Statistica Sinica Preprint No: SS-14-171R1

Title	Feature screening in ultrahigh dimensional cox’s model
Manuscript ID	SS-14-171
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.2014.171
Complete List of Authors	Guangren Yang Ye Yu Runze Li and Anne Buu
Corresponding Author	Runze Li
E-mail	rzli@psu.edu

Feature Screening in Ultrahigh Dimensional Cox's Model

Guangren Yang, Ye Yu, Runze Li and Anne Buu

Jinan University, Pennsylvania State University
University of Michigan

Abstract: Survival data with ultrahigh dimensional covariates such as genetic markers have been collected in medical studies and other fields. In this work, we propose a feature screening procedure for the Cox model with ultrahigh dimensional covariates. The proposed procedure is distinguished from the existing sure independence screening (SIS) procedures (Fan, Feng and Wu, 2010, Zhao and Li, 2012) in that the proposed procedure is based on joint likelihood of potential active predictors, and therefore is not a marginal screening procedure. The proposed procedure can effectively identify active predictors that are jointly dependent but marginally independent of the response without performing an iterative procedure. We develop a computationally effective algorithm to carry out the proposed procedure and establish the ascent property of the proposed algorithm. We further prove that the proposed procedure possesses the sure screening property. That is, with the probability tending to one, the selected variable set includes the actual active predictors. We conduct Monte Carlo simulation to evaluate the finite sample performance of the proposed procedure and further compare the proposed procedure and existing SIS procedures. The proposed methodology is also demonstrated through an empirical analysis of a real data example.

Key words and phrases: Cox's model, penalized likelihood, partial likelihood, ultrahigh dimensional survival data.

1. Introduction

Modeling high dimensional data has become the most important research topic in literature. Variable selection is fundamental in analysis of high dimensional data. Feature screening procedures that can effectively reduce ultrahigh dimensionality become indispensable for ultrahigh dimensional data and have attracted considerable attentions in recent literature. Fan and Lv (2008) proposed a marginal screening procedure for ultrahigh dimensional Gaussian linear models, and further demonstrated that the marginal screening procedure may possesses a sure screening property under certain conditions. Such a marginal screening procedure has been referred to as a sure independence screening (SIS) procedure. The SIS procedure has been further developed for generalized linear models and robust linear models in the presence of ultrahigh dimensional covariates (Fan, Samworth and Wu, 2009; Li et al. 2012). The SIS procedure has also been proposed for ultrahigh dimensional additive models (Fan, Feng and Song, 2011) and ultrahigh dimensional varying coefficient models (Liu, Li and Wu, 2014, Fan, Ma and Dai, 2014). These authors showed that their procedures enjoy sure screening property in the language of Fan and Lv (2008) under the settings in which the sample consists of independently and identically distributed observations from a population.

Analysis of survival data is inevitable since the primary outcomes or responses are subject to be censored in many scientific studies. The Cox model (Cox, 1972) is the most commonly-used regression model for survival data, and the partial likelihood method (Cox, 1975) has become a standard approach to parameter estimation and statistical inference for the Cox model. The penalized partial likelihood method has been proposed for variable selection in the Cox model (Tibshirani, 1997; Fan and Li, 2002; Zhang and Lu, 2007; Zou, 2008). Many studies collect survival data as well as a huge number of covariates such as genetic markers. Thus, it is of great interest to develop new data analytic tools for analysis of survival data with ultrahigh dimensional covariates. Bradic, Fan and Jiang (2011) extended the penalized partial likelihood approach for the Cox model with ultrahigh dimensional covariates. Huang, *et al* (2013) studied the penalized partial likelihood with the L_1-penalty for the Cox model with high dimensional covariates. In theory, the penalized partial likelihood may be used to select significant variables in ultrahigh dimensional Cox models. However, in practice, the penalized partial likelihood may suffer from algorithm instability, statistical inaccuracy and highly computational cost when the dimension of covariate vector is much greater than the sample size. Feature screening may play a fundamental role in analysis of ultrahigh dimensional survival data. Fan, Feng and Wu (2010) proposed a SIS procedure for the Cox model by measuring the importance of predictors based on marginal partial likelihood. Zhao and Li (2012) further developed a principled Cox SIS procedure which essentially ranks the importance of a covariate by its t-value of marginal partial likelihood estimate and selects a cutoff to control the false discovery rate.

In this paper, we propose a new feature screening procedure for ultrahigh dimensional Cox models. The proposed procedure is distinguished from the SIS procedures (Fan, Feng and Wu, 2010; Zhao and Li, 2012) in that it is based on the joint partial likelihood of potential important features rather than the marginal partial likelihood of individual feature. Non-marginal screening procedures have been demonstrated their advantage over the SIS procedures in the context of generalized linear models. For example, Wang (2009) proposed a forward regression approach to feature screening in ultrahigh dimensional linear models. Xu and Chen (2014) proposed a feature screening procedure for generalized linear models via the sparsity-restricted maximum likelihood estimator. Both Wang (2009) and Xu and Chen (2014) demonstrated their approaches can perform significantly better than the SIS procedures under some scenarios. However, their methods are merely for linear and generalized linear models. In this paper, we will show that the newly proposed procedure can outperform the sure independence screening procedure for the Cox model. This work makes the following major contribution to the literature.

(a) We propose a sure joint screening (SJS) procedure for ultrahigh dimensional Cox model. We further propose an effective algorithm to carry out the proposed screening procedure, and demonstrate the ascent property of the proposed algorithm.

(b) We establish the screening property for the SJS procedure. This indeed is challenging because the theoretical tools for penalized partial likelihood for the ultrahigh dimensional Cox model cannot be utilized in our context. This work is the first to employ Hoeffding inequality for a sequence of martingale differences to establish concentration inequality for the score function of partial likelihood.

We further conduct Monte Carlo simulation studies to assess the finite sample performance of the proposed procedure and compare its performance with existing sure screening procedure for ultrahigh dimensional Cox models. Our numerical results indicate that the proposed SJS procedure outperforms the existing SIS procedures. We also demonstrate the proposed joint screening procedure by an empirical analysis of a real data example.

The rest of this paper is organized as follows. In Section 2, we propose a new feature screening for the Cox model, and further demonstrate the ascent property of our proposed algorithm to carry out the proposed feature screening procedure. We also study the sampling property of the proposed procedure and establish its sure screening property. In Section 3, we present numerical comparisons and an empirical analysis of a real data example. Some discussion and conclusion remarks are given in Section 4. Technical proofs are given in the Appendix.

2. New feature screening procedure for Cox's model

Let T and $\mathbf{x}$ be the survival time and its p-dimensional covariate vector, respectively. Throughout this paper, we consider the following Cox proportional hazard model:

$$h(t|\mathbf{x}) = h_0(t)\exp(\mathbf{x}^T\boldsymbol{\beta}), \tag{2.1}$$

where $h_0(t)$ is an unspecified baseline hazard functions and $\boldsymbol{\beta}$ is an unknown parameter vector. In survival data analysis, the survival time may be censored by the censoring time C. Denote the observed time by $Z = \min\{T, C\}$ and the event indicator by $\delta = I(T \leq C)$. We assume the censoring mechanism is noninformative. That is, given $\mathbf{x}$, T and C are conditionally independent.

Suppose that $\{(\mathbf{x}_i, Z_i, \delta_i) : i = 1, \cdots, n\}$ is an independently and identically distributed random sample from model (2.1). Let $t_1^0 < \cdots < t_N^0$ be the ordered observed failure times. Let (j) provide the label for the subject failing at t_j^0 so that the covariates associated with the N failures are $\mathbf{x}_{(1)}, \cdots, \mathbf{x}_{(N)}$. Denote the risk set right before the time t_j^0 by R_j:

$$R_j = \{i : Z_i \geq t_j^0\}.$$

The partial likelihood function (Cox, 1975) of the random sample is

$$\ell_p(\boldsymbol{\beta}) = \sum_{j=1}^{N} [\mathbf{x}_{(j)}^T \boldsymbol{\beta} - \log\{\sum_{i \in R_j} \exp(\mathbf{x}_i^T \boldsymbol{\beta})\}]. \tag{2.2}$$

2.1 A new feature screening procedure

Suppose that the effect of $\mathbf{x}$ is sparse. Denote the true value of $\boldsymbol{\beta}$ by $\boldsymbol{\beta}^*$. The sparsity implies that $\|\boldsymbol{\beta}^*\|_0$ is small, where $\|\mathbf{a}\|_0$ stands for the L_0-norm of $\mathbf{a}$ (i.e. the number of nonzero elements of $\mathbf{a}$). In the presence of ultrahigh dimensional covariates, one may consider to reduce the ultrahigh dimensionality to a moderate one by an effective feature screening method. In this section, we propose screening features in the Cox model by the constrained partial likelihood

$$\widehat{\boldsymbol{\beta}}_m = \arg\max_{\boldsymbol{\beta}} \ell_p(\boldsymbol{\beta}) \text{ subject to } \|\boldsymbol{\beta}\|_0 \leq m \tag{2.3}$$

for a pre-specified m which is assumed to be greater than the number of nonzero elements of $\boldsymbol{\beta}^*$. For high dimensional problems, it becomes almost impossible to solve the constrained maximization problem (2.3) directly. Alternatively, we consider a proxy of the partial likelihood function. It follows by the Taylor expansion for the partial likelihood function $\ell_p(\boldsymbol{\gamma})$ at $\boldsymbol{\beta}$ lying within a neighbor of $\boldsymbol{\gamma}$ that

$$\ell_p(\boldsymbol{\gamma}) \approx \ell_p(\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^T \ell_p'(\boldsymbol{\beta}) + \frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\beta})^T \ell_p''(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta}),$$

where $\ell_p'(\boldsymbol{\beta}) = \partial \ell_p(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}|_{\boldsymbol{\gamma}=\boldsymbol{\beta}}$ and $\ell_p''(\boldsymbol{\beta}) = \partial^2 \ell_p(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T|_{\boldsymbol{\gamma}=\boldsymbol{\beta}}$. When $p < n$ and $\ell_p''(\boldsymbol{\beta})$ is invertible, the computational complexity of calculating the inverse of $\ell_p''(\boldsymbol{\beta})$ is $O(p^3)$. For the setting of large p and small n, $\ell_p''(\boldsymbol{\beta})$ is not invertible. Low computational costs are always desirable for feature screening. To deal with singularity of the Hessian matrix and save computational costs, we propose to use the following approximation for $\ell_p''(\boldsymbol{\gamma})$

$$g(\boldsymbol{\gamma}|\boldsymbol{\beta}) = \ell_p(\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^T \ell_p'(\boldsymbol{\beta}) - \frac{u}{2}(\boldsymbol{\gamma} - \boldsymbol{\beta})^T W (\boldsymbol{\gamma} - \boldsymbol{\beta}), \tag{2.4}$$

where u is a scaling constant to be specified and W is a diagonal matrix. Throughout this paper, we use $W = \text{diag}\{-\ell_p''(\boldsymbol{\beta})\}$, the matrix consisting of the diagonal elements of $-\ell_p''(\boldsymbol{\beta})$. This implies that we approximate $\ell_p''(\boldsymbol{\beta})$ by $u\,\text{diag}\{\ell_p''(\boldsymbol{\beta})\}$.

Remark. Xu and Chen (2014) proposed a feature screening procedure by iterative hard-thresholding algorithm (IHT) for generalized linear models with independently and identically distributed (iid) observations. They approximated the likelihood function $\ell(\boldsymbol{\gamma})$ of the observed data by a linear approximation $\ell(\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^T \ell'(\boldsymbol{\beta})$, but they also introduced a regularization term $-u\|\boldsymbol{\gamma} - \boldsymbol{\beta}\|^2$. Thus, the $g(\boldsymbol{\gamma}|\boldsymbol{\beta})$ in Xu and Chen (2014) would coincide with the one in (2.4) if one set $W = I_p$, the $p \times p$ identity matrix, but the motivation of our proposal indeed is different from theirs, and the working matrix W is not set to be I_p throughout this paper.

It can be seen that $g(\boldsymbol{\beta}|\boldsymbol{\beta}) = \ell_p(\boldsymbol{\beta})$, and under some conditions, $g(\boldsymbol{\gamma}|\boldsymbol{\beta}) \le \ell_p(\boldsymbol{\beta})$ for all $\boldsymbol{\gamma}$. This ensures the ascent property. See Theorem 1 below for more details. Since W is a diagonal matrix, $g(\boldsymbol{\gamma}|\boldsymbol{\beta})$ is an additive function of γ_j for any given $\boldsymbol{\beta}$. The additivity enables us to have a closed form solution for the following maximization problem

$$\max_{\boldsymbol{\gamma}} g(\boldsymbol{\gamma}|\boldsymbol{\beta}) \qquad \text{subject to} \quad \|\boldsymbol{\gamma}\|_0 \le m \tag{2.5}$$

for given $\boldsymbol{\beta}$ and m. Note that the maximizer of $g(\boldsymbol{\gamma}|\boldsymbol{\beta})$ is $\tilde{\boldsymbol{\gamma}} = \boldsymbol{\beta} + u^{-1}W^{-1}\ell_p'(\boldsymbol{\beta})$. Denote $r_j = w_j\tilde{\gamma}_j^2$ with w_j being the j-th diagonal element of W for $j = 1, \cdots, p$, and sort r_j so that $|r_{(1)}| \ge |r_{(2)}| \ge \cdots \ge |r_{(p)}|$. The solution of maximization problem (2.5) is the hard-thresholding rule defined below

$$\widehat{\gamma}_j = \tilde{\gamma}_j I\{|r_j| > |r_{(m+1)}|\} \widehat{=} H(\tilde{\gamma}_j; m). \tag{2.6}$$

This enables us to effectively screen features by using the following algorithm:

Step 1. Set the initial value $\boldsymbol{\beta}^{(0)} = \mathbf{0}$.

Step 2. Set $t = 0, 1, 2, \cdots$ and iteratively conduct Step 2a and Step 2b below until the algorithm converges.

Step 2a. Calculate $\tilde{\boldsymbol{\gamma}}^{(t)} = (\tilde{\gamma}_1^{(t)}, \cdots, \tilde{\gamma}_p^{(t)})^T = \boldsymbol{\beta}^{(t)} + u_t^{-1}W^{-1}(\boldsymbol{\beta}^{(t)})\ell_p'(\boldsymbol{\beta}^{(t)})$, and

$$\tilde{\boldsymbol{\beta}}^{(t)} = (H(\tilde{\gamma}_1^{(t)}; m), \cdots, H(\tilde{\gamma}_p^{(t)}; m))^T \widehat{=} \mathbf{H}(\tilde{\boldsymbol{\gamma}}^{(t)}; m). \tag{2.7}$$

Set $S_t = \{j : \tilde{\beta}_j^{(t)} \ne 0\}$, the nonzero index of $\tilde{\boldsymbol{\beta}}^{(t)}$.

Step 2b. Update $\boldsymbol{\beta}$ by $\boldsymbol{\beta}^{(t+1)} = (\beta_1^{(t+1)}, \cdots, \beta_p^{(t+1)})^T$ as follows. If $j \notin S_t$, set $\beta_j^{(t+1)} = 0$; otherwise, set $\{\beta_j^{(t+1)} : j \in S_t\}$ be the maximum partial likelihood estimate of the submodel S_t.

Unlike the screening procedures based on marginal partial likelihood methods proposed in Fan, Feng and Wu (2010) and further studied in Zhao and Li (2012), our proposed procedure is to iteratively update $\boldsymbol{\beta}$ using Step 2. This enables the proposed screening procedure to incorporate correlation information among the predictors through updating $\ell_p'(\boldsymbol{\beta})$ and $\ell_p''(\boldsymbol{\beta})$. Thus, the proposed procedure is expected to perform better than the marginal screening procedures when there are some predictors that are marginally independent of the survival time, but not jointly independent of the survival time. Meanwhile, since each iteration in Step 2 can avoid large-scale matrix inversion and, therefore, it can be carried out with low computational costs. Based on our simulation study, the proposed procedures can be implemented with less computing time than the marginal screening procedure studied in Fan, Feng and Wu (2000) and Zhao and Li (2012) in some scenarios (see Tables 3.2 and 3.3 for details). Theorem 1 below offers convergence behavior of the proposed algorithm.

Theorem 1. *Suppose that Conditions (D1)—(D4) in the Appendix hold. Denote*

$$\rho^{(t)} = \sup_{\boldsymbol{\beta}} \left[\lambda_{\max}\{W^{-1/2}(\boldsymbol{\beta}^{(t)})\{-\ell_p''(\boldsymbol{\beta})\}W^{-1/2}(\boldsymbol{\beta}^{(t)})\} \right]$$

where $\lambda_{\max}(A)$ stands for the maximal eigenvalue of a matrix A. If $u_t \geq \rho^{(t)}$, then

$$\ell_p(\boldsymbol{\beta}^{(t+1)}) \geq \ell_p(\boldsymbol{\beta}^{(t)}),$$

where $\boldsymbol{\beta}^{(t+1)}$ is defined in Step 2b in the above algorithm.

Theorem 1 claims the ascent property of the proposed algorithm if u_t is appropriately chosen. That is, the proposed algorithm may improve the current estimate within the feasible region (i.e. $\|\beta\|_0 \leq m$), and the resulting estimate in the current step may serve as a refinement of the last step. This theorem also provides us some insights about choosing u_t in practical implementation. In our numerical studies, this algorithm typically converges within six iterations. It is worth noting that Theorem 1 does not implies that the proposed algorithm converges to converge to the global optimizer.

2.2 Sure screening property

For the convenience of presentation, we use s to denote an arbitrary subset of $\{1,\ldots,p\}$, which amounts to a submodel with covariates $\mathbf{x}_s = \{x_j, j \in s\}$ and associated coefficients $\boldsymbol{\beta}_s = \{\beta_j, j \in s\}$. Also, we use $\tau(s)$ to indicate the size of model s. In particular, we denote the true model by $s^* = \{j : \beta_j^* \neq 0, 1 \leq j \leq p_n\}$ with $\tau(s^*) = \|\boldsymbol{\beta}^*\|_0 = q$. The objective of feature selection is to obtain a subset $\widehat{s}$ such that $s^* \subset \widehat{s}$ with a very high probability.

We now provide some theoretical justifications for the newly proposed feature screening procedure. The sure screening property (Fan and Lv, 2008) is referred to as

$$Pr(s^* \subset \widehat{s}) \longrightarrow 1, \quad \text{as} \quad n \to \infty, \tag{2.8}$$

To establish this sure screening property for the proposed SJS, we introduce some additional notations as follows. For any model s, let $\ell'(\boldsymbol{\beta}_s) = \partial \ell(\boldsymbol{\beta}_s)/\partial \boldsymbol{\beta}_s$ and $\ell''(\boldsymbol{\beta}_s) = \partial^2 \ell(\boldsymbol{\beta}_s)/\partial \boldsymbol{\beta}_s \partial \boldsymbol{\beta}_s^T$ be the score function and the Hessian matrix of $\ell(\cdot)$ as a function of $\boldsymbol{\beta}_s$, respectively. Assume that a screening procedure retains m out of p features such that $\tau(s^*) = q < m$. So, we define

$$S_+^m = \{s : s^* \subset s; \|s\|_0 \leq m\} \quad \text{and} \quad S_-^m = \{s : s^* \not\subset s; \|s\|_0 \leq m\}$$

as the collections of the over-fitted models and the under-fitted models. We investigate the asymptotic properties of $\widehat{\boldsymbol{\beta}}_m$ under the scenario where p, q, m and $\boldsymbol{\beta}^*$ are allowed to depend on the sample size n. We impose the following conditions, some of which are purely technical and only serve to facilitate theoretical understanding of the proposed feature screening procedure.

(C1) There exist $w_1, w_2 > 0$ and some non-negative constants τ_1, τ_2 such that $\tau_1 + \tau_2 < 1/2$ and

$$\min_{j \in s^*} |\beta_j^*| \geq w_1 n^{-\tau_1} \quad \text{and} \quad q < m \leq w_2 n^{\tau_2}.$$

(C2) $\log p = O(n^\kappa)$ for some $0 \leq \kappa < 1 - 2(\tau_1 + \tau_2)$.

(C3) There exist constants $c_1 > 0$, $\delta_1 > 0$, such that for sufficiently large n,

$$\lambda_{\min}[-n^{-1}\ell_p''(\boldsymbol{\beta}_s)] \geq c_1$$

for $\boldsymbol{\beta}_s \in \{\boldsymbol{\beta} : \|\boldsymbol{\beta}_s - \boldsymbol{\beta}_s^*\|_2 \leq \delta_1\}$ and $s \in S_+^{2m}$, where $\lambda_{\min}[\cdot]$ denotes the smallest eigenvalue of a matrix.

Condition (C1) states a few requirements for establishing the sure screening property of the proposed procedure. The first one is the sparsity of $\boldsymbol{\beta}^*$ which makes the sure screening possible with $\tau(\widehat{s}) = m > q$. Also, it requires that the minimal component in $\boldsymbol{\beta}^*$ does not degenerate too fast, so that the signal is detectable in the asymptotic sequence. Meanwhile, together with (C3), it confines an appropriate order of m that guarantees the identifiability of s^* over s for $\tau(s) \leq m$. Condition (C2) assumes that p diverges with n at up to an exponential rate; it implies that the number of covariates can be substantially larger than the sample size. We establish the sure screening property of the quasi-likelihood estimation by the following theorem.

Theorem 2. *Suppose that Conditions (C1)—(C3) and Conditions (D1)—(D7) in the Appendix hold. Let $\widehat{s}$ be the model obtained by the (2.3) of size m. We have*

$$Pr(s^* \subset \widehat{s}) \to 1, \quad as \quad n \to \infty.$$

The proof is given in the Appendix. The sure screening property is an appealing property of a screening procedure since it ensures that the true active predictors are retained in the model selected by the screening procedure. One has to specify the value of m in practical implementation. In the literature of feature screening, it is typical to set $m = [n/\log(n)]$ (Fan and Lv, 2008). Although it is an ad hoc choice, it works reasonably well in our numerical examples. With this choice of m, one is ready to further apply existing methods such as the penalized partial likelihood method (See, for example, Tibshirani, 1997, Fan and Li, 2002) to further remove inactive predictors. Thus, we set $m = [n/\log(n)]$ throughout the numerical studies of this paper. To be distinguished from the SIS procedure, the proposed procedure is referred to as sure joint screening (SJS) procedure.

3. Numerical studies

In this section, we evaluate the finite sample performance of the proposed feature screening procedure via Monte Carlo simulations. We further illustrate the proposed procedure via an empirical analysis of a real data set. All simulations were conducted by using R codes.

3.1 Simulation studies

The main purpose of our simulation studies is to compare the performance of the SJS with the SIS procedure for the Cox model (Cox-SIS) proposed by Fan, Feng and Wu (2010) and further studied by Zhao and Li (2012). To make a fair comparison, we set the model size of Cox-SIS to be the same as that of our new procedure. In our simulation, the predictor variable $\mathbf{x}$ is generated from a p-dimensional normal distribution with mean zero and covariance matrix $\Sigma = (\sigma_{ij})$. Two commonly-used covariance structures are considered.

(S1) Σ is compound symmetric. That is, $\sigma_{ij} = \rho$ for $i \neq j$ and equal 1 for $i = j$. We take $\rho = 0.25$, 0.50 and 0.75.

(S2) Σ has autoregressive structure. That is, $\sigma_{ij} = \rho^{|i-j|}$. We also consider $\rho = 0.25$, 0.5 and 0.75.

We generate the censoring time from an exponential distribution with mean 10, and the survival time from the Cox model with $h_0(t) = 10$ and two sets of $\boldsymbol{\beta}$s listed below:

(b1) $\beta_1 = \beta_2 = \beta_3 = 5$, $\beta_4 = -15\rho$, and other β_js equal 0.

(b2) $\beta_j = (-1)^U(a + |V_j|)$ for $j = 1, 2, 3$ and 4, where $a = 4\log n/\sqrt{n}$, $U \sim$ Bernoulli(0.4) and $V_j \sim \mathcal{N}(0, 1)$.

Under the setting (S1) and (b1), X_4 is jointly dependent but marginally independent of the survival time for all $\rho \neq 0$. Thus, this setting is designed to challenge the marginal SIS procedures. The coefficients in (b2) was used in Fan and Lv (2008), and here we adopt it for survival data.

In our simulation, we consider the sample size $n = 100$ and 200, and the dimension p=2000 and 5000. For each combination, we conduct 1000 replicates of Monte Carlo simulation. We compare the performance of feature screening procedures using the following two criteria:

1. P_s: the proportion that an individual active predictor is selected for a given model size m in the 1000 replications.
2. P_a: the proportion that all active predictors are selected for a given model size m in the 1000 replications.

The sure screening property ensures that P_s and P_a are both close to one when the estimated model size m is sufficiently large. We choose $m = [n/\log n]$ throughout our simulations, where $[a]$ denotes the integer that a is rounded to.

It is expected that the performance of SJS depends on the following factors: the structure of the covariance matrix, the values of $\boldsymbol{\beta}$, the dimension of all candidate features and the sample size n. In survival data analysis, the performance of a statistical procedure depends on the censoring rate. Table 3.1 depicts the censoring rates for the 12 combinations of covariance structure, the values of ρ and values of $\boldsymbol{\beta}$. It can be seen from Table 3.1 that the censoring rate ranges from 13% to 35%, which lies in a reasonable range to carry out simulation studies.

Table 3.1: Censoring Rates

	$\rho = 0.25$		$\rho = 0.50$		$\rho = 0.75$	
Σ	$\boldsymbol{\beta}$ in (b1)	$\boldsymbol{\beta}$ in (b2)	$\boldsymbol{\beta}$ in (b1)	$\boldsymbol{\beta}$ in (b2)	$\boldsymbol{\beta}$ in (b1)	$\boldsymbol{\beta}$ in (b2)
S1	0.329	0.163	0.317	0.148	0.293	0.239
S2	0.323	0.181	0.353	0.135	0.342	0.227

Table 3.2 reports $\mathcal{P}_s$ for the active predictors and $\mathcal{P}_a$ when the covariance matrix of $\mathbf{x}$ is the compound symmetric (i.e., S1). Table 3.2 also depicts the average computing time for each replication. Note that under the design of (S1) with (b1), X_4 is jointly dependent but marginally independent of the survival time for all $\rho \neq 0$. This setting is designed to challenge all screening procedures, in particularly the marginal screening procedures. As shown in Table 3.2, Cox-SIS fails to identify X_4 as an active predictor completely when $\boldsymbol{\beta}$ is set to be the

Table 3.2: The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ with $\Sigma = (1-\rho)I + \rho\mathbf{11}^T$

		Cox-SIS						SJS					
		$\mathcal{P}_s$				$\mathcal{P}_a$	Time	$\mathcal{P}_s$				$\mathcal{P}_a$	Time
ρ	$\boldsymbol{\beta}$	X_1	X_2	X_3	X_4	ALL	(s)	X_1	X_2	X_3	X_4	ALL	(s)
		$n = 100$ and $p = 2000$											
.25	b1	.984	.991	.991	0	0	13.07	.999	.995	.997	.981	.975	7.54
	b2	.826	.817	.826	.842	.437	12.94	.993	.992	.993	.997	.984	7.81
.50	b1	.951	.948	.937	.001	.001	13.07	.961	.962	.962	.983	.937	8.31
	b2	.73	.707	.707	.734	.236	12.95	.981	.976	.977	.976	.936	8.47
.75	b1	.761	.783	.775	.008	.005	12.09	.954	.943	.942	.987	.898	8.36
	b2	.611	.638	.619	.62	.134	9.22	.887	.891	.9	.898	.717	6.07
		$n = 100$ and $p = 5000$											
.25	b1	.977	.975	.981	0	0	32.00	0.988	0.981	0.984	0.925	0.912	26.00
	b2	.739	.788	.763	.769	.317	27.76	.972	.974	.978	.975	.938	54.98
0.50	b1	.892	.9	.894	0	0	42.82	0.871	0.861	0.862	0.948	0.805	31.89
	b2	.636	.619	.643	.629	.127	28.25	.919	.922	.934	.923	.812	59.68
.75	b1	.701	.696	.659	.008	.002	30.94	0.829	0.838	0.828	0.988	0.724	36.73
	b2	.501	.501	.488	.472	.045	25.90	.78	.799	.784	.783	.486	49.65
		$n = 200$ and $p = 2000$											
.25	b1	1	1	1	0	0	15.90	1	1	1	1	1	16.32
	b2	.977	.971	.979	.964	.897	6.99	1	1	1	1	1	5.94
.50	b1	.999	1	1	0	0	12.20	1	1	1	1	1	12.54
	b2	.95	.946	.932	.942	.786	16.29	1	1	1	1	1	16.46
.75	b1	.989	.99	.994	.001	.001	15.79	1	1	1	1	1	17.70
	b2	.887	.873	.883	.909	.597	18.34	1	.998	1	1	.998	20.33
		$n = 200$ and $p = 5000$											
.25	b1	1	1	1	0	0	34.32	1	1	1	1	1	160.33
	b2	.952	.962	.949	.958	.825	42.47	1	1	1	1	1	211.99
.50	b1	.999	.998	1	0	0	32.71	1	1	1	1	1	181.90
	b2	.904	.903	.892	.885	.637	30.38	1	1	1	1	1	152.62
.75	b1	.978	.976	.985	.004	.004	34.83	1	1	1	.999	.999	218.22
	b2	.823	.832	.832	.812	.431	28.40	.998	.999	.997	.999	.993	146.69

one in (b1). This is expected. The newly proposed SJS procedure, on the other hand, includes X_4 with nearly every simulation. In addition, SJS has the value of $\mathcal{P}_a$ very close to one for every case when $\boldsymbol{\beta}$ is set to be the one in (b1). There is no doubt that SJS outperforms Cox-SIS easily in this setting.

We next discuss the performance of the Cox-SIS and the SJS when the covariance matrix of $\mathbf{x}$ is compound symmetric and $\boldsymbol{\beta}$ is set to be the one in (b2). In this setting, there is no predictor that is marginally independent of, but jointly dependent with the response. Table 3.2 clearly shows that how the performance of Cox-SIS and SJS is affected by the sample size, the dimension of predictors and the value of ρ. Overall, the SJS outperforms the Cox-SIS in all cases in terms of $\mathcal{P}_s$ and $\mathcal{P}_a$. The improvement of SJS over Cox-SIS is quite significant when the sample size is small (i.e., $n = 100$) or when $\rho = 0.75$. The performance of SJS becomes better as the sample size increases. This is consistent with our theoretical analysis since the SJS has the sure screening property.

Table 3.3: The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ with $\Sigma = (\rho^{|i-j|})$

		Cox-SIS						SJS					
		$\mathcal{P}_s$				$\mathcal{P}_a$	Time	$\mathcal{P}_s$				$\mathcal{P}_a$	Time
ρ	$\boldsymbol{\beta}$	X_1	X_2	X_3	X_4	ALL	(s)	X_1	X_2	X_3	X_4	ALL	(s)
		$n = 100$ and $p = 2000$											
.25	b1	1	1	.997	.183	.182	10.46	1	1	1	.989	.989	5.84
	b2	.989	1	.999	.983	.971	10.60	1	1	1	1	1	5.55
.50	b1	1	1	.941	.446	.394	10.61	.998	.997	.936	.97	.931	5.91
	b2	1	1	1	.999	.999	12.29	1	1	1	1	1	6.31
.75	b1	1	1	.525	.364	.048	6.57	.985	.927	.641	.907	.615	3.77
	b2	1	1	1	1	1	10.71	1	1	1	1	1	5.47
		$n = 100$ and $p = 5000$											
.25	b1	1	1	.991	.135	.131	32.23	1	1	1	.965	.965	59.62
	b2	.981	.999	1	.975	.955	40.31	.999	1	1	.999	.999	74.80
0.50	b1	1	1	.888	.296	.214	38.82	.992	.981	.821	.896	.811	70.76
	b2	.999	1	1	.999	.998	42.13	1	1	1	1	1	71.58
.75	b1	1	1	.439	.23	.019	29.09	.959	.82	.449	.783	.415	53.55
	b2	1	1	1	1	1	31.05	1	1	1	1	1	52.37
		$n = 200$ and $p = 2000$											
.25	b1	1	1	1	.592	.592	12.93	1	1	1	1	1	11.62
	b2	1	1	1	1	1	13.20	1	1	1	1	1	13.11
0.50	b1	1	1	.999	.869	.868	12.96	1	1	1	1	1	10.47
	b2	1	1	1	1	1	12.78	1	1	1	1	1	11.39
.75	b1	1	1	.921	.757	.678	12.91	1	1	.999	.999	.998	11.17
	b2	1	1	1	1	1	14.26	1	1	1	1	1	12.39
		$n = 200$ and $p = 5000$											
.25	b1	1	1	1	.45	.45	37.59	1	1	1	1	1	192.79
	b2	1	1	1	1	1	35.63	1	1	1	1	1	166.09
.50	b1	1	1	1	.79	.79	38.47	1	1	1	1	1	166.29
	b2	1	1	1	1	1	27.90	1	1	1	1	1	132.96
.75	b1	1	1	.88	.674	.554	47.62	1	1	.993	.997	.991	235.95
	b2	1	1	1	1	1	34.52	1	1	1	1	1	163.85

Table 3.2 also indicates that the performance of Cox-SIS is better as the sample size increases, the feature dimension decreases or the value of ρ decreases. However, these factors have less impacts on the performance of SJS. For every case listed in Table 3.2, SJS outperforms Cox-SIS no matter whether there presents marginally independent but jointly dependent predictors or not. In terms of computing time, SJS and Cox-SIS are comparable. For $p = 2000$, SJS needs slightly less computing time than Cox-SIS, while SJS needs more computing time for $p = 5000$.

Table 3.3 depicts the simulation results for the AR covariance structure (S2). It is worth noting that with the AR covariance structure and $\boldsymbol{\beta}$ being set to the one in (b1) or (b2), none of the active predictors $X_1, \cdots, X_4$ is marginally independent of the survival time. Thus, one would expect that the Cox-SIS works well for both cases (b1) and (b2). Table 3.3 indicates that both Cox-SIS and SJS perform very well when $\boldsymbol{\beta}$ is set to be the one in (b2). On the other

Table 3.4: Comparison with Cox-ISIS

		Cox-ISIS						SJS					
		$\mathcal{P}_s$				$\mathcal{P}_a$	Time	$\mathcal{P}_s$				$\mathcal{P}_a$	Time
p	ρ	X_1	X_2	X_3	X_4	ALL	(s)	X_1	X_2	X_3	X_4	ALL	(s)
		$\tau = 1$											
2000	.25	.998	.998	.999	1	.996	23.34	.999	.996	.995	.979	.975	5.75
	.5	.898	.894	.897	1	.708	21.47	.97	.968	.975	.983	.952	6.05
	.75	.697	.696	.694	1	.303	19.03	.952	.949	.953	.993	.903	5.72
5000	.25	.998	.994	.999	.992	.983	36.47	.988	.981	.984	.925	.912	26.00
	.5	.819	.833	.853	1	.562	37.81	.871	.861	.862	.948	.805	31.89
	.75	.579	.583	.611	1	.177	38.81	.829	.838	.828	.988	.724	36.73
		$\tau = 0.75$											
2000	.25	1	.997	1	.999	.996	14.19	.999	.998	1	.98	.978	3.85
	.5	.896	.899	.904	1	.712	14.10	.97	.969	.97	.987	.952	4.47
	.75	.709	.687	.724	1	.334	22.99	.936	.938	.942	.99	.882	7.33
5000	.25	.991	.996	.99	.99	.972	42.64	.983	.985	.988	.931	.914	52.50
	.5	.84	.823	.844	1	.563	44.85	.895	.89	.896	.956	.848	43.96
	.75	.566	.584	.555	1	.167	50.80	.832	.819	.836	.985	.7	55.27
		$\tau = 0.5$											
2000	.25	.997	.997	.999	1	.994	14.45	1	.997	.998	.981	.978	3.99
	.5	.891	.888	.899	1	.702	26.78	.957	.962	.963	.987	.943	8.81
	.75	.672	.678	.665	1	.273	13.95	.883	.889	.889	.99	.772	4.79
5000	.25	.993	.995	.99	.993	.975	41.41	.977	.983	.989	.912	.897	34.82
	.5	.806	.847	.805	1	.527	56.10	.874	.867	.855	.946	.803	57.31
	.75	.56	.574	.544	1	.161	40.54	.738	.761	.746	.975	.564	61.49
		$\tau = 0.25$											
2000	.25	.97	.972	.976	.973	.902	14.40	.971	.971	.981	.853	.824	3.72
	.5	.822	.806	.819	1	.534	14.45	.866	.845	.833	.966	.748	5.00
	.75	.528	.536	.526	1	.126	14.48	.552	.566	.564	.952	.238	4.72
5000	.25	.941	.936	.934	.949	.805	43.85	.901	.914	.897	.675	.592	59.46
	.5	.731	.736	.709	.999	.366	45.25	.664	.671	.645	.86	.475	50.66
	.75	.466	.432	.419	1	.067	49.79	.427	.389	.372	.958	.1	118.30

hand, the Cox-SIS has very low $\mathcal{P}_a$ when $n = 100$ and $\boldsymbol{\beta}$ is set to be the one in (b1), although $\mathcal{P}_a$ becomes much higher when the sample size increases from 100 to 200. In summary, SJS outperform Cox-SIS in all cases considered in Table 3.3, in particular, when $\boldsymbol{\beta}$ is set to be the one in (b1). In terms of computing time, the pattern is similar to that in Table 3.2.

We next compare SJS with the iterative Cox-SIS. Table 3.2 indicates that Cox-SIS fails to identify the active predictor X_4 under the compound symmetric covariance (S1) when $\boldsymbol{\beta}$ is set to be the one in (b1) because this setting leads X_4 to be jointly dependent but marginally independent of the survival time. Fan, Feng and Wu (2010) proposed iterative SIS for Cox model (abbreviated as Cox-ISIS). Thus, it is of interest to compare the newly proposed procedure with the Cox-ISIS. To this end, we conduct simulation under the settings with S1, b1 and $n = 100$. In this simulation study, we also investigate the impact of signal strength to the performance of the proposed procedure by considering

$\beta_1 = \beta_2 = \beta_3 = 5\tau$, $\beta_4 = -15\tau\rho$, and other β_js equal 0. We take $\tau = 1$, 0.75, 0.5 and 0.25. To make a fair comparison, the Cox-ISIS is implemented by iterating Cox-SIS twice (each with the size $m/2$) so that the number of the included predictors equals $m = [n/\log(n)] = 22$ for both Cox-SIS and the SJS.

The simulation results are summarized in Table 3.4. In addition to the two criteria $\mathcal{P}_s$ and $\mathcal{P}_a$, we report the computing time consumed by both procedures due to their iterative essence. Table 3.4 indicates that when $\rho = 0.25$ is small, both Cox-ISIS and SJS work quite well while SJS takes less time than ISIS. When $\rho = 0.5$ and 0.75, SJS can significantly outperform Cox-ISIS in terms of $\mathcal{P}_s$ and $\mathcal{P}_a$. SJS has less computing time than Cox-ISIS when $p = 2000$, and is comparable in computing time to Cox-ISIS when $p = 5000$.

3.2 An application As an illustration, we apply the proposed feature screening procedure for an empirical analysis of microarray diffuse large-B-cell lymphoma (DLBCL) data (Rosenwald et al., 2002). Given that DLBCL is the most common type of lymphoma in adults and has a survival rate of only about 35 to 40 percent after the standard chemotherapy, there has been continuous interest to understand the genetic markers that may have impacts on the survival outcome.

Table 3.5: Four-three Gene IDs selected by Cox-SJS, Cox-ISIS and Cox-SIS

	SJS			Cox-ISIS			Cox-SIS		
Gene	269	3811	6156	427	2108	4548	1072	1841	5027
IDs	807	3818	6517	655	2109	4721	1188	2437	5054
	1023	3819	6607	1188	2244	4723	1439	2579	5055
	1191	3820	6758	1456	2246	5034	1456	2672	5297
	1662	3821	6844	1579	2361	5055	1660	3799	5301
	1664	3824	6908	1662	2579	5301	1662	3810	5614
	1682	3825	6956	1671	3799	5614	1663	3811	5950
	1825	3826	7068	1681	3811	5649	1664	3812	5953
	2115	4025	7070	1682	3813	5950	1671	3813	6365
	3332	4216	7175	1825	3822	6956	1672	3820	6519
	3372	4317	7343	1878	3824	7098	1678	3821	7096
	3373	4401	7357	1996	3825	7343	1680	3822	7343
	3497	4545	7380	2064	4131	7357	1681	3824	7357
	3791	4595		2106	4317		1682	3825	
	3810	5668		2107	4448		1825	4131	

This data set consists of the survival time of $n = 240$ DLBCL patients after chemotherapy, and $p = 7399$ cDNA microarray expressions of each individual patient as predictors. Given such a large number of predictors and the small sample size, feature screening seems to be a necessary initial step as a prelude to sophisticated statistical modeling procedure that cannot deal with high dimensional survival data. All predictors are standardized so that they have mean zero and variance one.

There are five patients with survival time being close to 0. After removing them from our analysis, our empirical analysis in this example is based on the sample of 235 patients. As a simple comparison, Cox-SIS, Cox-ISIS, and SJS are all applied to this data and obtain the reduced model with $[235/\log(235)] = 43$ genes. The IDs of genes selected by the three screening procedures are listed in Table 3.5. The maximum of partial likelihood function of three corresponding models obtained by SJS, Cox-ISIS and Cox-SIS procedures are -536.9838, -561.8795, and -600.0885, respectively. This implies that both SJS and Cox-ISIS performs much better than Cox-SIS with SJS performing the best.

Table 3.6: IDs of Selected Genes by SCAD and Lasso

	Gene IDs										
SJS-SCAD	1023	1662	1664	1682	1825	2115	3332	3373	3497	3791	3810
	3811	3818	3819	3820	3821	3824	4317	4545	4595	5668	6156
	6517	6607	6758	6844	6908	7343	7357	7380			
SJS-Lasso	269	807	1023	1191	1664	1682	1825	2115	3332	3373	3497
	3791	3810	3811	3819	3820	3821	4025	4216	4317	4401	4545
	4595	5668	6156	6517	6607	6758	6844	6908	7068	7070	7157
	7343	7357	7380								
ISIS-SCAD	1188	1456	1662	1681	1682	1825	1878	2108	3811	3812	3813
	3822	3824	3825	4317	4448	4548	4723	5034	5055	5649	5950
	6956	7098	7343	7357							
ISIS-Lasso	427	655	1188	1456	1579	1662	1671	1681	1825	1878	2106
	2107	2108	2109	2246	2361	3813	3822	3825	4131	4317	4448
	4548	4723	5034	5055	5301	5614	5649	5950	6956	7098	7343
	7357										
SIS-SCAD	1671	1672	1825	3799	3810	3822	3824	7069	7357		
SIS-Lasso	1188	1456	1664	1671	1825	2437	3821	4131	5027	5297	6519
	7069	7343	7357								

Table 3.7: Likelihood, DF, AIC and BIC of Resulting Models.

	Likelihood	df	BIC	AIC
SJS-SCAD	-546.1902	30	1256.168	1152.380
SJS-Lasso	-542.9862	36	1282.518	1157.972
ISIS-SCAD	-575.7148	26	1293.379	1203.430
ISIS-Lasso	-567.6035	34	1320.833	1203.207
SIS-SCAD	-622.5386	9	1294.213	1263.077
SIS-Lasso	-610.6605	14	1297.755	1249.321

We first apply penalized partial likelihood with the L_1 penalty (Tibshirani, 1997) and with the SCAD penalty (Fan and Li, 2002) for the models obtained from the screening stage. We refer to these two variable selection procedures as Lasso and SCAD for simplicity. The tuning parameter in the SCAD and the Lasso was selected by the BIC tuning parameter selector, a direct extension of Wang, Li and Tsai (2007). The IDs of genes selected by the SCAD and the Lasso are listed in Table 3.6. The likelihood, the degree of freedom (df), the BIC score and the AIC score of the resulting models are listed in Table 3.7, from which SJS-SCAD results in the best fit model in terms of the AIC and BIC. The partial likelihood ratio test for comparing the model selected by SJS-SCAD and SJS without SCAD is 18.41286 with df=13. This leads to the P-value of this partial likelihood ratio test to be 0.142. This implies the model selected by SJS-SCAD is in favor, compared with the one obtained in the screening stage. The resulting estimates and standard errors of the model selected by SJS-SCAD are depicted in Table 3.8, which indicates that most selected genes have significant impact on the survival time. We further compare Tables 3.5 and 3.8, and find that Gene 4317 was selected by both SJS and Cox-ISIS, but not by Cox-SIS. From Tables 3.6, this gene is also included in models selected by SJS-SCAD, SJS-Lasso, Cox-ISIS-SCAD and Cox-ISIS-Lasso. This motivates further investigation of this variable. Table 3.9 presents likelihoods and AIC/BIC scores for models with and without Gene 4317. The P-values of the likelihood ratio tests indicate that Gene 4317 should be included in the models. This clearly indicates that Cox-SIS fails to identify this significant gene.

Table 3.8: Estimates and Standard Errors (SE) based on SJS-SCAD

Gene ID	Estimate(SE)	P-value	Gene ID	Estimate(SE)	P-value
1023	0.4690(0.1289)	2.74e-04	3821	-0.8668(0.5901)	0.142
1662	-0.7950(0.3388)	1.90e-02	3824	0.2176(0.0791)	5.97e-03
1664	1.3437(0.3227)	3.14e-05	4317	0.4471(0.1153)	1.05e-04
1682	0.3468(0.1464)	1.79e-02	4545	0.04761(0.0181)	8.23e-03
1825	0.7459(0.1306)	1.13e-08	4595	0.4751(0.0977)	1.16e-06
2115	-0.5097(0.1168)	1.29e-05	5668	-0.6518(0.1314)	6.99e-07
3332	-0.4340(0.1100)	8.00e-05	6156	-0.4751(0.1142)	3.19e-05
3373	0.1713(0.0608)	4.84e-03	6517	-0.0156(0.0068)	2.15e-02
3497	0.4417(0.1076)	4.06e-05	6607	0.6265(0.1196)	1.64e-07
3791	0.1260(0.0454)	5.59e-03	6758	-0.5383(0.1075)	5.64e-07
3810	1.2120(0.3697)	1.05e-03	6844	0.7052(0.1171)	1.72e-9
3811	-0.9292(0.3262)	4.39e-03	6908	-0.3667(0.1221)	2.68e-03
3818	0.7600(0.4598)	0.098	7343	-0.3411(0.1143)	2.84e-03
3819	1.1895(0.3824)	1.87e-03	7357	-0.8760(0.1152)	2.88e-14
3820	-2.0650(0.4843)	2.01e-05	7380	0.3791(0.1031)	2.37e-04

4. Discussions

In this paper, we propose a sure joint screening (SJS) procedure for feature

Table 3.9: Likelihood, AIC and BIC of Models with and without Gene 4317.

	SJS	SJS-SCAD	SJS-Lasso	ISIS	ISIS-SCAD	ISIS-Lasso
LKHD with Gene4317	-536.9838	-546.1902	-542.9862	-561.8795	-575.7148	-567.6035
LKHD w/o Gene4317	-544.1571	-549.4587	-547.8609	-568.8975	-580.2026	-572.1035
df	1	1	1	1	1	1
BIC w/o Gene4317	1317.617	1257.245	1286.807	1367.098	1296.895	1324.373
AIC w/o Gene4317	1172.314	1156.917	1165.722	1221.795	1210.405	1210.207
p-value of LRT	1.50e-04	0.0106	0.0018	1.70e-04	0.0027	0.0027

screening in the Cox model with ultrahigh dimensional covariates. The proposed SJS is distinguished from the existing Cox-SIS and Cox-ISIS in that SJS is based on joint likelihood of potential candidate features. We propose an effective algorithm to carry out the feature screening procedure, and show that the proposed algorithm possesses ascent property. We study the sampling property of SJS, and establish the sure screening property for SJS. We conduct Monte Carlo simulation to evaluate the finite sample performance of SJS and compare it with Cox-SIS and Cox-ISIS. Our numerical comparison indicates that SJS outperforms Cox-SIS and Cox-ISIS, and SJS can effectively screen out inactive covariates and retain truly active covariates. We further illustrate the proposed procedure using a real data example.

Theorem 1 ensures the ascent property of the proposed algorithm under certain conditions. However, it does not implies that the proposed algorithm converges to the global optimizer. If the proposed algorithm converges to a global maximizer of (2.3), then Theorem 2 shows that such a solution enjoys the sure screen property. In this paper, we simply set $m = [n/\log(n)]$ in our numerical study. It may be of interest to derive a data-driven method to determine m and reduce false positive rate in the screening stage. This would be a good topic for future research.

Acknowledgment. Yang's research was supported by National Nature Science Foundation of China grant 11471086 and the Fundamental Research Funds for the Central Universities 15JNQM019 and 21615452. Li's research was supported by NIDA, NIH grants P50 DA10075 and P50 DA036107. Buu's research was supported by NIH grants, K01 AA016591 and R01 DA035183. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA or the NIH. The authors are indebted to the referees, the associate editor and the Co-editor for their valuable comments, which have significantly improved the paper.

Appendix

We need the following notation to present the regularity conditions for the partial likelihood and the Cox model. Most notations are adapted from Andersen and Gill (1982), in which counting processes were introduced for the Cox model and the consistency and asymptotic normality of the partial likelihood estimate were established. Denote $\overline{N}_i(t) = I\{T_i \le t, T_i \le C_i\}$ and $R_i(t) = \{T_i \ge t, C_i \ge t\}$. Assume that there are no two component processes $N_i(t)$ jumping at the same time. For simplicity, we shall work on the finite interval $[0, \tau]$. In Cox's model, properties of stochastic processes, such as being a local martingale or a predictable process, are relative to a right-continuous nondecreasing family $(\mathcal{F}_t : t \in [0, \tau])$ of sub σ-algebras on a sample space $(\Omega, \mathcal{F}, \mathcal{P})$; $\mathcal{F}_t$ represents everything that happens up to time t. Throughout this section, we define $\Lambda_0(t) = \int_0^t h_0(u)\, du$.

By stating that $\overline{N}_i(t)$ has intensity process $h_i(t) \widehat{=} h(t|\mathbf{x}_i)$, we mean that the processes $M_i(t)$ defined by

$$M_i(t) = \overline{N}_i(t) - \int_0^t h_i(u) du, \quad i = 1, \ldots, n,$$

are local martingales on the time interval $[0, \tau]$.

Define

$$\mathbf{A}^{(k)}(\boldsymbol{\beta}, t) = \frac{1}{n} \sum_{i=1}^{n} R_i(t) \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\} \mathbf{x}_i^{\otimes k}, \quad \mathbf{a}^{(k)}(\boldsymbol{\beta}, t) = E[\mathbf{A}^{(k)}(\boldsymbol{\beta}, t)] \quad \text{for} \quad k = 0, 1, 2,$$

and

$$E(\boldsymbol{\beta}, t) = \frac{\mathbf{A}^{(1)}(\boldsymbol{\beta}, t)}{\mathbf{A}^{(0)}(\boldsymbol{\beta}, t)}, \quad V(\boldsymbol{\beta}, t) = \frac{\mathbf{A}^{(2)}(\boldsymbol{\beta}, t)}{\mathbf{A}^{(0)}(\boldsymbol{\beta}, t)} - E(\boldsymbol{\beta}, t)^{\otimes 2}.$$

where $\mathbf{x}_i^{\otimes 0} = 1$, $\mathbf{x}_i^{\otimes 1} = \mathbf{x}_i$ and $\mathbf{x}_i^{\otimes 2} = \mathbf{x}_i \mathbf{x}_i^T$. Note that $\mathbf{A}^{(0)}(\boldsymbol{\beta}, t)$ is a scalar, $\mathbf{A}^{(1)}(\boldsymbol{\beta}, t)$ and $E(\boldsymbol{\beta}, t)$ are p-vector, and $\mathbf{A}^{(2)}(\boldsymbol{\beta}, t)$ and $V(\boldsymbol{\beta}, t)$ are $p \times p$ matrices.

Define

$$Q_j = \sum_{i=1}^{n} \int_0^{t_j} \left[\mathbf{x}_i - \frac{\sum_{i \in R_j} \mathbf{x}_i \exp(\mathbf{x}_i^T \boldsymbol{\beta})}{\sum_{i \in R_j} \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \right] dM_i.$$

Here, $E[Q_j | \mathcal{F}_{j-1}] = Q_{j-1}$ i.e. $E[Q_j - Q_{j-1} | \mathcal{F}_{j-1}] = 0$. Let $b_j = Q_j - Q_{j-1}$, then $(b_j)_{j=1,2,\ldots}$ is a sequence of bounded martingale differences on $(\Omega, \mathcal{F}, P)$. That is, b_j is bounded almost surely (a.s.) and $E[b_j | \mathcal{F}_{j-1}] = 0$ a.s. for $j = 1, 2, \ldots$.

(D1) (Finite interval). $\Lambda_0(\tau) = \int_0^\tau h_0(t) dt < \infty$

(D2) (Asymptotic stability). There exists a neighborhood $\mathcal{B}$ of $\boldsymbol{\beta}^*$ and scalar, vector and matrix functions $\mathbf{a}^{(0)}$,$\mathbf{a}^{(1)}$ and $\mathbf{a}^{(2)}$ defined on $\mathcal{B} \times [0, \tau]$ such that for $k = 0, 1, 2$

$$\sup_{t\in[0,\tau],\boldsymbol{\beta}\in\mathcal{B}} \|\mathbf{A}^{(k)}(\boldsymbol{\beta}, t) - \mathbf{a}^{(k)}(\boldsymbol{\beta}, t)\| \xrightarrow{p} 0.$$

(D3) (Lindeberg condition). There exists $\delta > 0$ such that

$$n^{-1/2} \sup_{i,t} |\mathbf{x}_i| R_i(t) I\{\boldsymbol{\beta}_0' \mathbf{x}_i > -\delta|\mathbf{x}_i|\} \xrightarrow{p} 0,$$

(D4) (Asymptotic regularity conditions). Let $\mathcal{B}$, $\mathbf{a}^{(0)}$, $\mathbf{a}^{(1)}$ and $\mathbf{a}^{(2)}$ be as in Condition (D2) and define $e = \mathbf{a}^{(1)}/\mathbf{a}^{(0)}$ and $v = \mathbf{a}^{(2)}/\mathbf{a}^{(0)} - e^{\otimes 2}$. For all $\boldsymbol{\beta} \in \mathcal{B}$,$t \in [0, \tau]$;

$$\mathbf{a}^{(1)}(\boldsymbol{\beta}, t) = \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{a}^{(0)}(\boldsymbol{\beta}, t), \quad \mathbf{a}^{(2)}(\boldsymbol{\beta}, t) = \frac{\partial^2}{\partial \boldsymbol{\beta}^2} \mathbf{a}^{(0)}(\boldsymbol{\beta}, t),$$

$\mathbf{a}^{(0)}(\cdot, t)$, $\mathbf{a}^{(1)}(\cdot, t)$ and $\mathbf{a}^{(2)}(\cdot, t)$ are continuous functions of $\boldsymbol{\beta} \in \mathcal{B}$, uniformly in $t \in [0, \tau]$, $\mathbf{a}^{(0)}$, $\mathbf{a}^{(1)}$ and $\mathbf{a}^{(2)}$ are bounded on $\mathcal{B} \times [0, \tau]$; $\mathbf{a}^{(0)}$ is bounded away from zero on $\mathcal{B} \times [0, \tau]$, and the matrix

$$\mathbf{A} = \int_0^\tau v(\boldsymbol{\beta}_0, t) \mathbf{a}^{(0)}(\boldsymbol{\beta}_0, t) h_0(t) dt$$

is positive definite.

(D5) The function $\mathbf{A}^{(0)}(\boldsymbol{\beta}^*, t)$ and $\mathbf{a}^{(0)}(\boldsymbol{\beta}^*, t)$ are bounded away from 0 on $[0, \tau]$.

(D6) There exist constants $C_1, C_2 > 0$, such that $\max_{ij} |x_{ij}| < C_1$ and $\max_i |\mathbf{x}_i^T \boldsymbol{\beta}^*| < C_2$.

(D7) $\{b_j\}$ is a sequence of martingale differences and there exit nonnegative constants c_j such that for every real number t,

$$E\{\exp(tb_j)|\mathcal{F}_{j-1}\} \leq \exp(c_j^2 t^2/2) \quad a.s. \quad (j = 1, 2, \ldots, N)$$

For each j, the minimum of those c_j is denoted by $\eta(b_j)$.

$$|b_j| \leq K_j \quad a.s. \quad \text{for} \quad j = 1, 2, \ldots, N$$

and $E\{b_{j_1}, b_{j_2}, \ldots, b_{j_k}\} = 0$ for $b_{j_1} < b_{j_2} < \cdots < b_{j_k}$;$k = 1, 2, \ldots$.

Note that the partial derivative conditions on $\mathbf{a}^{(0)}$, $\mathbf{a}^{(1)}$ and $\mathbf{a}^{(2)}$ are satisfied by $\mathbf{A}^{(0)}$, $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$; and that $\mathbf{A}$ is automatically positive semidefinite. Furthermore the interval $[0, \tau]$ in the conditions may everywhere be replaced by the set $\{t : h_0(t) > 0\}$.

Condition (D1)—(D5) is a standard condition for the proportional hazards model (Anderson and Gill, 1982), which is weaker than the one required by Bradic et al (2011) and $\mathbf{A}^{(k)}(\boldsymbol{\beta}_0, t)$ converges uniformly to $\mathbf{a}^{(k)}(\boldsymbol{\beta}_0, t)$. Condition (D6) is a routine one, which is needed to apply the concentration inequality for general empirical processes. For example, the bounded covariate assumption is used by Huang et al. (2013) for discussing the Lasso estimator of proportional hazards models. Condition (D7) is needed for the asymptotic behavior of the score function $\ell_p'(\boldsymbol{\beta})$ of partial likelihood because the score function cannot be represented as a sum of independent random vectors, but it can be represented as sum of a sequence of martingale differences.

Proof of Theorem 1. Applying the Taylor expansion to $\ell_p(\boldsymbol{\gamma})$ at $\boldsymbol{\gamma} = \boldsymbol{\beta}$, it follows that

$$\ell_p(\boldsymbol{\gamma}) = \ell_p(\boldsymbol{\beta}) + \ell_p'(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta}) + \frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\beta})^T \ell_p''(\tilde{\boldsymbol{\beta}})(\boldsymbol{\gamma} - \boldsymbol{\beta}),$$

where $\tilde{\boldsymbol{\beta}}$ lies between $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$.

$$(\boldsymbol{\gamma}-\boldsymbol{\beta})^T\{-\ell_p''(\tilde{\boldsymbol{\beta}})\}(\boldsymbol{\gamma}-\boldsymbol{\beta}) \le (\boldsymbol{\gamma}-\boldsymbol{\beta})^T W(\boldsymbol{\beta})(\boldsymbol{\gamma}-\boldsymbol{\beta})\lambda_{\max}[W^{-1/2}(\boldsymbol{\beta})\{-\ell_p''(\tilde{\boldsymbol{\beta}})\}W^{-1/2}(\boldsymbol{\beta})]$$

Thus, if $u > \lambda_{\max}[W^{-1/2}(\boldsymbol{\beta})\{-\ell_p''(\tilde{\boldsymbol{\beta}})\}W^{-1/2}(\boldsymbol{\beta})] \ge 0$ since $-\ell_p''(\boldsymbol{\beta})$ is non-negative definite, then

$$\ell_p(\boldsymbol{\gamma}) \ge \ell_p(\boldsymbol{\beta}) + \ell_p'(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta}) - \frac{u}{2}(\boldsymbol{\gamma} - \boldsymbol{\beta})^T W(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta})$$

Thus it follows that $\ell_p(\boldsymbol{\gamma}) \ge g(\boldsymbol{\gamma}|\boldsymbol{\beta})$ and $\ell_p(\boldsymbol{\beta}) = g(\boldsymbol{\beta}|\boldsymbol{\beta})$ by definition of $g(\boldsymbol{\gamma}, \boldsymbol{\beta})$. Hence, under the conditions of Theorem 1, it follows that

$$\ell_p(\boldsymbol{\beta}_*^{(t+1)}) \ge g(\boldsymbol{\beta}_*^{(t+1)}|\boldsymbol{\beta}^{(t)}) \ge g(\boldsymbol{\beta}^{(t)}|\boldsymbol{\beta}^{(t)}) = \ell(\boldsymbol{\beta}^{(t)}).$$

The second inequality is due to the fact that $\|\boldsymbol{\beta}_*^{(t+1)}\|_0 = \|\boldsymbol{\beta}^{(t)}\|_0 = m$, and $\boldsymbol{\beta}_*^{(t+1)} = \arg\max_{\boldsymbol{\gamma}} g(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(t)})$ subject to $\|\boldsymbol{\gamma}\|_0 \le m$. By definition of $\boldsymbol{\beta}^{(t+1)}$, $\ell_p(\boldsymbol{\beta}^{(t+1)}) \ge \ell_p(\boldsymbol{\beta}_*^{(t+1)})$ and $\|\boldsymbol{\beta}^{(t+1)}\|_0 = m$. This proves Theorem 1.

Proof of Theorem 2. Let $\widehat{\boldsymbol{\beta}}_s$ be the partial likelihood estimate of $\boldsymbol{\beta}_s$ based on model s. The theorem is implied if $Pr\{\widehat{s} \in S_+^m\} \to 1$. Thus, it suffices to show that

$$Pr\left\{\max_{s\in S_-^m} \ell_p(\widehat{\boldsymbol{\beta}}_s) \ge \min_{s\in S_+^m} \ell_p(\widehat{\boldsymbol{\beta}}_s)\right\} \to 0,$$

as $n \to \infty$.

For any $s \in S_-^m$, define $s' = s \cup s^* \in S_+^{2m}$. Under (C1) condition, we consider

$\boldsymbol{\beta}_{s'}$ close to $\boldsymbol{\beta}^*_{s'}$ such that $\|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}\| = w_1 n^{-\tau_1}$ for some $w_1, \tau_1 > 0$. Clearly, when n is sufficiently large, $\boldsymbol{\beta}_{s'}$ falls into a small neighborhood of $\boldsymbol{\beta}^*_{s'}$, so that Condition (C3) becomes applicable. Thus, it follows Condition (C3) and the Cauchy-Schwarz inequality that

$$\begin{aligned}
\ell_p(\boldsymbol{\beta}_{s'}) - \ell_p(\boldsymbol{\beta}^*_{s'}) &= [\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}]^T \ell_p'(\boldsymbol{\beta}^*_{s'}) + (1/2)[\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}]^T \ell_p''(\tilde{\boldsymbol{\beta}}_{s'})[\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}] \\
&\leq [\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}]^T \ell_p'(\boldsymbol{\beta}^*_{s'}) - (c_1/2)n\|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}\|_2^2 \\
&\leq w_1 n^{-\tau_1}\|\ell_p'(\boldsymbol{\beta}^*_{s'})\|_2 - (c_1/2)w_1^2 n^{1-2\tau_1}, \qquad \text{(A.1)}
\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}_{s'}$ is an intermediate value between $\boldsymbol{\beta}_{s'}$ and $\boldsymbol{\beta}^*_{s'}$. Thus, we have

$$\begin{aligned}
Pr\{\ell_p(\boldsymbol{\beta}_{s'}) - \ell_p(\boldsymbol{\beta}^*_{s'}) \geq 0\} &\leq Pr\{\|\ell_p'(\boldsymbol{\beta}^*_{s'})\|_2 \geq (c_1 w_1/2)n^{1-\tau_1}\} \\
&= Pr\left\{\sum_{j\in s'}[\ell_j'(\boldsymbol{\beta}^*_{s'})]^2 \geq (c_1 w_1/2)^2 n^{2-2\tau_1}\right\} \\
&\leq \sum_{j\in s'} Pr\{[\ell_j'(\boldsymbol{\beta}^*_{s'})]^2 \geq (2m)^{-1}(c_1 w_1/2)^2 n^{2-2\tau_1}\}
\end{aligned}$$

Also, by (C1), we have $m \leq w_2 n^{\tau_2}$, and also the following probability inequality

$$\begin{aligned}
Pr\{\ell_j'(\boldsymbol{\beta}^*_{s'}) \geq (2m)^{-1/2}(c_1 w_1/2)n^{1-\tau_1}\} &\leq Pr\{\ell_j'(\boldsymbol{\beta}^*_{s'}) \geq (2w_2 n^{\tau_2})^{-1/2}(c_1 w_1/2)n^{1-\tau_1}\} \\
&= Pr\left\{\ell_j'(\boldsymbol{\beta}^*_{s'}) \geq c n^{1-\tau_1-0.5\tau_2}\right\} \\
&= Pr\left\{\ell_j'(\boldsymbol{\beta}^*_{s'}) \geq n c n^{-\tau_1-0.5\tau_2}\right\} \qquad \text{(A.2)}
\end{aligned}$$

where $c = c_1 w_1/(2\sqrt{2w_2})$ denotes some generic positive constant. Recall (2.2), by differentiation and rearrangement of terms, it can be shown as in Andersen and Gill (1982) that the gradient of $\ell_p(\boldsymbol{\beta})$ is

$$\ell_p'(\boldsymbol{\beta}) \equiv \frac{\partial \ell_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{1}{n}\sum_{i=1}^{n}\int_0^\infty [\mathbf{x}_i - \bar{\mathbf{x}}_n(\boldsymbol{\beta}, t)]\, d\overline{N}_i(t). \qquad \text{(A.3)}$$

where $\bar{\mathbf{x}}_n(\boldsymbol{\beta}, t) = \sum_{i\in R_j} \mathbf{x}_i \exp(\mathbf{x}_i^T\boldsymbol{\beta}) / \sum_{i\in R_j} \exp(\mathbf{x}_i^T\boldsymbol{\beta})$. As a result, the partial score function $\ell_p'(\boldsymbol{\beta})$ no longer has a martingale structure, and the large deviation results for continuous time martingale in Bradic et al (2011) and Huang et al (2013) are not directly applicable. The martingale process associated with $\overline{N}_i(t)$ is given by $M_i(t) = \overline{N}_i(t) - \int_0^t R_i(s)\exp(\mathbf{x}^T\boldsymbol{\beta}^*)d\Lambda_0(u)$.

Let t_j be the time of the jth jump of the process $\sum_{i=1}^n \int_0^\infty R_i(t) d\overline{N}_i(t)$,

$j = 1, \ldots, N$ and $t_0 = 0$. Then, t_j are stopping times. For $j = 0, 1, \ldots, N$, define

$$Q_j = \sum_{i=1}^{n} \int_0^{t_j} b_i(u) d\overline{N}_i(u) = \sum_{i=1}^{n} \int_0^{t_j} b_i(u) dM_i(u) \tag{A.4}$$

where $b_i(u) = \mathbf{x}_i - \bar{\mathbf{x}}_n(\boldsymbol{\beta}, u)$, $i = 1, \ldots, n$ are predictable, under no two component processes jumping at the same time and (D6), and satisfy $|b_i(u)| \leq 1$.

Since $M_i(u)$ are martingales and $b_i(u)$ are predictable, $\{Q_j, j = 0, 1, \ldots\}$ is a martingale with the difference $|Q_j - Q_{j-1}| \leq \max_{u,i} |b_i(u)| \leq 1$. Recall definition of N in Section 2, we define $C_0^2 n \leq N$, where C_0 is a constant. So, by the martingale version of the Hoeffding's inequality (Azuma, 1967) and under Condition (D7), we have

$$Pr(|Q_N| > nC_0 x) \leq 2\exp\{-n^2 C_0^2 x^2/(2N)\} \leq 2\exp(-nx^2/2) \tag{A.5}$$

By (A.4), $Q_N = n\ell_p'(\boldsymbol{\beta})$ if and only if $\sum_{i=1}^n \int_0^\infty R_i(t) d\overline{N}_i(t) \leq N$. Thus, the left-hand side of (3.15) in Lemma 3.3 of Huang *et al* (2013) is no greater than $Pr(|Q_N| > nC_0 x) \leq 2\exp(-nx^2/2)$.

So, (A.2) can be rewritten as follows.

$$Pr\left\{\ell_j'(\boldsymbol{\beta}_{s'}^*) \geq ncn^{-\tau_1 - 0.5\tau_2}\right\} \leq \exp\{-0.5nn^{-2\tau_1 - \tau_2}\} = \exp\{-0.5n^{1-2\tau_1-\tau_2}\} \tag{A.6}$$

Also, by the same arguments, we have

$$Pr\{\ell_j'(\boldsymbol{\beta}_{s'}^*) \leq -m^{-1/2}(c_1 w_1/2) n^{1-\tau_1}\} \leq \exp\{-0.5n^{1-2\tau_1-\tau_2}\} \tag{A.7}$$

The inequalities (A.6) and (A.7) imply that,

$$Pr\{\ell_p(\boldsymbol{\beta}_{s'}) \geq \ell_p(\boldsymbol{\beta}_{s'}^*)\} \leq 4m\exp\{-0.5n^{1-2\tau_1-\tau_2}\}$$

Consequently, by Bonferroni inequality and under conditions (C1) and (C2), we have

$$\begin{aligned}
Pr\left\{\max_{s \in S_-^m} \ell_p(\boldsymbol{\beta}_{s'}) \geq \ell_p(\boldsymbol{\beta}_{s'}^*)\right\} &\leq \sum_{s \in S_-^m} Pr\{\ell_p(\boldsymbol{\beta}_{s'}) \geq \ell_p(\boldsymbol{\beta}_{s'}^*)\} \\
&\leq 4mp^m \exp\{-0.5n^{1-2\tau_1-\tau_2}\} \\
&= 4\exp\{\log m + m\log p - 0.5n^{1-2\tau_1-\tau_2}\} \\
&\leq 4\exp\{\log w_2 + \tau_2 \log n + w_2 n^{\tau_2} \tilde{c} n^{\kappa} - 0.5n^{1-2\tau_1-\tau_2}\} \\
&= 4w_2 \exp\{\tau_2 \log n + w_2 \tilde{c} n^{\tau_2+\kappa} - 0.5n^{1-2\tau_1-\tau_2}\} \\
&= a_1 \exp\{\tau_2 \log n + a_2 n^{\tau_2+\kappa} - 0.5n^{1-2\tau_1-\tau_2}\} \\
&= o(1) \quad \text{as} \quad n \to \infty
\end{aligned} \tag{A.8}$$

for some generic positive constants $a_1 = 4w_2$ and $a_2 = w_2\tilde{c}$. By Condition (C3), $\ell_p(\boldsymbol{\beta}_{s'})$ is concave in $\boldsymbol{\beta}_{s'}$, (A.8) holds for any $\boldsymbol{\beta}_{s'}$ such that $\|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}\| \geq w_1 n^{-\tau_1}$.

For any $s \in S^m_-$, let $\breve{\boldsymbol{\beta}}_{s'}$ be $\widehat{\boldsymbol{\beta}}_s$ augmented with zeros corresponding to the elements in s'/s^* (i.e. $s' = \{s \cup (s^*/s)\} \cup (s'/s^*)$). By Condition (C1), it is seen that $\|\breve{\boldsymbol{\beta}}_{s'} - \boldsymbol{\beta}^*_{s'}\|_2 = \|\breve{\boldsymbol{\beta}}_{s^*\cup(s'/s^*)} - \boldsymbol{\beta}^*_{s^*\cup(s'/s^*)}\|_2 = \|\breve{\boldsymbol{\beta}}_{s^*\cup(s'/s^*)} - \boldsymbol{\beta}^*_{s^*}\|_2 \geq \|\boldsymbol{\beta}^*_{s^*\cup(s'/s^*)} - \boldsymbol{\beta}^*_{s^*}\|_2 \geq \|\boldsymbol{\beta}^*_{s'/s^*}\|_2 \geq w_1 n^{-\tau_1}$. Consequently,

$$Pr\left\{\max_{s\in S^m_-} \ell_p(\widehat{\boldsymbol{\beta}}_s) \geq \min_{s\in S^m_+} \ell_p(\widehat{\boldsymbol{\beta}}_s)\right\} \leq Pr\left\{\max_{s\in S^m_-} \ell_p(\breve{\boldsymbol{\beta}}_{s'}) \geq \ell_p(\boldsymbol{\beta}^*_{s'})\right\} = o(1)$$

The theorem is proved.

References

Andersen, P. K. and Gill, R. D. (1982). Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics*, **10**, 1033–1311.

Azuma, K. (1967). Weighted Sums of Certain Dependent Random Variables. *Tohoku Mathematical Journal*, **19**, 357–367.

Bradic, J., Fan, J. and Jiang, J. (2011). Regularization for Cox's Proportional Hazards Model with NP-dimensionality. *The Annals of Statistics*, **39**, 3092–3120.

Cox, D. R. (1972). Regression Models and Life Tables (with Discussion). *Journal of the Royal Statistical Society, Series B,* **34**, 187–220.

Cox, D. R. (1975). Partial Likelihood. *Biometrika,* **62**, 269–276.

Fan, J., Feng, Y. and Song, R. (2011). Nonparametric Independence Screening in Sparse Ultra-high Dimensional Additive Models. *Journal of the American Statistical Association*, **116**, 544-557.

Fan, J., Feng, Y. and Wu, Y. (2010). High-dimensional Variable Selection for Cox's Proportional Hazards Model. *IMS Collections, Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown*, **6**, 70–86.

Fan, J. and Li, R. (2002). Variable Selection for Cox's Proportional Hazards Model and Frailty Model. *The Annals of Statistics*, **30**, 74–99.

Fan, J. and Lv, J. (2008). Sure Independence Screening for Ultrahigh Dimensional Feature Space (with discussion). *Journal of the Royal Statistical Society, Series B*, **70**, 849–911.

Fan, J., Ma, Y. and Dai, W. (2014). Nonparametric Independence Screening in Sparse Ultra-High Dimensional Varying Coefficient Models *Journal of the American Statistical Association.* **109**, 1270 – 1284.

Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh Dimensional Feature Selection: Beyond the Linear Model. *Journal of Machine Learning Research*, **10**, 1829–1853.

Huang, J., Sun, T., Ying, Z., Yu, Y. and Zhang, C.-H. (2013). Oracle Inequalities for the Lasso in the Cox Model. *The Annals of Statistics*, **41**, 1142–1165.

Li, G., Peng, H., Zhang, J. and Zhu, L.-X. (2012). Robust Rank Correlation Based Screening. *The Annals of Statistics*, **40**, 1846–1877.

Liu, J., Li, R. and Wu, R. (2014). Feature Selection for Varying Coefficient Models with Ultrahigh Dimensional Covariates. *Journal of American Statistical Association*, **109**, 266 – 274.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Hermelink, H. K., Smeland, E. B. and Staudt, L. M. (2002). The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-cell Lymphoma. *The New England Journal of Medicine*, **346**, 1937–1947.

Tibshirani, R. (1997). The Lasso Method for Variable Selection in the Cox Model. *Statistics in Medicine*, **16**, 385–395.

Wang, H. (2009). Forward Regression for Ultra-high Dimensional Variable Screening. *Journal of the American Statistical Association*, **104**, 1512–1524.

Wang, H., Li, R. and Tsai, C.-L. (2007). Tuning Parameter Selectors For the Smoothly Clipped Absolute Deviation Method. *Biometrika*, **94**, 553–568.

Xu, C. and Chen, J. (2014). The Sparse MLE for Ultra-High-Dimensional Feature Screening. *Journal of the American Statistical Association*, **109**, 1257–1269.

Zhang, H. and Lu, W. (2007). Adaptive-LASSO for Cox's Proportional Hazards Model. *Biometrika*, **94**, 1–13.

Zhao, S. D. and Li, Y. (2012). Principled Sure Independence Screening for Cox Models with Ultra-High-Dimensional Covariates. *Journal of Multivariate Analysis,* **105**, 397–411.

Zou, H. (2008). A Note on Path-based Variable Selection in The Penalized Proportional Hazards Model. *Biometrika*, **95**, 241–247.

Guangren Yang
School of Economics, Jinan University, Guangzhou, P.R. China.
Email: tygr@jnu.edu.cn.

Ye Yu
Department of Statistics, The Pennsylvania State University,
University Park, PA 16802.
E-mail: ywy5092@psu.edu

Runze Li
Department of Statistics and The Methodology Center,
The Pennsylvania State University, University Park, PA 16802.
E-mail: rzli@psu.edu

Anne Buu
School of Nursing, University of Michigan, Ann Arbor, MI 48109, USA.
Email: buu@umich.edu