SCAD	0.526(0.157)	0.092(0.079)	0.297(0.099)	0.015(0.011)	6.185
True model	1	0	1	0	0
			Case 2		
CMCP	0.539(0.123)	0.454(0.099)	0.118(0.074)	0.071(0.018)	8.151
CSICA	0.247(0.087)	0.156(0.058)	0.049(0.057)	0.036(0.018)	6.498
CSCAD	0.606(0.112)	0.491(0.093)	0.126(0.066)	0.072(0.016)	8.403
MSICA	0.470(0.134)	0.323(0.086)	0.098(0.062)	0.066(0.022)	7.716
MCP	0.476(0.144)	0.391(0.106)	0.076(0.076)	0.058(0.018)	7.282
SICA	0.630(0.114)	0.499(0.090)	0.125(0.125)	0.072(0.016)	7.586
SCAD	0.469(0.145)	0.383(0.107)	0.074(0.074)	0.057(0.018)	7.250
True model	1	0	1	0	0

The simulation results based on 200 replicates are summarized in Tables 1–3. We compared four types of composite penalties and three types of variable selection penalties, i.e., composite MCP (CMCP), composite SICA (CSICA), composite SCAD (CSCAD), MCP SICA (MSICA), MCP, SICA and SCAD penalties. The tuning parameter was chosen using the 5-fold cross-validation principle. In the tables, we report the rates of correctly identifying the important groups (GTPR), the rates of incorrectly selecting unimportant groups (GFPR), the rates of correctly identifying the important variables (TPR), the rates of incorrectly selecting unimportant variables (FPR), and the L_2-loss for estimation accuracy for two cases. Table 1 indicates that all methods for two cases perform well for the number of selected important groups and variables when $p < n$. Table 2 shows that when the model is sparse enough but $p > n$, CSICA performs better than others correctly excluding unimportant groups. In addition, the computed L_2-loss results indicate that CSICA is more efficient in group selection with higher correction rate, thus providing higher accuracy. Table 3 provides further evidence showing that CSICA tends to select a sparser model than other selectors as the number of important variables or important groups increases.

Assuming the true sparse model is known, to compare the performance of the best model that ever exists on the solution path, we recorded the maximum number of correctly selected variables among all models on the solution path and averaged it over all replicates. The performance results from Cases 1 and 2 are

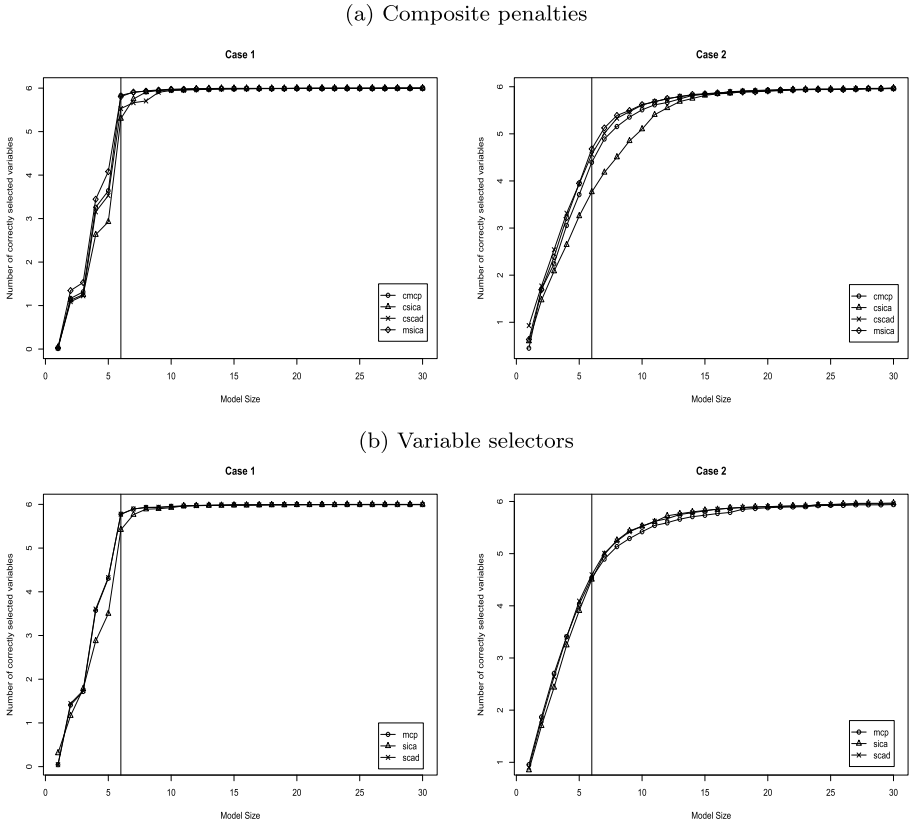


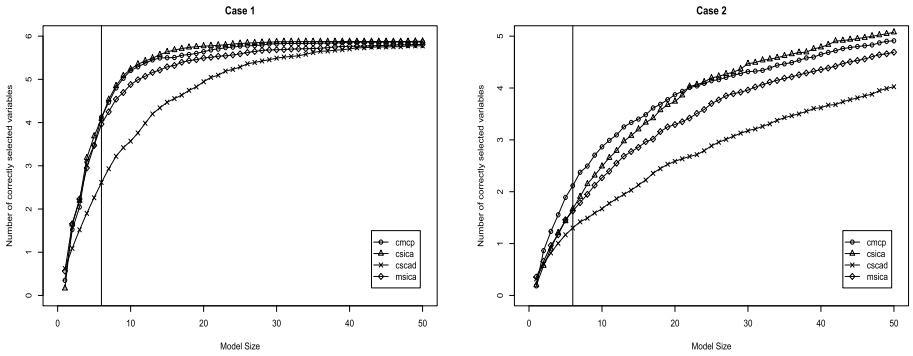
FIG 1. *Variable selection performance of various methods for two cases in Example 1. The vertical lines indicate the true sparsity dimension.*

presented in Figures 1–3. Figures 1 and 2 show that under the sparsity model, all of the proposed selectors perform well since they can identify all important variables immediately after the model size reaches the true sparsity dimension. On the contrary, when $p > n$ and s is large, Figure 3 shows that the proposed bi-selection method is comparable to the variable selection method in Case 1 and performs better than it in Case 2.

5. An application

We apply the proposed method to analyze the breast cancer data set containing the metastasis-free survival time. In the study of [24], 295 patients with primary breast carcinomas were classified as having a gene-expression signature associated with either a poor or a good prognosis. We focus on 144 patients having lymph node positive disease with censor rate at 66%. The data set is available

(a) Composite penalties



(b) Variable selectors

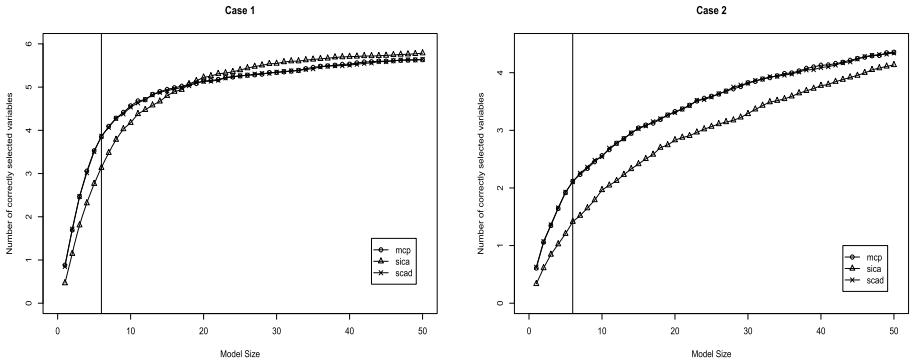
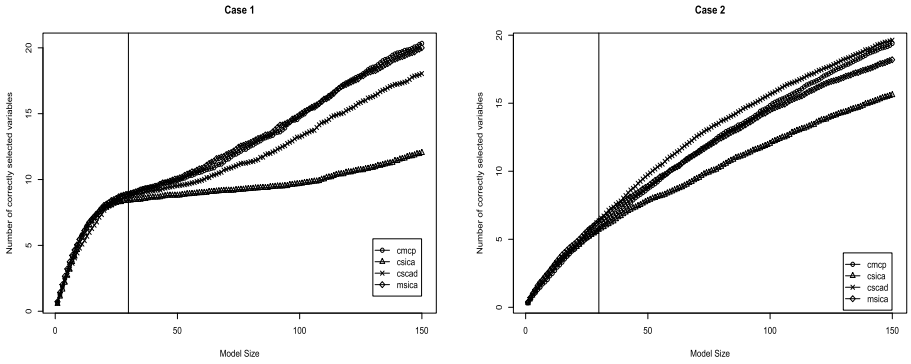


FIG 2. *Variable selection performance of various methods for two cases in Example 2. The vertical lines indicate the true sparsity dimension.*

in the **R** package "*penalized*". The primary objective of this study is to identify key risk factors impacting the survival time of breast cancer patients. We consider 5 clinical risk factors and 70 gene expression measurements, including *diameter* of the tumor (1 for >= 2 cm and 0 for < 2 cm), *number* of affected lymph nodes (1 for 1–3 and 0 for >= 4), estrogen receptor *status* (1 for positive and 0 for negative), *grade* of the tumor (1 for Well diff and 0 otherwise), *age* of the patient at diagnosis, and *gene expression measurements* of 70 prognostic genes.

Huang et al. [11] analyzed this dataset using the group bridge penalty and considered the multiplicative effects of covariates on the survival time. Suppose the survival time follows the additive hazards model (2.1). We conduct the group selection and variable selection procedures by minimizing the penalized psudoscore function (2.2) with composite penalties (CMCP, CSCAD, CSICA, MSICA) and variable selectors (MCP, SICA and SCAD).

(a) Composite penalties



(b) Variable selectors

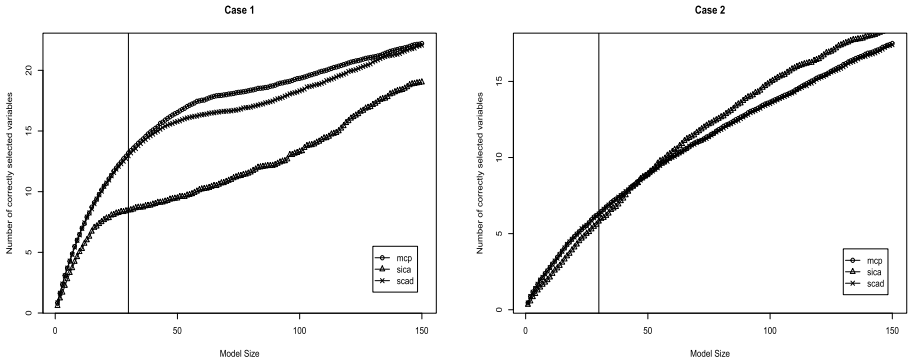
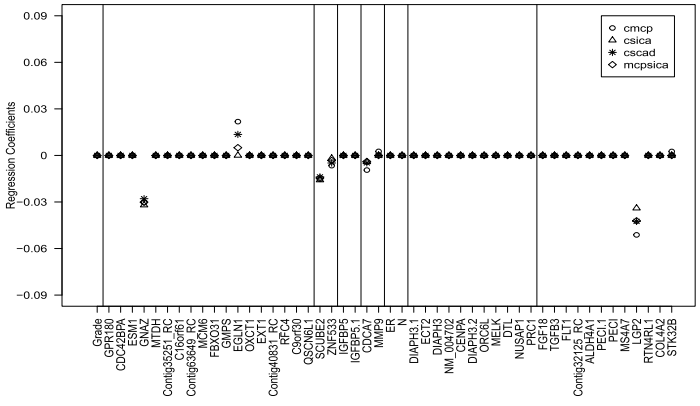


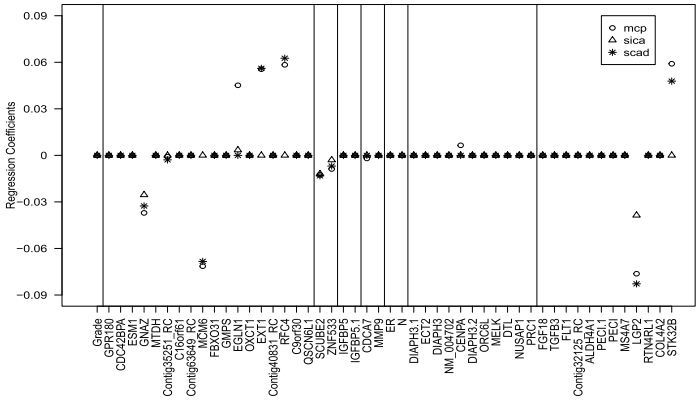
FIG 3. *Variable selection performance of various methods for two cases in Example 3. The vertical lines indicate the true sparsity dimension.*

We begin the statistical analysis by first reducing the model dimension to 50 through screening out the 25 most unimportant variables, and then group the remaining 25 relatively important variables into 8 distinct categories using dynamic clustering. In Figure 4, we show the selection results in panels (a) and (b), where the optimal tuning parameter is chosen by the 10-fold cross-validation. For comparison, in panel (c), we display the results using group bridge, adaptive lasso, and group lasso with the AIC principle in Huang et al. [11], where each block represents a group. The following findings are easily observed from these figures: (i) the selectors identify more important variables in the multiplicative hazards model than the additive hazards model; (ii) under the additive hazards model, all selectors identify genes $GNAZ$ and $SCUBE2$ as important variables; (iii) most of the selectors can identify 4 important groups in the additive hazards model compared to 8 important groups in the Cox model; (iv) the proposed method with the CSICA selects sparser model than others.

(a) Results in the additive hazards model with composite penlaties



(b) Results in the additive hazards model with variable selectors



(c) Results in the Cox model by AIC principle

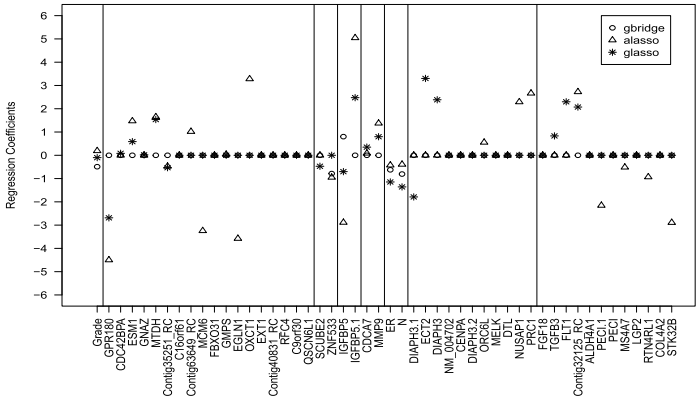


FIG 4. *Plots of variable and group selection in the breast cancer data by using composite selectors and variable selectors compared with the selection results in Cox model. gbridge represents the group bridge penalty, alasso represents the adaptive lasso penalty, glasso represents the group lasso penalty.*

6. Concluding remarks

In this article, we studied a class of regularized regression methods under the additive hazards model. The proposed approach can consistently identify both important groups and important variables within selected groups. Furthermore, we established the asymptotic properties of the proposed estimators for high-dimensional data. To efficiently compute the penalized estimator, we developed the local coordinate descent (LCD) algorithm through approximating the penalties by its first order linear part. The cross validation principle was adopted to determine the optimal tuning parameter. The numerical studies demonstrated that the proposed penalized approaches and the LCD algorithm work well.

The assumption of covariates exhibiting solely linear effects may be unrealistic. The true covariate effects may be in fact more complex, hence it is important to consider potential nonlinearity, especially when continuous covariates are involved. [10, 4] studied a partially linear Cox model including linear and nonlinear components. Under the additive hazards model, a further study is to consider the partially linear additive hazards model conditional on covariates $\boldsymbol{Z}$ and $\boldsymbol{X}$:

$$\lambda(t|\boldsymbol{Z},\boldsymbol{X}) = \lambda_0(t) + \boldsymbol{\beta}^T\boldsymbol{Z}(t) + \phi_1(X_1(t)) + \ldots + \phi_d(X_d(t)),$$

where λ_0 is an unspecified baseline hazard function, $\boldsymbol{\beta}$ is a p-dimensional regression parameter, and $\phi_1, \ldots, \phi_d$ are unknown smooth functions with d much smaller than p. A future research opportunity is to extend the approach proposed in this paper to a model with covariates exhibiting both linear and nonlinear effects, attaining the goal of simultaneously selecting both important groups and important variables.

Appendix A: Proofs of asymptotic results

A.1. Proof of Lemma 3.1

We first consider the $|\hat{\mathcal{A}}|$-dimensional subspace $\mathbb{B} = \{\boldsymbol{\beta} \in \mathbb{R}^p | \boldsymbol{\beta}_{\hat{\mathcal{D}}} = 0\}$. Inequality (3.3) ensures that $Q(\boldsymbol{\beta})$ is strictly convex in a neighborhood of $\hat{\boldsymbol{\beta}}$ in $\mathbb{B}$. Equation (3.1) implies that $\hat{\boldsymbol{\beta}}$ is a stationary point and therefore it is a strict local minimizer of $Q(\boldsymbol{\beta})$ in $\mathbb{B}$. Similar to the proof of Lemma A.1 in [16], we then only need to show that for any $\boldsymbol{\beta}_1 \in \mathbb{R}^p \backslash \mathbb{B}$ that lies in a sufficiently small neighborhood of $\hat{\boldsymbol{\beta}}$, and $\boldsymbol{\beta}_2$ which is the projection of $\boldsymbol{\beta}_1$ onto the subspace $\mathbb{B}$, we have $Q(\boldsymbol{\beta}_1) \geq Q(\boldsymbol{\beta}_2)$. Note that

$$\begin{aligned} Q(\boldsymbol{\beta}_1) - Q(\boldsymbol{\beta}_2) &= \sum_{j \in \hat{\mathcal{D}}: \beta_{1j} \neq 0} \frac{\partial Q(\boldsymbol{\beta}^*)}{\partial \beta_j} \beta_{1j} \\ &= \sum_{j \in \hat{\mathcal{D}}: \beta_{1j} \neq 0} \{-U_j(\boldsymbol{\beta}^*) + \lambda \frac{\partial \rho_\lambda(|\boldsymbol{\beta}^*|)}{\partial |\beta_j|} \operatorname{sgn}(\beta_j^*)\} \beta_{1j}, \end{aligned}$$

where $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^T$ is a point on the line segment between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Using (3.2) and the fact that $\text{sgn}(\beta_j^*) = \text{sgn}(\beta_{1j})$, the desired conclusion is drawn.

A.2. Proof of Theorem 3.2

To check conditions (3.1) and (3.3) in Lemma 3.1, we determine $\hat{\boldsymbol{\beta}}$ on the subspace $\mathbb{B} = \{\boldsymbol{\beta} \in \mathbb{R}^p | \boldsymbol{\beta}_{\hat{\mathcal{D}}} = 0\}$. Since $\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{b} - \boldsymbol{V}\boldsymbol{\beta}$, we have

$$\boldsymbol{U}_{\mathcal{A}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{U}_{\mathcal{A}}(\boldsymbol{\beta}_0) - \boldsymbol{V}_{\mathcal{A}\mathcal{A}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}).$$

Substituting this equation into (3.1) gives that

$$\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}} = \boldsymbol{V}_{\mathcal{A}\mathcal{A}}^{-1}\left\{\boldsymbol{U}_{\mathcal{A}}(\boldsymbol{\beta}_0) - \lambda\frac{\partial\rho_\lambda(|\hat{\boldsymbol{\beta}}|)}{\partial|\beta_j|_{j\in\mathcal{A}}} \circ \text{sgn}(\hat{\boldsymbol{\beta}}_{\mathcal{A}})\right\}. \tag{A.1}$$

According to the proof of Theorem 1 in Lin and Lv [16], the following inequalities hold with large enough probability:

$$\begin{aligned}
&\|\boldsymbol{U}_{\mathcal{A}}(\boldsymbol{\beta}_0)\|_\infty < \frac{1}{2cn^\gamma}\frac{\alpha}{4}\lambda\rho'(\mathbf{0}+),\ \|\boldsymbol{U}_{\mathcal{C}}(\boldsymbol{\beta}_0)\|_\infty < \frac{\alpha}{4}\lambda\rho'(\mathbf{0}+),\\
&\|\boldsymbol{V}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty < 2\phi,\ \|\boldsymbol{V}_{\mathcal{C}\mathcal{A}}\boldsymbol{V}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty < \left\{(1-\frac{\alpha}{2})\frac{\rho'(\mathbf{0}+)}{c_n^*\rho'_\lambda(\boldsymbol{d})}\right\} \wedge (2cn^\gamma),\\
&\Lambda_{\min}(\boldsymbol{V}_{\mathcal{A}\mathcal{A}}) > \lambda\kappa_0.
\end{aligned} \tag{A.2}$$

The subsequent steps are proceeded based on these inequalities.

Define function $f : \mathbb{R}^s \to \mathbb{R}^s$ by

$$f(\boldsymbol{\theta}) = \boldsymbol{\beta}_{0\mathcal{A}} + \boldsymbol{V}_{\mathcal{A}\mathcal{A}}^{-1}\{\boldsymbol{U}_{\mathcal{A}}(\boldsymbol{\beta}_0) - \overline{p}(|\boldsymbol{\theta}|) \circ \text{sgn}(\boldsymbol{\theta})\},$$

where

$$\overline{p}(|\boldsymbol{\theta}|) = \lambda\frac{\partial\rho_\lambda(|\boldsymbol{\beta}|)}{\partial|\beta_j|_{j\in\mathcal{A}}}$$

with $\boldsymbol{\beta} = (\boldsymbol{\theta}^T, 0, \ldots, 0)^T$. Let

$$\mathcal{K} = \{\boldsymbol{\theta} \in \mathbb{R}^s |\ \ \|\boldsymbol{\theta} - \boldsymbol{\beta}_{0\mathcal{A}}\|_\infty \le c_1\phi\lambda\rho'(\mathbf{0}+)\}$$

for some constant c_1. Then for $\boldsymbol{\theta} \in \mathcal{K}$,

$$\begin{aligned}
\|f(\boldsymbol{\theta}) - \boldsymbol{\beta}_{0\mathcal{A}}\|_\infty &\le \|\boldsymbol{V}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty\{\|\boldsymbol{U}_{\mathcal{A}}(\boldsymbol{\beta}_0)\|_\infty + \|\overline{p}(\boldsymbol{\theta})\|_\infty\}\\
&\le 2\phi\lambda\rho'(\mathbf{0}+)\Big\{\frac{1}{2cn^\gamma}\frac{\alpha}{4} + c_n^*\Big\} \le c_1\phi\lambda\rho'(\mathbf{0}+).
\end{aligned}$$

So we have $f(\mathcal{K}) \subset \mathcal{K}$, where f is a continuous function on the convex compact hypercube $\mathcal{K}$, which yields that (A.1) has a solution $\hat{\boldsymbol{\beta}}_{\mathcal{A}} \in \mathcal{K}$ by using Brouwer's fixed point theorem. Furthermore, from the definition of $\mathcal{K}$ and (3.4) that $d \ge c_1\phi\lambda\rho'(\mathbf{0}+)$, we have $\|\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}\|_\infty \le d$. This shows that $\text{sgn}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}) = \text{sgn}(\boldsymbol{\beta}_{0\mathcal{A}})$.

Therefore, we get an estimator $\hat{\boldsymbol{\beta}}$ satisfying (3.1). In addition, since $\|\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}\|_\infty \le d$, it follows from $\Lambda_{\min}(\boldsymbol{V}_{\mathcal{A}\mathcal{A}}) \ge \lambda\kappa_0$ that $\Lambda_{\min}(\boldsymbol{V}_{\mathcal{A}\mathcal{A}}) \ge \lambda\kappa(\rho_\lambda; \hat{\boldsymbol{\beta}}_{\mathcal{A}})$, and then (3.3) holds. Therefore, $\mathcal{A} \subset \hat{\mathcal{A}}$, and then $\mathcal{B} \subset \hat{\mathcal{B}}$. To conclude the group selection consistency, it remains to check (3.2) by taking $\mathcal{A}$ and $\mathcal{D}$ in Lemma 3.1 as $\mathcal{B}$ and $\mathcal{C}$ respectively. Since

$$\boldsymbol{U}_{\mathcal{C}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{U}_{\mathcal{C}}(\boldsymbol{\beta}_0) - \boldsymbol{V}_{\mathcal{C}\mathcal{A}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}),$$

using inequalities (A.2) we can get that

$$\begin{aligned}
&\|\boldsymbol{U}_{\mathcal{C}}(\hat{\boldsymbol{\beta}})\|_\infty \\
\le&\|\boldsymbol{U}_{\mathcal{C}}(\boldsymbol{\beta}_0)\|_\infty + \|\boldsymbol{V}_{\mathcal{C}\mathcal{A}}\boldsymbol{V}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \left\{ \|\boldsymbol{U}_{\mathcal{A}}(\boldsymbol{\beta}_0)\|_\infty + \lambda \left|\left| \frac{\partial \rho_\lambda(|\hat{\boldsymbol{\beta}}|)}{\partial |\beta_j|_{j\in\mathcal{A}}} \right|\right|_\infty \right\} \\
\le&\frac{\alpha}{4}\lambda\rho'(\mathbf{0}+) + 2cn^\gamma \cdot \frac{1}{2cn^\gamma} \cdot \frac{\alpha}{4}\lambda\rho'(\mathbf{0}+) + (1-\frac{\alpha}{2})\lambda \frac{\rho'(\mathbf{0}+)}{c_n^*\rho'_\lambda(\boldsymbol{d})} c_n^*\rho'_\lambda(\boldsymbol{d}) \\
\le&\lambda\rho'(\mathbf{0}+),
\end{aligned}$$

which means that $\hat{\mathcal{B}} \subset \mathcal{B}$. Theorem 1 is concluded.

A.3. Proof of Theorem 3.3

Since we have verified (3.1) and (3.3) in the proof of Theorem 3.2, so it suffices to check (3.2) by Lemma 3.1. Similar to (A.2), we note that

$$\|\boldsymbol{U}_{\mathcal{D}}(\boldsymbol{\beta}_0)\|_\infty < \frac{\alpha}{4}\lambda\rho'(\mathbf{0}+),\ \|\boldsymbol{V}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty < 2\phi.$$

Since

$$\boldsymbol{U}_{\mathcal{D}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{U}_{\mathcal{D}}(\boldsymbol{\beta}_0) - \boldsymbol{V}_{\mathcal{D}\mathcal{A}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}),$$

we have

$$\begin{aligned}
&\|\boldsymbol{U}_{\mathcal{D}}(\hat{\boldsymbol{\beta}})\|_\infty \\
\le&\|\boldsymbol{U}_{\mathcal{D}}(\boldsymbol{\beta}_0)\|_\infty + \|\boldsymbol{V}_{\mathcal{D}\mathcal{A}}\boldsymbol{V}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty \left\{ \|\boldsymbol{U}_{\mathcal{A}}(\boldsymbol{\beta}_0)\|_\infty + \lambda \left|\left| \frac{\partial \rho_\lambda(|\hat{\boldsymbol{\beta}}|)}{\partial |\beta_j|_{j\in\mathcal{A}}} \right|\right|_\infty \right\} \\
\le&\frac{\alpha}{4}\lambda\rho'(\mathbf{0}+) + 2cn^\gamma \cdot \frac{1}{2cn^\gamma} \cdot \frac{\alpha}{4}\lambda\rho'(\mathbf{0}+) + (1-\frac{\alpha}{2})\lambda \frac{\rho'(\mathbf{0}+)}{c_n^*\rho'_\lambda(\boldsymbol{d})} c_n^*\rho'_\lambda(\boldsymbol{d}) \\
\le&\lambda\rho'(\mathbf{0}+).
\end{aligned}$$

This relation yields (3.2) by (3.7). So we can draw the conclusion of variable selection consistency.

For the asymptotic normality, we note that

$$\begin{aligned}
&\sqrt{n}\boldsymbol{u}^T\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-\frac{1}{2}}\boldsymbol{D}_{\mathcal{A}\mathcal{A}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}-\boldsymbol{\beta}_{0\mathcal{A}})\\
=&\sqrt{n}\boldsymbol{u}^T\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1/2}\boldsymbol{U}_{\mathcal{A}}(\boldsymbol{\beta}_0)\\
&+\sqrt{n}\boldsymbol{u}^T\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1/2}\boldsymbol{D}_{\mathcal{A}\mathcal{A}}(\boldsymbol{V}_{\mathcal{A}\mathcal{A}}^{-1}-\boldsymbol{D}_{\mathcal{A}\mathcal{A}}^{-1})\boldsymbol{U}_{\mathcal{A}}(\boldsymbol{\beta}_0)\\
&-\sqrt{n}\boldsymbol{u}^T\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1/2}\boldsymbol{D}_{\mathcal{A}\mathcal{A}}\boldsymbol{V}_{\mathcal{A}\mathcal{A}}^{-1}\lambda\frac{\partial\rho_\lambda(|\hat{\boldsymbol{\beta}}|)}{\partial|\beta_j|_{j\in\mathcal{A}}}\circ\operatorname{sgn}(\hat{\boldsymbol{\beta}}_{\mathcal{A}})\\
\hat{=}&T_1+T_2+T_3.
\end{aligned}$$

We consider term T_1. It is obviously that

$$\boldsymbol{u}^T\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1/2}\boldsymbol{W}_{\mathcal{A}\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1/2}\boldsymbol{u}=1+\boldsymbol{u}^T\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1/2}(\boldsymbol{W}_{\mathcal{A}\mathcal{A}}-\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}})\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1/2}\boldsymbol{u},$$

where

$$\boldsymbol{W}=\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau(\boldsymbol{Z}_i(t)-\overline{\boldsymbol{Z}}(t))^{\otimes 2}dN_i(t).$$

By Lemma A.5 in [16], the second term of the above expression is bounded by

$$\begin{aligned}
&\|\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1/2}\|_2\|\boldsymbol{W}_{\mathcal{A}\mathcal{A}}-\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}\|_2\|\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1/2}\|_2\\
=&\Lambda_2^{-1/2}sO_p(n^{-1/2})\Lambda_2^{-1/2}=\frac{s}{\Lambda_2}O_p(n^{-1/2})=o_p(1).
\end{aligned}$$

Thus, T_1 is asymptotically standard normal by using the martingale central limit theorem [1].

We then consider term T_2. Let Ω_L represent the event that $\max_j\sup_{t\in[0,\tau]}|Z_j(t)|\leq L$ for $L>0$. Since

$$\begin{aligned}
P(\Lambda_{\min}(\boldsymbol{V}_{\mathcal{A}\mathcal{A}})\leq\Lambda_1/2|\Omega_L))=&P(|\Lambda_{\min}(\boldsymbol{V}_{\mathcal{A}\mathcal{A}})-\Lambda_1|\geq\Lambda_1/2|\Omega_L))\\
\leq&s^2M\exp\Big\{-K\frac{n}{L^4}\Big(\frac{\Lambda_1^2}{s^2}\wedge 1\Big)\Big\},
\end{aligned}$$

it follows that $\Lambda_{\min}(\boldsymbol{V}_{\mathcal{A}\mathcal{A}})>\Lambda_1/2$ with probability at least

$$1-s^2M\exp\Big\{-K\frac{n}{L^4}\Big(\frac{\Lambda_1^2}{s^2}\wedge 1\Big)\Big\}-pM\exp(-KL^r),$$

and then

$$\|\boldsymbol{V}_{\mathcal{A}\mathcal{A}}^{-1}\|_2=\frac{1}{\Lambda_{\min}(\boldsymbol{V}_{\mathcal{A}\mathcal{A}})}<\frac{2}{\Lambda_1}. \tag{A.3}$$

Moreover, by Lemma A.3 and Lemma A.4 in [16], we obtain that $\|\boldsymbol{V}_{\mathcal{A}\mathcal{A}}-\boldsymbol{D}_{\mathcal{A}\mathcal{A}}\|_2=sO_p(n^{-1/2})$ and $\|\boldsymbol{U}_{\mathcal{A}}(\boldsymbol{\beta}_0)\|_2=\sqrt{s}O_p(n^{-1/2})$. Thus, it follows from $\|\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1/2}\boldsymbol{D}_{\mathcal{A}\mathcal{A}}\|_2=\Lambda_3^{-1/2}$ and $\|\boldsymbol{D}_{\mathcal{A}\mathcal{A}}^{-1}\|_2=1/\Lambda_1$ that

$$\begin{aligned}
|T_2|&\leq\sqrt{n}\|\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1/2}\boldsymbol{D}_{\mathcal{A}\mathcal{A}}\|_2\|\boldsymbol{D}_{\mathcal{A}\mathcal{A}}^{-1}\|_2\|\boldsymbol{V}_{\mathcal{A}\mathcal{A}}-\boldsymbol{D}_{\mathcal{A}\mathcal{A}}\|_2\|\boldsymbol{V}_{\mathcal{A}\mathcal{A}}^{-1}\|_2\|\boldsymbol{U}_{\mathcal{A}}(\boldsymbol{\beta}_0)\|_2\\
&\leq\sqrt{n}\Lambda_3^{-1/2}\Lambda_1^{-1}sO_p(n^{-1/2})2\Lambda_1^{-1}\sqrt{s}O_p(n^{-1/2})\\
&=\frac{2s^{3/2}}{\Lambda_1^2\Lambda_3^{1/2}}O_p(n^{-1/2})=o_p(1).
\end{aligned}$$

Thus, by (A.3) and (3.6), we have

$$|T_3| \le \sqrt{n}\|\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}}^{-1/2}\boldsymbol{D}_{\mathcal{A}\mathcal{A}}\|_2 \cdot \|\boldsymbol{V}_{\mathcal{A}\mathcal{A}}^{-1}\|_2 \cdot \lambda c_n^* \sqrt{s}\rho_\lambda'(\boldsymbol{d}) \le \frac{2\sqrt{ns}c_n^*\lambda\rho_\lambda'(\boldsymbol{d})}{\Lambda_1\Lambda_3^{1/2}} \to 0.$$

Theorem 3.3 is concluded by choosing the optimal L.

A.4. Proof of Theorem 3.4

Since

$$\begin{aligned}\frac{\partial\rho_\lambda(|\boldsymbol{\beta}^\omega|)}{\partial|\beta_j|}\Big|_{\beta_j=0} &= \sum_{k:A_k\ni j} f_O'\Big(\sum_{i=1}^{K} f_I(|\omega_{k_i}\beta_{k_i}|)\Big) f_I'(|\omega_j\beta_j|)\Big|_{\beta_j=0}\cdot|\omega_j| \\ &= f_I'(0+)|\omega_j| \sum_{k:A_k\ni j} f_O'\Big(\sum_{i:\beta_{k_i}\neq 0} f_I(|\omega_{k_i}\beta_{k_j}|)\Big),\end{aligned}$$

we have

$$\min_{j\in\mathcal{D}} \frac{\partial\rho_\lambda(|\boldsymbol{\beta}^\omega|)}{\partial|\beta_j|} \ge c f_I'(0+) f_O'(0+)\omega_{min}^{\mathcal{D}} = c\rho'(\mathbf{0}+)\omega_{min}^{\mathcal{D}}$$

by using the upper bound of $f_O'(\cdot)$, where c is a constant. Hence condition $\omega_{min}^{\mathcal{D}} \to \infty$ implies that (3.7) holds. Theorem 3.4 is concluded from Theorem 3.3.

Acknowledgements

The authors would like to thank the Editor, the Associate Editor and the two reviewers for their constructive and insightful comments and suggestions that greatly improved the paper.

References

[1] Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics* **10** 1100–1120. MR0673646

[2] Bradic, J., Fan J. and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionalilty. *The Annals of Statistics* **39** 3092–3120. MR3012402

[3] Breheny, P. and Huang, J. (2009). Penalized methods for bi-level selection. *Statistics and its Interface* **2** 369–380. MR2540094

[4] Cao, Y., Huang, J., Liu, Y. and Zhao, X. (2016). Sieve estimation of Cox models with latent structures. *Biometrics* **72** 1086–1097. MR3591593

[5] Cox, D. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London. MR0751780

[6] Fan, J. (1997). Comments on "Wallets in Statistics: A Review," by A. Antoniadis. *Journal of the Italian Statistical Society* **6** 131–138.

[7] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360. MR1946581
[8] Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics* **30** 74–99. MR1892656
[9] Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory* **57** 5467–5484. MR2849368
[10] Huang, J. (1999). Efficient estimation of the partly linear additive Cox model. *The Annals of Statistics* **27** 1536–1563. MR1742499
[11] Huang, J., Liu, L., Liu, Y. and Zhao, X. (2014). Group selection in the Cox model with a diverging number of covariates. *Statistica Sinica* **24** 1787–1810 MR3308663
[12] Huang, J., Ma, S., Xie, H. and Zhang, T. (2009). A group bridge approach for variable selection. *Biometrika* **96** 339–355. MR2507147
[13] Kim, Y., Kim, J. and Kim, Y. (2006). The blockwise sparse regression. *Statistica Sinica* **16** 375–390. MR2267240
[14] Leng, C. and Ma, S. (2007). Path consistent model selection in additive risk model via Lasso. *Statistics in Medicine* **26** 3753–3770. MR2395831
[15] Lin, D. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika* **81** 61–71. MR1279656
[16] Lin, W. and Lv, J. (2013). High-dimensional sparse additive hazards regression. *Journal of the American Statistical Association* **108** 247–264. MR3174617
[17] Loh, P.-L. and Wainwright, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research* **16** 559–616. MR3335800
[18] Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics* **37** 3498–3528. MR2549567
[19] Ma, S., Song, X. and Huang, J. (2007). Supervised group lasso with applications to microarray data analysis. *BMC Bioinformatics* **8** 60.
[20] Martinussen, T. and Scheike, T.H. (2009). Covariate selection for the semiparametric additive risk model. *Scandinavian Journal of Statistics* **36** 602–619. MR2572578
[21] Park, C. and Yoon, Y. (2011). Bridge regression: adaptivity and group selection. *Journal of Statistical Planning and Inference* **141** 3506–3519. MR2817359
[22] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58** 267–288. MR1379242
[23] Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine* **16** 385–395.
[24] van de Vijver, M., et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine* **47** 1999–2009.
[25] Wang, S. and Xiang, L. (2017). Penalized empirical likelihood inference

for sparse additive hazards regression with a diverging number of covariates. *Statistics and Computing* **27** 1347–1364. MR3647101

[26] 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. MR2212574

[27] Zhao, P., Rocha, G. and Yu, B. (2009). Grouped and hierarchical model selection through composite absolute penalties. *The Annals of Statistics* **37** 3468–3497. MR2549566

[28] Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38** 894–942. MR2604701

[29] Zhang, H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika* **94** 691–703. MR2410017

[30] Zhang, H., Sun, L., Zhou, Y. and Huang, J. (2017). Oracle inequalities and selected consistency for weighted lasso in high-dimensional additive hazards model. *Statistica Sinica* **27** 1903–1920. MR3726771

[31] Zhong, P.-S., Hu, T. and Li, J. (2015). Tests for coefficients in high-dimensional additive hazard models. *Scandinavian Journal of Statistics* **42** 649–664. MR3391684

[32] Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429. MR2279469