Statistics and Its Interface Volume 14 (2021) 197–209

A non-marginal variable screening method for the varying coefficient Cox model

Lianqiang Qu and Liuquan Sun*

The varying coefficient model has become a very popular statistical tool for describing the dynamic effects of covariates on the response. In this article, we develop a new variable screening method for the varying coefficient Cox model based on the kernel smoothing and group learning methods. The sure screening property is established for ultrahigh-dimensional settings. In addition, an iterative groupwise hard-thresholding algorithm is developed to implement our method. Simulation studies are conducted to evaluate the finite sample performances of the proposed method. An application to an ovarian cancer dataset is provided.

AMS 2000 subject classifications: 62N01, 62G08.
Keywords and phrases: Cox model, Kernel smoothing, Non-marginal screening, Ultrahigh-dimensionality, Varying coefficient.

1. INTRODUCTION

The Cox proportional hazards model is one of the most popular semiparametric regression models for analyzing survival data, which assumes that the regression coefficients are constant over time. In reality, however, the regression coefficients may vary over time [31, 38, 45]. Varying coefficient models provide more flexibility in modelling covariate effects, and also can reveal deep insights into functional and complex interactive effects of covariates [14, 25, 27]. This study is motivated by an analysis of an ovarian cancer dataset from the Cancer Genome Atlas [30]. Some preliminary analysis on this data indicates that the effects of genes on the patient's survival time are age-dependent (see Section 5.3 for more details). In this article, we consider the following varying coefficient Cox model for the analysis:

$$\lambda\big(t|Z,V\big) = \lambda_0(t)\exp\{\beta(V)^T Z\}, \tag{1}$$

where Z is a $p \times 1$ vector of candidate covariates, V is an exposure variable, $\lambda_0(t)$ is an unknown baseline hazard function, and $\beta(V)$ is a vector of unspecified smooth functions, which characterizes the varying effects of Z with respect to V.

With the advancement in technology for data collection and storage, the ultrahigh-dimensional data are frequently encountered in many research areas, such as genetic microarray, biomedical imaging and economics. The penalized methods have been available for the standard Cox model, such as Lasso [32], SCAD [11] and adaptive Lasso [42]. For Cox models with time-varying coefficients, the group penalized methods have also been extensively studied. For example, [38] proposed an adaptive Lasso procedure for variable selection and structure identification. [16] considered variable selection by using the group SCAD-type and adaptive group Lasso estimators. For other related works, we can refer to [17, 24, 36] and the references within. However, when the number of covariates is larger than the sample size, the penalized variable selection methods may be computationally infeasible, and challenging to achieve model selection consistency [13]. Thus, a variable screening step is necessary before variable selection is carried out.

[13] proposed the sure independence screening (SIS) method to select important variables for the ultrahigh-dimensional linear regression model. Because of its good numerical performance and novel theoretical properties, the SIS idea has been extensively studied in varying coefficient regression models [9, 10, 25, 29, 35]. In addition, the SIS idea has been quickly adapted in survival analysis. Examples include [15, 18, 23, 28, 41, 43]. [19] gave a selective review of variable screening procedures for survival data with ultrahigh-dimensional covariates.

However, the simple SIS methods face a number of challenges. For example, it may miss some important variables that are marginally unrelated but jointly related to the response, and may give misleading results when there exist strongly correlations among the covariates. To overcome these issues, [39] proposed a sure joint screening procedure for the Cox model. [1] suggested a sure group joint screening method for clustered survival data. [20] considered a conditional screening method by computing the marginal contribution of each covariate. [21] introduced a forward variable selection procedure based on the partial likelihood. Since these methods consider the joint effects of candidate covariates in the variable screening process rather than the marginal effect of individual variable, we call them as the non-marginal variable screening (NOVAS) methods. However, these existing NOVAS methods, without considering the dynamic effects of covariates, can not be directly applied to model (1). Recently, [40] proposed a non-marginal variable screening procedure for the ultrahigh-dimensional vary-

*Corresponding author.

ing coefficient Cox model, in which the spline-based method was employed.

In this article, we develop a NOVAS approach for model (1), which naturally combines the kernel smoothing and group learning methods. Although our method shares the same spirit as the work of [40], it is not a trivial extension of the existing methods. Specifically, we approximate the varying coefficient vector $\beta(\cdot)$ using a local constant fitting method, and estimate $\beta(\cdot)$ by applying a mixture of local partial likelihood with a group sparse constraint. Then we select the important variables based on the proposed estimator. This procedure can be regarded as a nonparametric version of the group variable screening [10]. Moreover, we show that the proposed method enjoys the sure screening property [13] without assuming the asymptotic stability and Lindeberg conditions that were adopted by [40]. The sure screening property guarantees that with probability tending to 1, the selected model includes the true model. In addition, an efficient iterative groupwise hard-thresholding (IGHT for short) algorithm is developed to carry out our screening method, and the convergence properties of the IGHT algorithm is also established. The high efficiency of our method is demonstrated through extensive simulation studies.

The rest of this article is organized as follows. Section 2 presents our NOVAS approach for model (1), and an IGHT algorithm is proposed. Section 3 establishes the sure screening properties. Section 4 discusses some tuning parameter selection issues for implementation of the proposed method. Section 5 presents simulation studies to evaluate the empirical performance of the proposed method. Also, an application to an ovarian cancer dataset is provided, and a screening-based penalized method is developed. Section 6 concludes the article with discussion. Proofs and technical details are given in the Appendix.

2. NON-MARGINAL VARIABLE SCREENING METHOD

2.1 Screening method

Let T be the failure time and C be the censoring time. Define $X = \min(T, C)$, and $\Delta = I(T \le C)$, where $I(\cdot)$ is the indictor function. Assume that T and C are independent given Z and V. The observed data consist of n independent and identically distributed replicates of (X, Δ, Z, V), denoted by $(X_i, \Delta_i, Z_i, V_i)$ $(i = 1, \ldots, n)$. Define the counting process $N_i(t) = I(X_i \le t, \Delta_i = 1)$, and the at-risk process $Y_i(t) = I(X_i \ge t)$. For model (1), following [12], for each given v, the logarithm of the local partial likelihood is given by

$$\begin{aligned}(2)\quad &\ell_n\{\beta(v)\}\\ &=\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} K_h(V_i - v)\beta(v)^T Z_i dN_i(t)\\ &\quad - \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} K_h(V_i - v)\\ &\quad \times \log\Bigg(\sum_{j=1}^{n} K_h(V_j - v)Y_j(t)\exp\{\beta(v)^T Z_j\}\Bigg)dN_i(t),\end{aligned}$$

where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function, h is a bandwidth, and τ is a prespecified constant such that $P(X_i \ge \tau) > 0$.

When p $(< n)$ is fixed, we can estimate $\beta(v)$ using the maximizer of (2). However, it is challenging to maximize (2) when p is larger than n. To overcome this difficulty, we assume that the true parameter vector $\beta^*(v) = (\beta_1^*(v), \ldots, \beta_p^*(v))^T$ is sparse. That is, the cardinality of the set $M^* = \{j : \beta_j^*(v) \not\equiv 0,\ 1 \le j \le p\}$ is less than n. Let $D = (\beta(V_1), \ldots, \beta(V_n)) = (d_1, \ldots, d_p)^T \in R^{p\times n}$, and define

$$\mathcal{L}(D) \equiv \frac{1}{n}\sum_{i=1}^{n}\ell_n\big\{\beta(V_i)\big\}.$$

Based on the sparsity assumption on $\beta^*(v)$, we propose to minimize the following objective function with a group sparse constraint:

$$(3)\qquad -\mathcal{L}(D) \ \text{ subject to } \ ||D||_{\text{row}} \le k,$$

where $k < n$ is a pre-specified positive integer, $||A||_{\text{row}} = \sum_{j=1}^{p} I(\|a_j\|_2 \ne 0)$ for any matrix $A = (a_1, \ldots, a_p)^T \in R^{p\times n}$, and $\|a_j\|_2$ denotes the Euclidean norm, i.e., $\|a_j\|_2 = (a_j^T a_j)^{1/2}$ for $1 \le j \le p$. Because the objective function (3) is constructed via the kernel smoothing technique, we refer to the proposed method as the kernel non-marginal variable screening (KNOVAS for short) method. Let $\hat{D}$ be the minimizer of (3) for a given k.

2.2 Iterative groupwise hard thresholding algorithm

Let $\dot{f}(x)$ denote the first derivative of any function $f(\cdot)$. For an arbitrary matrix A, denote the trace by $\text{tr}(A)$, and the Frobenius norm by $||A||_F$. For each given D and B around D, consider a quadratic approximation of $\mathcal{L}(B)$:

$$\mathcal{Q}_t(B|D) = \mathcal{L}(D) + \frac{1}{n}\text{tr}\{(B - D)^T\tilde{\mathcal{L}}(D)\} - \frac{t}{2n}||B - D||_F^2,$$

where t is a pre-specified positive constant, and $\tilde{\mathcal{L}}(D) = (\dot{\ell}_n\{\beta(V_1)\}, \ldots, \dot{\ell}_n\{\beta(V_n)\}) \in R^{p\times n}$. It can be seen that $\mathcal{Q}_t(D|D) = \mathcal{L}(D)$, and thus $\mathcal{Q}_t(B|D)$ approximates $\mathcal{L}(D)$ well for B close to D. Based on $\mathcal{Q}_t(B|D)$, we can obtain an iterative algorithm to solve (3):

$$\begin{aligned}(4)\qquad \hat{B} =& \arg\min_{B\in R^{p\times n}}\{-\mathcal{Q}_t(B|D)\}\\ &\text{subject to } \ ||B||_{\text{row}} \le k.\end{aligned}$$

After ignoring constant terms, (4) can be rewritten as

$$\hat{B} = \arg\min_{B\in R^{p\times n}} \frac{t}{2n}||B-(D+t^{-1}\tilde{\mathcal{L}}(D))||_F^2 \quad \text{subject to } ||B||_{\text{row}} \le k. \tag{5}$$

Proposition 2.1. *Let $\tilde{D} = (\tilde{d}_1,\ldots,\tilde{d}_p)^T \in R^{p\times n}$ be any matrix with $\tilde{d}_j = (\tilde{d}_{j1},\ldots,\tilde{d}_{jn})^T \in R^n$. If $\hat{B} = (\hat{b}_1,\ldots,\hat{b}_p)^T$ is an optimal solution to the following problem:*

$$\min_{B\in R^{p\times n}} ||B-\tilde{D}||_F^2 \quad \textit{subject to} \quad ||B||_{row} \le k,$$

then it has a closed form and its jth row is defined as

$$\hat{b}_j = \tilde{d}_j I(\tilde{d}_j^* \ge \tilde{d}_{(k)}^*), \quad 1\le j\le p, \tag{6}$$

where $\tilde{d}_j^* = n^{-1}\sum_{i=1}^n \tilde{d}_{ji}^2$, and $\tilde{d}_{(k)}^*$ is the kth largest value of $\tilde{d}_1^*,\ldots,\tilde{d}_p^*$.

This proposition is a groupwise version of Proposition 3 in [4]. The proof is given in the Appendix. From Proposition 1, we see that (6) yields a sparse solution by a hard-thresholding rule. In addition, (6) suggests that the KNOVAS first ranks the importance of covariates in a decreasing order according to the current estimates of $E\{\beta_j(V)^2\}$, and then it filters out the covariates whose corresponding estimates are smaller than $\tilde{d}_{(k)}^*$. This procedure is similar to the nonparametric variable screening procedure [10], which fitted p marginal regressions of the response against each covariate separately. However, (6) is based on the joint effect estimates of candidate covariates, which makes our method differ from the marginal screening methods.

Based on Proposition 2.1 and (5), we propose an IGHT algorithm to solve (3). To be specific, define $D^{[l]}$ as the estimate obtained at the lth iteration. Let L be the maximum iterative number. Our proposed IGHT algorithm is summarized below.

Algorithm 1 IGHT algorithm

Step 1. Choose the initial value $D^{[0]} = 0$;
Step 2. For each $l \in \{0,1,\ldots,L\}$,
Step 2a. Choose an initial step size $t^{[l]}$;
Step 2b. Compute $D^{[l+1]}$ by replacing $\tilde{D}$ with $[D^{[l]} + t^{-1}\tilde{\mathcal{L}}(D^{[l]})]$ in (6);
Step 2c. Stop Step 2 while the following linear search criterion is met:

$$\mathcal{L}(D^{[l+1]}) \ge \mathcal{L}(D^{[l]}) + \frac{\sigma t^{[l]}}{2n}||D^{[l+1]}-D^{[l]}||_F^2, \tag{7}$$

where $\sigma \in (0,1)$. Otherwise, let $t^{[l]} \leftarrow 2t^{[l]}$, and return to Step 2b;
Step 3. Stop the algorithm if $||D^{[l+1]}-D^{[l]}||_F^2 < \epsilon||D^{[l]}||_F^2$; Otherwise, let $l \leftarrow l+1$.

The estimate $\hat{D}$ is obtained at the convergence. The IGHT algorithm involves no heavy-duty operations such as $[\ddot{\ell}_n\{\beta(V)\}]^{-1}$, and thus it is computationally efficient. Step 2c is a backtracking method to find one value of $t^{[l]}$ such that $\mathcal{L}(D)$ monotonically increases with steps. The details about the setups of tuning parameters can be found in Section 4.2. Next, we turn to the convergence properties of the IGHT algorithm. Assume that the function $\ell_n\{\beta(\cdot)\}$ is Lipschitz continuous:

$$\|\dot{\ell}_n\{\tilde{\beta}(v)\} - \dot{\ell}_n\{\beta(v)\}\|_2 \le \phi\|\tilde{\beta}(v)-\beta(v)\|_2,$$

where $\phi > 0$ is a constant free of v, which is satisfied if the largest eigenvalue of $\ddot{\ell}_n\{\beta(v)\}$ is uniformly bounded in v. The Lipschitz condition is used to guarantee the boundedness of the step size $t^{[l]}$. Let D^* be the true value of D. The convergence properties of the IGHT algorithm are given in the following theorem.

Theorem 2.1. *Let $\{D^{[l]}\}$ be the sequence generated by the IGHT algorithm. If $t^{[l]} \ge \phi/(1-\sigma)$, then*

(a) There exists a subsequence $\mathcal{S}$ such that $\{D^{[l]} : l \in \mathcal{S}\}$ is convergent;
(b) After L iterations, the sequence $\{D^{[l]}\}$ satisfies that

$$\min_{0\le l\le L} \frac{1}{n}||D^{[l+1]}-D^{[l]}||_F^2 \le \frac{t^*}{L}\{\mathcal{L}(D^*)-\mathcal{L}(D^{[0]})\},$$

where $t^* = 2(1-\sigma)/(\phi\sigma)$, and $\mathcal{L}(D^{[l]}) \to \mathcal{L}(D^*)$ as $l\to\infty$.

The proof of Theorem 2.1 can be found in the Appendix. The part (a) describes the asymptotic convergence property of the IGHT algorithm. The part (b) implies that for any $\varepsilon > 0$, there exists $L = O(1/\varepsilon)$ such that for some $1 \le l^* \le L$, $n^{-1}\|D^{[l^*+1]}-D^{[l^*]}\|_F^2 \le \varepsilon$. In other words, the IGHT algorithm stops in a finite number of steps. Let $s = \text{card}(M^*)$, where $\text{card}(A)$ is the cardinality of a set A. The next theorem gives an upper bound for the estimation error in each iteration.

Theorem 2.2. *If $s \le k$ and $\phi < t^{[l]} < \rho/(1-1/\sqrt{32})$, then*

$$\|D^{[l]}-D^*\|_F \le 2^{-l}\|D^{[0]}-D^*\|_F + \sqrt{8/\phi}\ \|\tilde{\mathcal{L}}(D^*)\|_F,$$

where ρ is defined in Condition (C4) of Section 3.

Theorem 2.2, combining with the part (a) of Theorem 2.1, implies that there exists at least one subsequence such that the difference between the limit point and D^* can be bounded by a scaled version of $\|\tilde{\mathcal{L}}(D^*)\|_F$. Moreover, if we take the initial value $D^{[0]} = 0$, then after at most $l = \log_2(\|D^*\|_F/\|\tilde{\mathcal{L}}(D^*)\|_F)$ iterations, $D^{[l]}$ satisfies that $\|D^{[l]}-D^*\|_F \le (1+\sqrt{8/\phi})\|\tilde{\mathcal{L}}(D^*)\|_F$. Thus, the estimation error can also be controlled by a scaled version of $\|\tilde{\mathcal{L}}(D^*)\|_F$ in a finite number of steps.

3. SURE SCREENING PROPERTY

Let $\hat{M}$ denote the submodel index set given by

$$\hat{M} = \left\{ 1 \le j \le p : \frac{1}{n} \sum_{i=1}^{n} \hat{\beta}_j(V_i)^2 \neq 0 \right\}.$$

Define the collections of the under-fitted models and the over-fitted models as

$$M_-^k = \{M : M^* \not\subset M, \ \text{card}(M) \le k\}, \ \text{and}$$
$$M_+^k = \{M : M^* \subset M, \ \text{card}(M) \le k\},$$

respectively. The following conditions are imposed to establish the sure screening properties.

(C1) The kernel function $K(\cdot)$ is a symmetric density with a compact support, and has bounded variation. In addition, $\int v^2 K(v)dv < \infty$, and $h = O(n^{-\gamma})$ with $1/6 < \gamma < 1/2$.

(C2) The density function of V, denoted by $f_V(v)$, is twice continuously differentiable and positively bounded away from 0 on its support $\mathcal{J}$. Moreover, $\beta^*(v)$ is twice continuously differentiable on $\mathcal{J}$.

(C3) $\inf_{t\in[0,\tau]} \inf_{\beta(v)\in\mathcal{C}} \inf_{v\in\mathcal{J}} s^{(0)}(t;\beta(v),v) > 0$, where $\mathcal{C} \subseteq R^p$ is a convex and compact set, $s^{(0)}(t;\beta(v),v) = f_V(v)E\{S_T(t;Z,v)Z|V = v\}$, and $S_T(t;Z,V)$ is the survival function of T conditional on the covariates Z and V.

Conditions (C1) and (C2) are mild conditions on the density function $f_V(\cdot)$ and the kernel function $K(\cdot)$, which are satisfied by most commonly used distributions and kernels. These two conditions also imply that $f_V(v)$ and $K(v)$ are uniformly bounded on their supports. Condition (C3) is a standard assumption in the context of survival analysis [12]. Let $\beta_M(v)$ be a subvector of $\beta(v)$ associated with the components in an arbitrary subset M of $\{1,\ldots,p\}$. The following conditions are required for the sure screening properties.

(C4) There exist ρ and $\delta > 0$ such that for sufficiently large n,

$$\inf_{v\in\mathcal{J}} \inf_{\beta_M(v)\in\mathcal{M}(v)} \lambda_{\min}[-\ddot{\ell}_n\{\beta_M(v)\}] > \rho,$$

where $\mathcal{M}(v) = \{\beta_M(v) : ||\beta_M(v) - \beta_M^*(v)||_2 < \delta\}$ with $M \in M_+^{2k}$, and $\lambda_{\min}[A]$ denotes the minimum eigenvalue of a matrix A.

(C5) There exist some positive constants m_0, m_1 and α such that for sufficiently large η,

$$P\{|Z_j| \ge \eta\} \le m_1 \exp\{-m_0\eta^\alpha\} \quad \text{for} \quad j = 1,\ldots,p.$$

Condition (C4) corresponds to the uniform uncertainty principle given by [5], which is a relatively mild condition used in the literature of high dimensional methods (e.g., [7, 33, 37]. Also note that Condition (C4) is dependent on the sparse level k, and the $2k$-restricted minimum eigenvalue can be much bigger than the minimum eigenvalue of $\lambda_{\min}[-\ddot{\ell}_n\{\beta(v)\}]$. Condition (C5) ensures the tails of covariates to be exponentially light. This condition holds for a variety of distributions, such as the normal distribution and the distributions with bounded support.

Theorem 3.1. *Suppose that Conditions (C1)–(C5) hold, and there exist some positive constants $\omega_1, \omega_2, \kappa_1$ and κ_2 such that*

$$\inf_{v\in\mathcal{J}} \min_{j\in M^*} |\beta_j^*(v)| \ge \omega_1 n^{-\kappa_1}, \tag{8}$$

and $s \le k < \omega_2 n^{\kappa_2}$. Then for some constants c_1 and $c_2 > 0$,

$$P\{M^* \subseteq \hat{M}\} \ge 1 - c_2 n^{2+\kappa_2} p^k \exp\left\{-c_1 n^{\frac{(1-2\gamma-2\kappa_1-\kappa_2)\alpha}{\alpha+2}}\right\}.$$

Remark 3.1. Condition (8) states that the minimum signal strength of relevant covariates is uniformly bounded away from zero, but the lower bound may converge to zero. Together with (C4), it confines an appropriate order of k that guarantees the identifiability of M^* over M for $\text{card}(M) = k$. To establish the sure screening property, [40] required the asymptotic stability and Lindeberg conditions, which may be difficult to verify in the ultrahigh-dimensional case. However, we do not impose these assumptions in Theorem 3.1.

Remark 3.2. Theorem 3.1 states that the proposed screening procedure satisfies the sure screening property [13]. The number of covariates is allowed to be

$$\log(p) = o(n^{\frac{(1-2\gamma-2\kappa_1-\kappa_2)\alpha}{\alpha+2}-\kappa_2})$$

for $\kappa_1 + (1+1/\alpha)\kappa_2 \in (0, 0.5-\gamma)$. Accordingly, p grows with the sample size n at an exponential rate. It is worth noting that the probability bound is dependent on the order of the bandwidth h. If h is selected in a reasonable range and satisfies Condition (C1), the larger the bandwidth is, the higher the dimensionality we can handle. It can be seen that the number of covariates can be taken as $p = o(\exp\{n^{2(0.3-\kappa_1-\kappa_2)}\})$ for $\kappa_1 + \kappa_2 \in (0, 0.3)$ with the optimal bandwidth $h = O(n^{-1/5})$ and bounded covariates ($\alpha = \infty$). Thus, Theorem 3.1 generalizes the results of [14] and [25] to the varying coefficient Cox model.

4. IMPLEMENT ISSUES

4.1 Selection of thresholding parameter

In practice, the thresholding parameter k plays a very important role in the KNOVAS. A large k will lead to a large number of false positives in the selected model, while a small k may prevent sure screening. By the conditions of Theorem 3.1, we know that k is the order of $O(n^{\kappa_2})$. One can take $k = [n^{4/5}/\log(n^{4/5})]$ as suggested in [25], where $[x]$ denotes the integer part of $x \ge 0$. This is a hard cutoff

rule that retains a fixed number of variables in the selected model. The best choice of k is the true model size s. It can be seen from (6) that determining s is equivalent to estimate $\tilde{d}^*_{(s)}$ or $\tilde{d}^*_{(s+1)}$. The latter can be regarded as the maximum signal among the irrelevance covariates in practice. In what follows, we adopt the ideas of [2] and [44] to determine $\tilde{d}^*_{(s+1)}$.

For the ith subject ($i = 1, \ldots, n$), we first generate q-dimensional auxiliary covariates W_i from a standard multinormal distribution, which are independent of the observed data. We then carry out the KNOVAS procedure based on the extended data $\{(X_i, \Delta_i, Z_i^*, V_i), i = 1, \ldots, n\}$, where $Z_i^* = (Z_i^T, W_i^T)^T$ are $(p+q)$-dimensional covariates. Since the last q covariates are unrelated to the survival time T, $\hat{\beta}_j(v)$ $(p+1 \le j \le p+q)$ obtained based on the extended data are estimates of zero. Define

$$d^* = \max_{p+1 \le j \le p+d} \frac{1}{n} \sum_{i=1}^{n} \hat{\beta}_j(V_i)^2,$$

which can be viewed as the minimum thresholding level that makes no false positives. By replacing $\tilde{d}^*_{(k)}$ with d^* in (6), we obtain

$$\hat{b}_j = \tilde{d}_j I(\tilde{d}^*_j > d^*), \quad 1 \le j \le p.$$

Note that d^* depends on the q auxiliary covariates. To stabilize the model selection results, we repeat the above procedure N times, and obtain N submodels. Then we choose a best fitting model from the N submodels by using some information criteria such as AIC or BIC. In the simulation studies below, we take $q = p$, $N = 5$, and use a BIC-type information criterion to determine the final model. More details about this information criterion can be found in Section 5.2.

4.2 Parameter settings of IGHT algorithm

A good setup for the step size t can greatly reduce the cost of the IGHT algorithm, and hence it is critical for the fast convergence of the algorithm. By Theorem 2.1, we know that only t is large enough that can guarantee $\mathcal{L}(D)$ to increase after each iteration. However, our empirical studies suggest that a larger value of t often leads to a slower convergence of the IGHT algorithm. We choose an initial $t^{[l]}$ in Step 2a by adopting the Barzilai-Borwein rule [3], which uses a diagonal matrix $tI_{p\times p}$ to approximate the Hessian matrix $-\ddot{\ell}_n\{\beta(V_i)\}$ at $\beta(V_i) = \hat{\beta}^{[l]}(V_i)$. Specifically, we choose t at the $(l+1)$th iteration of the IGHT algorithm as

$$t^{[l+1]} = \arg\min_t \frac{1}{n} \sum_{i=1}^{n} ||tx_i^{[l]} - y_i^{[l]}||_2^2 = \frac{\sum_{i=1}^n (x_i^{[l]})^T y_i^{[l]}}{\sum_{i=1}^n (x_i^{[l]})^T x_i^{[l]}},$$

where $x_i^{[l]} = \beta^{[l]}(V_i) - \beta^{[l-1]}(V_i)$, and $y_i^{[l]} = -[\dot{\ell}_n\{\beta^{[l]}(V_i)\} - \dot{\ell}_n\{\beta^{[l-1]}(V_i)\}]$. In addition, we set $\sigma = \epsilon = 10^{-5}$, and $L = 1000$ in the simulation studies below.

Remark 4.1. As shown in Theorem 2.1, when the step size $t^{[l]} \ge \phi/(1-\sigma)$, the IGHT algorithm converges to the optimal solution, and the estimate $\hat{D}$ is obtained at the convergence. In our simulation studies and real data analysis, the initial step size $t^{[l]}$ selected by the Barzilai-Borwein rule works well, and the algorithm always converges to the optimal solution.

5. NUMERICAL EXAMPLES

5.1 Monte Carlo simulation

We conduct simulation studies to illustrate the sure screening property of the proposed procedure. For comparison, we also consider three alternative methods: FAST [15], CR [28], and the spline-based method [40], denoted by YZLH. For the three methods, we also use the data-driven procedure presented in Section 4.1 to determine the thresholding level. The total number of covariates is taken to be $p = 500$ and 1000 for all examples. The censoring time follows an exponential distribution with mean μ, where μ is selected to yield censoring rates of 26% and 42% for different settings, respectively. The covariates $(V^*, Z^T)^T$ are generated from a multi-normal distribution with mean zero and covariance Σ, where $\Sigma = (\sigma_{ij})_{(p+1)\times(p+1)}$ and $\sigma_{ij} = \rho^{|i-j|}$. The exposure variable is taken as $V = \Phi(V^*)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Thus, V follows a uniform distribution $U(0,1)$ and is related to Z. The kernel function is chosen as $K(x) = 0.75(1-x^2)I(|x| \le 1)$, and the bandwidth is determined by the thumb-rule with $h = \varpi\hat{\sigma}_V n^{-1/5}$, where ϖ is a pre-specified constant and $\hat{\sigma}_V$ is the estimated standard error of V for each setting. We take $\varpi = 2$, which works well in the simulation studies. The results presented below are based on 500 replications with sample size $n = 200$.

We mainly consider the following criteria to assess the performances of these methods: (1) the proportion of submodels $\hat{M}$ that contain all important covariates, denoted by p_a; (2) the proportion of submodels $\hat{M}$ that contain Z_j, denoted by p_j; (3) the proportion of correct fitting, that is, $\hat{M} = M^*$, denoted by CF; (4) the average number of true positives (TP) and false positives (FP). A powerful screening procedure should guarantee that p_a, p_j and CF are close to one, TP is close to the number of covariates with nonzero coefficients, and FP is close to zero.

Example 1. The true model index sets are taken to be $M^* = \{10, 100, 200, 400, 500\}$ and $M^* = \{10, 100, 200, 800, 1000\}$ for $p = 500$ and $p = 1000$, respectively. The correlation parameter ρ is set to be 0.1, 0.5 and 0.9. Moreover, we consider the following three cases for $\beta(V)$:

Case I. $\beta_{10}(v) = 1.5$, $\beta_{100}(v) = 1$, $\beta_{200}(v) = -1$, $\beta_{800}(v) = 1.2$ and $\beta_{1000}(v) = -1.5$.

Case II. $\beta_{10}(v) = (v-2)^2$, $\beta_{100}(v) = -2I(v > 0.3)$, $\beta_{200}(v) = 3\sin(2\pi v)$, $\beta_{800}(v) = 3v$

Table 1. The selection results for Case I of Example 1

ρ	Method	Censoring rate of 26%									Censoring rate of 42%								
		p_{10}	p_{100}	p_{200}	p_{800}	p_{1000}	p_a	TP	FP	CF	p_{10}	p_{100}	p_{200}	p_{800}	p_{1000}	p_a	TP	FP	CF
							$p = 500$												
0.1	KNOVAS	1.00	1.00	1.00	1.00	1.00	1.00	5.00	3.84	0.16	1.00	1.00	1.00	1.00	1.00	1.00	5.00	3.91	0.20
	YZLH	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.61	0.12	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.66	0.10
	FAST	1.00	0.92	0.93	0.92	1.00	0.79	4.76	0.91	0.38	0.99	0.91	0.90	0.91	0.99	0.74	4.71	1.01	0.31
	CR	0.84	0.56	0.42	0.56	0.64	0.10	3.01	1.00	0.02	0.82	0.55	0.18	0.50	0.42	0.02	2.47	0.99	0.01
0.5	KNOVAS	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.25	0.12	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.38	0.15
	YZLH	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.78	0.08	1.00	1.00	0.99	1.00	1.00	1.00	4.99	5.04	0.07
	FAST	1.00	0.93	0.94	0.92	0.99	0.81	4.79	3.96	0.03	0.99	0.91	0.90	0.87	0.99	0.71	4.66	3.24	0.04
	CR	0.84	0.55	0.39	0.55	0.64	0.08	2.96	1.96	0.01	0.79	0.55	0.22	0.52	0.40	0.04	2.48	1.73	0.00
0.9	KNOVAS	1.00	0.98	0.99	0.99	1.00	0.97	4.97	17.56	0.00	0.99	0.97	0.98	0.98	1.00	0.93	4.93	16.51	0.00
	YZLH	1.00	0.81	0.76	0.93	0.99	0.56	4.49	24.40	0.00	0.99	0.76	0.70	0.92	0.99	0.50	4.36	23.20	0.00
	FAST	1.00	0.88	0.90	0.88	0.99	0.69	4.65	30.38	0.00	0.98	0.87	0.89	0.87	0.98	0.66	4.60	29.51	0.00
	CR	0.83	0.54	0.42	0.54	0.67	0.08	3.00	16.97	0.00	0.81	0.53	0.22	0.54	0.39	0.03	2.49	14.74	0.00
							$p = 1000$												
0.1	KNOVAS	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.01	0.18	1.00	1.00	1.00	1.00	1.00	1.00	5.00	3.48	0.19
	YZLH	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.77	0.12	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.85	0.09
	FAST	1.00	0.88	0.92	0.89	0.99	0.72	4.68	1.05	0.32	0.99	0.87	0.87	0.87	0.98	0.65	4.58	0.95	0.28
	CR	0.83	0.50	0.35	0.50	0.61	0.05	2.79	1.02	0.01	0.75	0.45	0.16	0.49	0.34	0.01	2.19	0.90	0.00
0.5	KNOVAS	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.71	0.17	1.00	1.00	1.00	1.00	1.00	1.00	5.00	3.74	0.17
	YZLH	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.78	0.09	1.00	0.99	1.00	1.00	1.00	0.99	4.99	5.04	0.07
	FAST	0.99	0.91	0.89	0.89	0.99	0.71	4.67	3.14	0.04	0.99	0.88	0.86	0.88	0.99	0.66	4.60	2.95	0.06
	CR	0.78	0.52	0.37	0.49	0.60	0.05	2.76	1.82	0.01	0.77	0.49	0.17	0.46	0.35	0.01	2.24	1.68	0.00
0.9	KNOVAS	1.00	0.99	1.00	0.99	1.00	0.98	4.98	16.81	0.00	1.00	0.98	0.99	0.98	1.00	0.95	4.95	15.96	0.00
	YZLH	0.99	0.79	0.77	0.93	0.99	0.57	4.47	23.60	0.00	0.99	0.75	0.71	0.92	0.99	0.51	4.36	22.00	0.00
	FAST	0.99	0.85	0.89	0.87	0.98	0.64	4.59	28.75	0.00	0.98	0.86	0.86	0.84	0.99	0.61	4.53	27.77	0.00
	CR	0.77	0.49	0.35	0.49	0.61	0.04	2.71	14.24	0.00	0.78	0.46	0.21	0.44	0.42	0.02	2.31	11.89	0.00

and $\beta_{1000}(v) = \exp(v)$.

Case III. $\beta_{10}(v) = 2v + 1,\ \beta_{100}(v) = 3\sin(2\pi v)$, $\beta_{200}(v) = 2,\ \beta_{800}(v) = \exp(v)$ and $\beta_{1000}(v) = 2$.

In Case I, all the nonzero coefficient functions are constants, and in Case II, the nonzero coefficient functions are truly varying over v. Case III is a more complicated example, which allows some covariates to have varying effects, but others not.

The simulation results are reported in Tables 1–3. We find that the proposed KNOVAS procedure performs well for all the situations considered here. Specifically, the proportions of p_j and p_a are close to one, and the values of TP are close to 5. Note that for Case II, the challenge is to identify the important covariate Z_{200}, since $\beta_{200}(V)$ has mean 0 when V is from a uniform distribution $U(0, 1)$. Table 2 indicates that the KNOVAS successfully identifies the covariate Z_{200}, and the values of TP are close to 5. These findings suggest that the KNOVAS with the data-driven thresholding method ensures the sure screening property, and can also handle the case where the coefficient functions are constants. We also observe that the KNOVAS performs stable in terms of p_a as ρ increases. This implies that our method can successfully identify the important covariates even when the covariates are highly correlated. In addition, the simulation results show that the proposed method still performs reasonably well for different p's and censoring rates.

Furthermore, compared with the YZLH's method, it can be seen from Tables 1–3 that when $\rho = 0.1$ and 0.5, the KNOVAS and YZLH's methods provide comparable results in terms of p_a, p_j and TP, but our method yields smaller values for FP. When $\rho = 0.9$, our method outperforms the YZLH's method in all the settings considered here. For example, under Case I with $p = 500$ and the censoring rate of 26%, the values of p_a are ranged from 0.50 to 0.57 for the YZLH's method, while the values are from 0.93 to 0.98 for the KNOVAS.

For the FAST and CR methods, as shown in Table 2 for Case II, the proportions of p_{200} and p_a are close to 0. Thus, the two methods fail to identify the covariate Z_{200}. This is because $E\{\beta_{200}(V)\} = 0$ for V following a uniform distribution $U(0, 1)$. When we treat $\beta_{200}(V)$ as a constant and apply the FAST and CR methods, Z_{200} is regarded as a false covariate, and thus is ranked behind. For Case III, similar results can be found for the covariate Z_{100}. In addition, in terms of p_j and p_a, our method has better overall performance compared to the marginal screening methods considered here, even though the coefficient

Table 2. The selection results for Case II of Example 1

ρ	Method	Censoring rate of 26%									Censoring rate of 42%								
		p_{10}	p_{100}	p_{200}	p_{800}	p_{1000}	p_a	TP	FP	CF	p_{10}	p_{100}	p_{200}	p_{800}	p_{1000}	p_a	TP	FP	CF
								$p = 500$											
0.1	KNOVAS	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.31	0.17	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.31	0.16
	YZLH	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.67	0.11	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.77	0.12
	FAST	1.00	0.72	0.00	0.79	0.94	0.00	3.45	1.13	0.00	1.00	0.65	0.00	0.73	0.90	0.00	3.28	0.95	0.00
	CR	0.94	0.16	0.00	0.43	0.61	0.00	2.13	0.95	0.00	0.91	0.05	0.01	0.41	0.57	0.00	1.96	0.96	0.00
0.5	KNOVAS	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.69	0.14	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.45	0.13
	YZLH	1.00	1.00	1.00	1.00	1.00	1.00	5.00	5.17	0.07	1.00	1.00	1.00	1.00	1.00	1.00	5.00	5.39	0.05
	FAST	1.00	0.72	0.01	0.79	0.91	0.01	3.43	2.86	0.00	1.00	0.69	0.00	0.75	0.89	0.00	3.33	2.66	0.00
	CR	0.93	0.17	0.01	0.42	0.58	0.00	2.11	1.81	0.00	0.93	0.04	0.00	0.42	0.58	0.00	1.97	1.58	0.00
0.9	KNOVAS	1.00	0.99	0.99	0.99	1.00	0.97	4.97	17.51	0.00	1.00	0.97	0.99	0.97	0.99	0.94	4.94	17.17	0.00
	YZLH	1.00	0.82	0.98	0.87	0.85	0.57	4.52	25.40	0.00	1.00	0.81	0.98	0.85	0.87	0.60	4.51	24.80	0.00
	FAST	1.00	0.74	0.00	0.79	0.91	0.00	3.44	25.06	0.00	1.00	0.67	0.00	0.76	0.86	0.00	3.30	23.99	0.00
	CR	0.95	0.15	0.00	0.46	0.60	0.00	2.17	14.38	0.00	0.94	0.06	0.00	0.42	0.56	0.00	1.98	13.19	0.00
								$p = 1000$											
0.1	KNOVA	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.24	0.15	1.00	1.00	1.00	0.99	1.00	0.99	4.99	3.32	0.19
	YZLH	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.68	0.12	1.00	1.00	1.00	1.00	1.00	1.00	5.00	5.16	0.11
	FAST	1.00	0.64	0.01	0.76	0.88	0.01	3.29	1.05	0.00	1.00	0.63	0.00	0.71	0.88	0.00	3.22	1.01	0.00
	CR	0.93	0.13	0.00	0.40	0.56	0.00	2.02	1.12	0.00	0.92	0.03	0.00	0.36	0.54	0.00	1.86	0.91	0.00
0.5	KNOVA	1.00	1.00	1.00	1.00	1.00	1.00	5.00	5.04	0.12	1.00	1.00	1.00	1.00	1.00	1.00	4.99	3.73	0.16
	YZLH	1.00	1.00	1.00	1.00	1.00	1.00	5.00	5.25	0.11	1.00	1.00	1.00	1.00	1.00	1.00	5.00	5.27	0.11
	FAST	1.00	0.64	0.00	0.76	0.89	0.00	3.29	2.34	0.00	1.00	0.63	0.00	0.68	0.87	0.00	3.18	2.52	0.00
	CR	0.91	0.14	0.00	0.36	0.54	0.00	1.95	1.52	0.00	0.90	0.06	0.00	0.33	0.52	0.00	1.81	1.55	0.00
0.9	KNOVA	1.00	0.98	1.00	0.97	1.00	0.95	4.95	17.13	0.00	1.00	0.96	0.99	0.98	0.99	0.91	4.91	15.90	0.00
	YZLH	1.00	0.85	0.98	0.83	0.90	0.60	4.56	24.70	0.00	1.00	0.83	0.97	0.83	0.85	0.58	4.48	24.30	0.00
	FAST	1.00	0.64	0.00	0.71	0.90	0.00	3.25	21.73	0.00	1.00	0.58	0.00	0.70	0.88	0.00	3.16	21.77	0.00
	CR	0.94	0.13	0.01	0.36	0.53	0.00	1.97	11.39	0.00	0.90	0.04	0.00	0.38	0.46	0.00	1.77	10.85	0.00

functions are indeed constants. This is not surprising because the KNOVAS employs the joint effects of candidate covariates in the screening procedure, and has a good potential to outperform the marginal screening methods. It is also worth pointing out that all methods tend to be conservative in model selection, since all of them have high false positives, especially when $\rho = 0.9$. The main reason is that some irrelevant covariates will be included in the selected model due to their strong association with the relevant ones.

5.2 Variable selection stage

As pointed out in Section 5.1, there are still many irrelevant covariates retained in $\hat{M}$. Thus, a penalized procedure is needed to further recover the final sparse model. For this, let $\beta_*(V_i)$ be the subvector of $\beta(V_i)$ defined by $\beta_*(V_i) = \{\beta_j(V_i), j \in \hat{M}\}$ for each $1 \le i \le n$. Further define $D_* = (\beta_*(V_1), \ldots, \beta_*(V_n)) = (d_{1*}, \ldots, d_{k*})^T \in R^{k\times n}$. Then we can obtain an estimator of D_* by

$$\hat{D}_*(\lambda) = \arg\min_{D_* \in R^{k\times n}} \left\{ -\mathcal{L}(D_*) + \lambda \sum_{j\in\hat{M}} w_j ||d_{j*}||_2 \right\}, \tag{9}$$

where λ is a tuning parameter, and w_j are data-driven weights. We can solve the minimization problem (9) via the local quadratic approximation [11], and choose the tuning parameter λ by minimizing an extended Bayesian information criterion (EBIC):

$$\mathrm{EBIC}(\lambda) = -\mathcal{L}\{\hat{D}_*(\lambda)\} + df_\lambda \frac{\log(nh)}{nh} C_n,$$

where df_λ is the number of nonzero coefficients, C_n is a positive constant that diverges to infinity as the sample size n increases. Here we use nh but not n because in the kernel regression setting, the effective sample size is nh rather than n. The EBIC has been studied for the linear regression and quantile regression models [6, 22, 34]. As suggested in [22], we use $C_n = \log(p)/3$ in the following simulation study.

Example 2. We consider a more challenging model, where the true model index set is taken to be $M^* = \{1, 2, 3, 4, 5, 6\}$ for $p = 500$ and 1000. The covariates $(V^*, Z^T)^T$ are generated from a multi-normal distribution with mean zero and covariance Σ, where $\Sigma = (\sigma_{ij})_{(p+1)\times(p+1)}$ with $\sigma_{ij} = 0.4^{|i-j|}$ for $i, j \ne 7$, $\sigma_{7j} = \sigma_{j7} = 0$ for $j \ne 7$, and $\sigma_{77} = 1$. The exposure variable V is generated as $V = \Phi(V^*)$. The nonzero coefficient functions are given by

$$\beta_1(v) = 2 + \cos\{\pi(6v - 5)/3\},\ \beta_2(v) = 3 - 3v,$$

$$\beta_3(v) = -2 + 0.25(2 - 3v)^3,\ \beta_4(v) = \sin(9v^2/2) + 1,$$

Table 3. The selection results for Case III of Example 1

		Censoring rate of 26%									Censoring rate of 42%								
ρ	Method	p_{10}	p_{100}	p_{200}	p_{800}	p_{1000}	p_a	TP	FP	CF	p_{10}	p_{100}	p_{200}	p_{800}	p_{1000}	p_a	TP	FP	CF
							$p = 500$												
0.1	KNOVAS	1.00	1.00	1.00	1.00	1.00	1.00	5.00	3.79	0.15	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.14	0.15
	YZLH	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.53	0.12	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.29	0.13
	FAST	0.99	0.01	0.98	0.91	0.97	0.01	3.86	0.87	0.00	0.97	0.01	0.99	0.87	0.99	0.01	3.82	1.02	0.00
	CR	0.78	0.00	0.82	0.63	0.81	0.00	3.04	1.02	0.00	0.75	0.01	0.78	0.53	0.77	0.00	2.83	0.99	0.00
0.5	KNOVAS	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.58	0.14	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.68	0.15
	YZLH	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.87	0.08	1.00	1.00	1.00	1.00	1.00	1.00	5.00	5.24	0.08
	FAST	0.99	0.00	0.99	0.91	0.98	0.00	3.86	3.28	0.00	0.98	0.00	0.98	0.89	0.99	0.00	3.84	2.95	0.00
	CR	0.76	0.00	0.81	0.59	0.83	0.00	2.99	2.07	0.00	0.73	0.00	0.79	0.60	0.77	0.00	2.89	1.81	0.00
0.9	KNOVAS	1.00	1.00	1.00	0.99	1.00	0.99	4.98	17.98	0.00	1.00	1.00	1.00	0.99	1.00	0.98	4.98	18.18	0.00
	YZLH	0.97	0.97	0.98	0.86	0.96	0.76	4.74	24.50	0.00	0.97	0.97	0.95	0.82	0.95	0.67	4.66	24.00	0.00
	FAST	0.98	0.00	0.99	0.90	1.00	0.00	3.87	28.46	0.00	0.98	0.00	0.98	0.87	0.97	0.00	3.80	27.20	0.00
	CR	0.81	0.00	0.77	0.60	0.81	0.00	2.99	17.55	0.00	0.77	0.00	0.73	0.55	0.76	0.00	2.81	16.12	0.00
							$p = 1000$												
0.1	KNOVAS	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.12	0.16	1.00	1.00	1.00	1.00	1.00	1.00	5.00	3.23	0.20
	YZLH	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.71	0.12	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.95	0.09
	FAST	0.97	0.00	0.99	0.89	0.98	0.00	3.82	1.09	0.00	0.97	0.00	0.97	0.84	0.98	0.00	3.75	0.91	0.00
	CR	0.74	0.00	0.75	0.52	0.76	0.00	2.77	0.98	0.00	0.71	0.00	0.75	0.50	0.71	0.00	2.67	0.87	0.00
0.5	KNOVAS	1.00	1.00	1.00	1.00	1.00	1.00	5.00	5.07	0.12	1.00	1.00	1.00	1.00	1.00	1.00	5.00	3.83	0.15
	YZLH	1.00	1.00	1.00	1.00	1.00	1.00	5.00	4.83	0.07	1.00	1.00	1.00	1.00	1.00	1.00	5.00	5.30	0.10
	FAST	0.98	0.00	0.98	0.88	0.98	0.00	3.82	2.68	0.00	0.96	0.00	0.96	0.87	0.97	0.00	3.77	2.53	0.00
	CR	0.72	0.00	0.72	0.51	0.75	0.00	2.71	1.68	0.00	0.67	0.00	0.72	0.49	0.72	0.00	2.60	1.74	0.00
0.9	KNOVAS	0.99	0.99	1.00	0.99	1.00	0.98	4.98	17.41	0.00	0.99	1.00	0.99	0.98	1.00	0.96	4.96	16.59	0.00
	YZLH	0.97	0.99	0.99	0.87	0.98	0.81	4.80	24.00	0.00	0.95	0.97	0.97	0.85	0.97	0.73	4.71	23.40	0.00
	FAST	0.97	0.00	0.98	0.84	0.99	0.00	3.78	26.06	0.00	0.95	0.00	0.98	0.83	0.95	0.00	3.71	25.16	0.00
	CR	0.72	0.00	0.73	0.51	0.74	0.00	2.70	13.97	0.00	0.70	0.00	0.72	0.51	0.73	0.00	2.65	14.58	0.00

$\beta_5(v) = \exp\{3v^2/(3v^2+1)\}$ and $\beta_6(v) = 1$.

This setup is also considered in [25] with $\beta_6(v) \equiv 0$. Note that the covariate Z_3 is marginally unrelated to T but is an important covariate. In addition, the signal strength of the covariate Z_6 is relatively weak compared to others, and its marginal signal strength is not enhanced by the correlations between covariates. The other setups are the same as in Section 5.1. The simulation results are reported in Table 4, in which the row labeled KNOVAS* refers to the KNOVAS method with the penalized stage. The results suggest that the KNOVAS* can significantly reduce the false positives after the screening procedure, and further recover the final sparse model. For example, when $p = 1000$ and the censoring rate is 42%, the FP of the KNOVAS* (= 0.47) is smaller than that of the KNOVAS (= 3.81), and the CF increases from 0.16 (KNOVAS) to 0.64 (KNOVAS*).

5.3 Ovarian cancer data

For illustration purposes, we apply the KNOVAS method to an ovarian cancer dataset. This dataset is from the Cancer Genome Atlas [30], which includes 486 patients and 11864 probe sets. The response of interest is the time to progression, and our analysis is restricted to 388 patients with observed responses. The follow-up time is $\tau = 110.75$ (in months), and 191 patients are censored. We take the patient's age as the exposure variable V, which is ranged from 30 to 87 years at diagnosis. The gene expression values are taken as the covariates. The bandwidth is selected as $h = 2\hat{\sigma}_V n^{-1/5}$, where $\hat{\sigma}_V$ is the estimated standard error of V. The covariates are normalized to have mean zero and variance one. The raw data are available at the TCGA website (https://tcga-data.nci.nih.gov/docs/publications/ov_2011/).

Using the KNOVAS with the data-driven thresholding method, we find five genes, named as CD79A, MUC13, PLA2G2D, PUF60 and TBX2, with strong effects on the progression-free survival time. Figure 1 depicts the estimated coefficient functions of the selected 5 genes. This means that their effects are nonlinear, and vary with the age. For example, the effects of PLA2G2D and PUF60 show a similar parabola shape: a negative impact on the progression-free survival time; TBX2 has a decreasing pattern, changing from positive to negative values as the age increases.

6. CONCLUSION

For the sparse varying coefficient Cox model, we introduced a non-marginal variable screening method that combines the kernel smoothing and group learning methods. An IGHT algorithm was developed for fast implementation

Table 4. The selection results for Example 2

Method	p_1	p_2	p_3	p_4	p_5	p_6	p_a	TP	FP	CF	p_1	p_2	p_3	p_4	p_5	p_6	p_a	TP	FP	CF
	Censoring rate of 26%										Censoring rate of 42%									
$p = 500$																				
KNOVAS*	1.00	1.00	1.00	1.00	1.00	0.99	0.99	5.99	0.31	0.72	1.00	1.00	1.00	1.00	1.00	0.99	0.99	5.98	0.44	0.67
KNOVAS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	6.00	3.39	0.19	1.00	1.00	1.00	1.00	1.00	0.99	0.99	5.99	3.42	0.19
YZLH	1.00	1.00	1.00	1.00	1.00	1.00	1.00	6.00	5.28	0.10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	6.00	5.81	0.06
FAST	1.00	1.00	0.00	0.93	0.99	0.00	0.00	4.31	0.98	0.00	1.00	1.00	0.00	0.93	0.99	0.41	0.00	4.33	1.12	0.00
CR	0.99	0.87	0.00	0.69	0.84	0.19	0.00	3.58	0.89	0.00	0.96	0.83	0.00	0.61	0.79	0.21	0.00	3.41	1.00	0.00
$p = 1000$																				
KNOVAS*	1.00	1.00	1.00	1.00	1.00	1.00	1.00	6.00	0.42	0.65	0.99	0.99	0.99	0.98	0.99	0.96	0.96	5.92	0.47	0.64
KNOVAS	1.00	1.00	1.00	1.00	1.00	1.00	1.00	6.00	3.37	0.17	1.00	1.00	1.00	0.99	1.00	0.99	0.98	5.97	3.81	0.16
YZLH	1.00	1.00	0.99	0.99	1.00	0.99	0.98	5.97	5.06	0.07	1.00	1.00	0.99	0.99	1.00	0.99	0.98	5.96	5.06	0.07
FAST	1.00	1.00	0.00	0.94	1.00	0.35	0.00	4.29	0.94	0.00	1.00	1.00	0.00	0.92	0.99	0.00	0.34	4.24	1.09	0.00
CR	0.97	0.81	0.00	0.63	0.83	0.14	0.00	3.37	1.07	0.00	0.97	0.81	0.00	0.63	0.81	0.00	0.12	3.35	1.01	0.00

of the proposed method. We established the convergence properties of the IGHT algorithm, and the sure screening properties of the variable screening procedure. The simulation studies showed that the proposed method performed well in the following cases: (i) the covariates are ultrahigh-dimensional and highly correlated; (ii) the covariate effects vary with an exposure variable; and (iii) the covariates are marginally unrelated but jointly related to the response.

For computationally feasibility, we used the thumb-rule to determine the bandwidth. Further research is needed to develop some data-based methods for the selection of the bandwidth. In addition, it would be interested to generalize the proposed method to other regression models, such as proportional odds models with varying coefficients, varying coefficient transformation models and nonparametric proportional hazards models [8].

APPENDIX A. PROOFS

Proof of Proposition 2.1. Let B be an optimal solution and $M = \{j : ||b_j||_2 \neq 0\}$. Then the objective function can be rewritten as

$$||B - \tilde{D}||_F^2 = \sum_{j \in M} ||b_j - \tilde{d}_j||_2^2 + \sum_{j \notin M} ||\tilde{d}_j||_2^2.$$

By taking $b_j = \tilde{d}_j$ for $j \in M$, we have $||B - \tilde{D}||_F^2 = \sum_{j \notin M} ||\tilde{d}_j||_2^2$. Thus, the objective function achieves the minimum if and only if M corresponds to the indices of the largest k values of $\tilde{d}_j^*$. □

Proof of Theorem 2.1(a). We first show the convergence of $\{\mathcal{L}(D^{[l]})\}$. Note that the linear criteria (7) ensures that the value of $\mathcal{L}(D^{[l]})$ is monotonically increasing after each iteration. Thus, it suffices to show that (7) holds because of the boundedness of $\mathcal{L}(D)$. By the Lipschitz continuous condition of $\ell_n\{\beta(\cdot)\}$, we have that for any $t \geq \phi$,

(10)
$$\mathcal{L}(B) \geq \mathcal{L}(D) + \frac{1}{n} \operatorname{tr}\{(B - D)^T \tilde{\mathcal{L}}(D)\} - \frac{t}{2n} \|B - D\|_F^2.$$

By the definitions of $\mathcal{Q}_t(B|D)$ and $D^{[l+1]}$, we have

(11)
$$\begin{aligned}
\mathcal{L}(D^{[l]}) =& \mathcal{Q}_{t^{[l]}}(D^{[l]}|D^{[l]}) \leq \mathcal{Q}_{t^{[l]}}(D^{[l+1]}|D^{[l]}) \\
=& \mathcal{L}(D^{[l]}) + \frac{1}{n} \operatorname{tr}\{(D^{[l+1]} - D^{[l]})^T \tilde{\mathcal{L}}(D^{[l]})\} \\
& - \frac{t^{[l]}}{2n} \|D^{[l+1]} - D^{[l]}\|_F^2 \\
=& \mathcal{L}(D^{[l]}) + \frac{1}{n} \operatorname{tr}\{(D^{[l+1]} - D^{[l]})^T \tilde{\mathcal{L}}(D^{[l]})\} \\
& - \frac{t^{[l]}}{2n} \|D^{[l+1]} - D^{[l]}\|_F^2 - \frac{t^{[l]} - \phi}{2n} \|D^{[l+1]} - D^{[l]}\|_F^2 \\
=& \mathcal{Q}_{\phi}(D^{[l+1]}|D^{[l]}) - \frac{t^{[l]} - \phi}{2n} \|D^{[l+1]} - D^{[l]}\|_F^2.
\end{aligned}$$

Then it follows from (10) and (11) that

(12)
$$\mathcal{L}(D^{[l]}) \leq \mathcal{L}(D^{[l+1]}) - \frac{t^{[l]} - \phi}{2n} \|D^{[l+1]} - D^{[l]}\|_F^2.$$

Hence the monotone line search criterion (7) is satisfied whenever $t^{[l]} \geq \phi/(1 - \sigma)$. (12), together with the boundedness of $\mathcal{L}(D)$, implies that $\{\mathcal{L}(D^{[l]})\}$ has at least one limiting point in the feasible region.

Next, we show that there exists a subsequence such that $D^{[l]}$ is convergent. When $t^{[l]} \to \infty$ as l goes to infinity, the result is trivial. In what follows, we assume that $\{t^{[l]}\}$ is bounded. Thus, there exists a subsequence $\mathcal{S}$ such that $t^{[l]} \to \tilde{t}$ for $l \in \mathcal{S}$. For each $t^{[l]} \in \mathcal{S}$, let $\mathcal{M}^{[l]} = \{j : n^{-1} \sum_{i=1}^n \beta_j^{[l]}(V_i) \neq 0\}$. By (12) and the fact that $\mathcal{L}(D^{[l]})$ is convergent, we know that $\|D^{[l+1]} - D^{[l]}\|_F^2 \to 0$ as $l \in \mathcal{S}$ goes to infinity, which implies that $\mathcal{M}^{[l]}$ is also convergent. Since $\mathcal{M}^{[l]}$ is a discrete sequence, there exists an $l^* \in \mathcal{S}$ such that $\mathcal{M}^{[l]} = \mathcal{M}^{[l^*]}$ for all $l \in \mathcal{S}$ and $l \geq l^*$. Thus, the IGHT algorithm is a gradient descent algorithm on the space $\mathcal{M}^{[l]}$ for all $l \in \mathcal{S}$ and $l \geq l^*$. Since a gradient descent algorithm for minimizing a convex function over a closed convex set

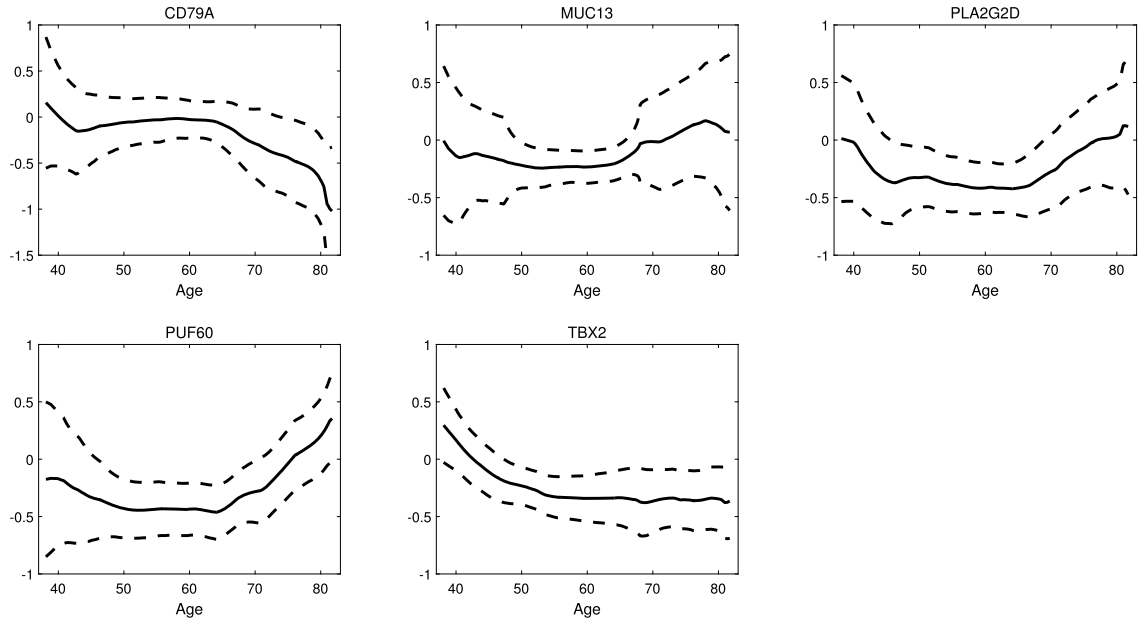

Figure 1. Estimated varying coefficients of the selected 5 genes. The solid lines are the estimated curves, and the dashed lines are the 95% pointwise confidence intervals.

yields a sequence of iterations that converges [26], we conclude that the subsequence $\{D^{[l]} : l \in \mathcal{S}\}$ is convergent, and the proof is completed. □

Proof of Theorem 2.1(b). It follows from (12) and $t^{[l]} \geq \phi/(1-\sigma)$ that

$$\sum_{l=0}^{L}\{\mathcal{L}(D^{[l]}) - \mathcal{L}(D^{[l+1]})\} \leq \frac{\sigma\phi}{2(\sigma-1)n}\sum_{l=0}^{L}\|D^{[l+1]} - D^{[l]}\|_F^2,$$

which implies that

$$\min_{0\leq l\leq L}\frac{1}{n}\|D^{[l+1]} - D^{[l]}\|_F^2 \leq \frac{2(1-\sigma)}{\sigma\phi L}\{\mathcal{L}(D^{[L+1]}) - \mathcal{L}(D^{[0]})\}.$$

Let $t^* = 2(1-\sigma)/(\sigma\phi)$. In view of Theorem 2.1(a), we obtain that the nondecreasing sequence $\mathcal{L}(D^{[l]})$ has at least one limiting point, denoted by $\mathcal{L}(D^*)$. Thus, we get

$$\begin{aligned}\min_{0\leq l\leq L}\frac{1}{n}\|D^{[l+1]} - D^{[l]}\|_F^2 \leq& \frac{t^*}{L}\{\mathcal{L}(D^{[L+1]}) - \mathcal{L}(D^{[0]})\}\\ \leq& \frac{t^*}{L}\{\mathcal{L}(D^*) - \mathcal{L}(D^{[0]})\}.\end{aligned}$$

This completes the proof of Theorem 2.1(b). □

Proof of Theorem 2.2. It can be checked that

$$\|D^{[l+1]} - D^*\|_F \leq 2\|D^{[l]} - D^* + (t^{[l]})^{-1}\tilde{\mathcal{L}}(D^{[l]})\|_F.$$

Then by the Taylor's expansion, we have

$$\begin{aligned}&\|D^{[l]} - D^* - (t^{[l]})^{-1}\tilde{\mathcal{L}}(D^{[l]})\|_F\\ =&\Big[\sum_{i=1}^{n}\|\beta^{[l]}(V_i) - \beta^*(V_i) + \frac{1}{t^{[l]}}\dot{\ell}_n\{\beta^*(V_i)\}\\ &+ \frac{1}{t^{[l]}}\ddot{\ell}_n\{\tilde{\beta}(V_i)\}\{\beta^{[l]}(V_i) - \beta^*(V_i)\}\|_2^2\Big]^{1/2}\\ \leq&\sqrt{2}\Big[\sum_{i=1}^{n}\Big\{\|[I + \frac{1}{t^{[l]}}\ddot{\ell}_n\{\tilde{\beta}(V_i)\}]\{\beta^{[l]}(V_i) - \beta^*(V_i)\}\|_2^2\\ &+ \frac{1}{t^{[l]}}\|\dot{\ell}_n\{\beta^*(V_i)\}\|_2^2\Big\}\Big]^{1/2}\\ \leq&\sqrt{2}\Big[\sum_{i=1}^{n}\|[I + \frac{1}{t^{[l]}}\ddot{\ell}_n\{\tilde{\beta}(V_i)\}]\{\beta^{[l]}(V_i) - \beta^*(V_i)\}\|_2^2\Big]^{1/2}\\ &+ \sqrt{\frac{2}{t^{[l]}}}\Big[\sum_{i=1}^{n}\|\dot{\ell}_n\{\beta^*(V_i)\}\|_2^2\Big]^{1/2},\end{aligned}$$

where $\tilde{\beta}(V_i)$ lies between $\beta^{[l]}(V_i)$ and $\beta^*(V_i)$. Thus, by Condition (C4) with $\phi < t^{[l]} < \rho/(1-1/\sqrt{32})$, we obtain

$$\|D^{[l+1]} - D^*\|_F \leq 2^{-1}\|D^{[l]} - D^*\|_F + \sqrt{\frac{2}{\phi}}\|\tilde{\mathcal{L}}(D^*)\|_F.$$

By iterating this relationship, we get

$$\|D^{[l]} - D^*\|_F \leq 2^{-l}\|D^{[0]} - D^*\|_F + \sqrt{\frac{8}{\phi}}\|\tilde{\mathcal{L}}(D^*)\|_F.$$

This completes the proof of Theorem 2.2. □

To show Theorem 3.1, we need the following lemma, which gives a concentration inequality of the objective function $\dot{\ell}_{nj}\{\beta(v)\}$ for $1 \leq j \leq p$. Let $\dot{\ell}_{nj}(\beta(v), v) = \dot{\ell}_{nj}\{\beta(v)\}$.

Lemma A.1. *Under Conditions (C1)–(C3), there exist positive constants k_0, k_1 and k_2 independent of n such that for a large $\eta > 0$ and $1 \le j \le p$,*

$$\begin{aligned}&\sup_{\beta(v)\in\mathcal{C}}\sup_{v\in\mathcal{J}} P\Big\{\big|\dot{\ell}_{nj}(\beta(v),v)\big| > k_0(nh^2)^{-1/2}(1+x)\Big\}\\ \le& k_1\exp\{-k_1x^2/(2\eta^2)\} + k_2\exp\{-k_2nh^2/(2\eta^2)\}\\ &+ nP\{|Z_{ij}| > \eta\}.\end{aligned}$$

The proof of Lemma A.1 is tedious, and the details can be found in the online Supplemental Material, http://intlpress.com/site/pub/files/_supp/sii/2021/0014/0002/SII-2021-0014-0002-s002.pdf.

Proof of Theorem 3.1. It suffices to show that

$$\begin{aligned}&P\Big\{\max_{M\in M_-^k}\mathcal{L}(D_M) \ge \min_{M\in M_+^k}\mathcal{L}(D_M)\Big\}\\ \le& c_2n^{2+\kappa_2}p^k\exp\Big\{-c_1n^{\frac{(1-2\gamma-2\kappa_1-\kappa_2)\alpha}{\alpha+2}}\Big\}.\end{aligned}$$

For any $M \in M_-^k$, let $M' = M \cup M^* \in M_+^{2k}$. Consider D_M close to D_M^* such that $\|\beta_{M'}(V_i) - \beta^*_{M'}(V_i)\|_2 = \omega_1 n^{-\kappa_1}$ for some $\omega_1, \kappa_1 > 0$. It follows from Condition (C4) and the Taylor's expansion that

$$\begin{aligned}&\mathcal{L}(D_M) - \mathcal{L}(D_M^*)\\ =&\frac{1}{n}\sum_{i=1}^n\Big\{\ell_n(\beta_{i,M'}) - \ell_n(\beta^*_{i,M'})\Big\}\\ =&\frac{1}{n}\sum_{i=1}^n\Big\{\dot{\ell}_n(\beta^*_{i,M'})^T(\beta_{i,M'} - \beta^*_{i,M'})\\ &+\frac{1}{2}(\beta_{i,M'} - \beta^*_{i,M'})^T\ddot{\ell}_n(\tilde{\beta}_{i,M'})(\beta_{i,M'} - \beta^*_{i,M'})\Big\}\\ \le&\frac{1}{n}\sum_{i=1}^n\Big\{||\dot{\ell}_n(\beta^*_{i,M'})||_2||\beta_{i,M'} - \beta^*_{i,M'}||_2\\ &-\frac{\rho}{2}||\beta_{i,M'} - \beta^*_{i,M'}||_2^2\Big\}\\ \le&\frac{1}{n}\sum_{i=1}^n\Big\{\omega_1n^{-\kappa_1}||\dot{\ell}_n(\beta^*_{i,M'})||_2 - \frac{\rho}{2}\omega_1^2n^{-2\kappa_1}\Big\},\end{aligned}$$

where $\tilde{\beta}_{i,M'}$ is an intermediate value between $\beta_{i,M'}$ and $\beta^*_{i,M'}$. Thus, we obtain

$$\begin{aligned}(13)\qquad &P\{\mathcal{L}(D_M) - \mathcal{L}(D_M^*) \ge 0\}\\ \le&P\Big\{\frac{1}{n}\sum_{i=1}^n||\dot{\ell}_n(\beta^*_{i,M'})||_2 \ge (\frac{\rho\omega_1}{2})n^{-\kappa_1}\Big\}\\ \le&\sum_{i=1}^n\sum_{j\in M'}P\Big\{|\dot{\ell}_{nj}(\beta^*_{i,M'})| \ge \frac{\rho\omega_1}{2}(2k)^{-1/2}n^{-\kappa_1}\Big\}.\end{aligned}$$

Also note that

$$\begin{aligned}&P\Big\{|\dot{\ell}_{nj}(\beta^*_{i,M'})| \ge \frac{\rho\omega_1}{2}(2k)^{-1/2}n^{-\kappa_1}\Big\}\\ \le&P\Big\{|\dot{\ell}_{nj}(\beta^*_{i,M'})| \ge \frac{\rho\omega_1}{2\sqrt{2\omega_2}}n^{-\kappa_1-0.5\kappa_2}, \max_{1\le i\le n}|Z_{ij}| \le \eta\Big\}\\ &+P\Big\{\max_{1\le i\le n}|Z_{ij}| > \eta\Big\}.\end{aligned}$$

By Lemma A.1, Condition (C5) and taking $\eta = n^{(1-2\gamma-2\kappa_1-\kappa_2)/(\alpha+2)}$, there exist c_0 and $c_1 > 0$ such that

$$\begin{aligned}(14)\qquad &P\Big\{|\dot{\ell}_{nj}(\beta^*_{i,M'})| \ge (\frac{\rho\omega_1}{2})(2k)^{-1/2}n^{-\kappa_1}\Big\}\\ \le&c_0n\exp\Big\{-c_1n^{\frac{(1-2\gamma-2\kappa_1-\kappa_2)\alpha}{\alpha+2}}\Big\}.\end{aligned}$$

The inequalities (13) and (14) imply that

$$P\{\mathcal{L}(D_M) \ge \mathcal{L}(D_M^*)\} \le c_0n^2k\exp\Big\{-c_1n^{\frac{(1-2\gamma-2\kappa_1-\kappa_2)\alpha}{\alpha+2}}\Big\}.$$

Therefore, the Bonferroni inequality yields that there is a constant $c_2 > 0$ such that

$$\begin{aligned}&P\Big\{\max_{M'\in M_-^k}\mathcal{L}(D_{M'}) \ge \mathcal{L}(D^*)\Big\}\\ \le&c_2n^{2+\kappa_2}p^k\exp\Big\{-c_1n^{\frac{(1-2\gamma-2\kappa_1-\kappa_2)\alpha}{\alpha+2}}\Big\}.\end{aligned}$$

By Condition (C4), we know that $\ddot{\ell}(\beta_{M'}(v))$ is positive definite for each $v \in \mathcal{J}$. This implies that the above result holds for any $\beta_{M'}(v)$ such that $\|\beta_{M'}(v) - \beta^*_{M'}(v)\|_2 \ge \omega_1n^{-\kappa_1}$. For any $M \in M_-^k$, let $\tilde{\beta}_{M'}(v)$ be $\beta_M(v)$ augmented with zeros corresponding to the elements in M'/M^*. Since $M' = \{M\cup(M^*/M)\}\cup\{M'/M^*\}$, it follows from Condition (C4) that $\|\tilde{\beta}_{M'}(v) - \beta^*_{M'}(v)\|_2 \ge \|\beta_{M'/M^*}(v)\|_2 \ge \omega_1n^{-\kappa_1}$. Thus, we have

$$\begin{aligned}&P\Big\{\max_{M'\in M_-^k}\mathcal{L}(D_M) \ge \min_{M'\in M_+^k}\mathcal{L}(D_M)\Big\}\\ \le&P\Big\{\max_{M'\in M_-^k}\mathcal{L}(\tilde{D}_M) \ge \mathcal{L}(D_M^*)\Big\}\\ \le&c_2n^{2+\kappa_2}p^k\exp\Big\{-c_1n^{\frac{(1-2\gamma-2\kappa_1-\kappa_2)\alpha}{\alpha+2}}\Big\}.\end{aligned}$$

The proof is completed. □

ACKNOWLEDGEMENTS

The authors would like to thank the Editor-in-Chief, Professor Yuedong Wang, an Associate Editor and two referees for their constructive and insightful comments and suggestions that greatly improve the article. This research was partly supported by the National Natural Science Foundation of China (Grant Nos. 11771431 and 11690015), Key Laboratory of RCSDS, CAS (No. 2008DP173182), and the Hubei Natural Science Foundation of China (Grant No. 2018CFB256).

Received 30 August 2019

REFERENCES

[1] Ahn, K., Sahr, N. and Kim, S. (2018). Screening group variables in the proportional hazards model. *Statist. Probab. Lett.* **135**, 20–25. MR3758256

[2] Barut, E., Fan, J. and Verhasselt, A. (2016). Conditional sure independence screening. *J. Amer. Statist. Assoc.* **111**, 1266–1277. MR3561948

[3] Barzilai, J. and Borwein, J. (1988). Two-point step size gradient methods. *IMA J. Numer. Anal.* **8**, 141–148. MR0967848

[4] Bertsimas, D., King, A. and Mazumder, R. (2016). Best subset selection via a modern optimization Lens. *Ann. Statist.* **44**, 813–852. MR3476618

[5] Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35**, 2313–2404. MR2382644

[6] Chen, J., Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759–771. MR2443189

[7] Chen, J., Chen, Z. (2012). Extended BIC for small-n-large-p sparse glm. *Statistica Sinica* **22**, 555–574. MR2954352

[8] Chen, K., Guo, S., Sun, L., Wang, J.-L. (2010). Global partial likelihood for nonparametric proportional hazards models. *J. Amer. Statist. Assoc.* **105**, 750–760. MR2724858

[9] Chu, W., Li, R., Reimherr, M. (2016). Feature screening for time-varying coefficient models with ultrahigh dimensional longitudinal data. *Ann. Appl. Stat.* **10**, 596–617. MR3528353

[10] Fan, J., Feng, Y., Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc.* **106**, 544–557. MR2847969

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

[12] Fan, J., Lin, H., Zhou, Y. (2006). Local partial-likelihood estimation for lifetime data. *Ann. Statist.* **34**, 290–325. MR2275243

[13] Fan, J., Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with Discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70**, 849–911. MR2530322

[14] Fan, J., Ma, Y., Dai, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *J. Amer. Statist. Assoc.* **109**, 1270–1284. MR3265696

[15] Gorst-Rasmussen, A., Scheike, T. (2013). Independent screening for single-index hazard rate models with ultrahigh dimensional features. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **75**, 217–245. MR3021386

[16] Honda, T., Härdle, W. (2014). Variable selection in Cox regression models with varying coefficients. *J. Statist. Plann. Inference* **148**, 67–81. MR3174148

[17] Honda, T., Yabe, R. (2017). Variable selection and structure identification for varying coefficient Cox models. *J. Multivariate Anal.* **161**, 103–122. MR3698118

[18] Hong, H., Chen, X., Christiani, D., Li, Y. (2018). Integrated powered density: screening ultrahigh dimensional covariates with survival outcomes. *Biometrics* **74**, 421–429. MR3825328

[19] Hong, H., Li, Y. (2017). Feature selection of ultrahigh-dimensional covariates with survival outcomes: a selective review. *Appl. Math. J. Chinese Univ. Ser. B* **32**, 379–396. MR3736443

[20] Hong, H., Kang, J., Li, Y. (2018). Conditional screening for ultra-high dimensional covariates with survival outcomes. *Lifetime Data Anal.* **24**, 45–71. MR3742906

[21] Hong, H., Zhang, Q., Li, Y. (2019). Forward regression for Cox models with high-dimensional covariates. *J. Multivariate Anal.* **173**, 268–290. MR3924485

[22] Lee, E., Noh, H., Park, B. (2014). Model selection via Bayesian information criterion for quantile regression models. *J. Amer. Statist. Assoc.* **109**, 216–229. MR3180558

[23] Li, J., Zheng, Q., Peng, L., Huang, Z. (2016). Survival impact index and ultrahigh-dimensional model-free screening with survival outcomes. *Biometrics* **72**, 1145–1154. MR3591599

[24] Lian, H., Lai, P., Liang, H. (2013). Partially linear structure selection in Cox models with varying coefficients. *Biometrics* **69**, 348–357. MR3071053

[25] Liu, J., Li, R., Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *J. Amer. Statist. Assoc.* **109**, 266–274. MR3180562

[26] Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course.* Applied Optimization 87. Kluwer Academic, Boston, MA.

[27] Qu, L., Song, X., Sun, L. (2018). Identification of local sparsity and variable selection for varying coefficient additive hazards models. *Comput. Statist. Data Anal.* **125**, 119–135. MR3800150

[28] Song, R., Lu, W., Ma, S., Jeng, J. (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika* **101**, 799–814. MR3286918

[29] Song, R., Yi, F., Zou, H. (2014). On varying-coefficient independence screening for high dimensional varying-coefficient models. *Statist. Sinica* **24**, 1735–1752. MR3308660

[30] The Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615.

[31] Tian, L., Zucker, D., Wei, L. (2005). On the Cox model with time-varying regression coefficients. *J. Amer. Statist. Assoc.* **100**, 172–183. MR2156827

[32] Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* **16**, 385–395.

[33] Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *J. Amer. Statist. Assoc.* **104**, 1512–1524. MR2750576

[34] Wang, H., Li, B., Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71**, 671–683. MR2749913

[35] Xia, X., Yang, H., Li, J. (2016) Feature screening for generalized varying-coefficient models with application to dichotomous response. *Comput. Statist. Data Anal.* **102**, 85–97. MR3506984

[36] Xiao, W., Lu, W., Zhang, H. (2016). Joint structure selection and estimation in the time-varying coefficient Cox model. *Statist. Sinica* **26**, 547–567. MR3497759

[37] Xu, C., Chen, J. (2014). The sparse MLE for ultrahigh-dimensional feature screening. *J. Amer. Statist. Assoc.* **109**, 1257–1269. MR3265695

[38] Yan, J., Huang, J. (2012). Model selection for Cox models with time-varying coefficients. *Biometrics* **68**, 419–428. MR2959608

[39] Yang, G., Yu, Y., Li, R., Buu, A. (2016). Feature screening in ultrahigh dimensional Cox's model. *Statist. Sinica* **26**, 881–901. MR3559935

[40] Yang, G., Zhang, L., Li, R., Huang, Y. (2019). Feature screening in ultrahigh-dimensional varying-coefficient Cox model. *J. Multivariate Anal.* **171**, 284–297. MR3899064

[41] Yue, M., Li, J. (2017). Improvement screening for ultra-high dimensional data with censored survival outcomes and varying coefficients. *Int. J. Biostat.* **13**, Article Number: 20170024, 1–16. MR3667106

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

[43] Zhao, D., Li, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *J. Multivariate Anal.* **105**, 397–411. MR2877525

[44] Zhu, L.-P., Li, L., Li, R., Zhu, L.-X. (2011). Model-free feature screening for ultrahigh dimensional data. *J. Amer. Statist. Assoc.* **106**, 1464–1475. MR2896849

[45] Zucker, D., Karr, A. (1990). Nonparametric survival analysis with time-dependent covariate effects: A penalized partial likelihood approach. *Ann. Statist.* **18**, 329–353. MR1041396

Lianqiang Qu
School of Mathematics and Statistics
Central China Normal University
Wuhan, 430079
P.R. China
E-mail address: qulianq@amss.ac.cn

Liuquan Sun
Institute of Applied Mathematics
Academy of Mathematics and Systems Science
Chinese Academy of Sciences
Beijing, 100190
P.R. China
E-mail address: slq@amt.ac.cn