Statistica Sinica **26** (2016), 881-901
doi:http://dx.doi.org/10.5705/ss.2014.171

FEATURE SCREENING IN ULTRAHIGH DIMENSIONAL COX'S MODEL

Guangren Yang[1], Ye Yu[2], Runze Li[2] and Anne Buu[3]

[1]*Jinan University,* [2]*Pennsylvania State University and* [3]*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 existing sure independence screening (SIS) procedures (Fan, Feng, and Wu (2010); Zhao and Li (2012)) in that it is based on the 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 it out and establish its ascent property. We further prove that the proposed procedure possesses the sure screening property: with probability tending to one, the selected variable set includes the actual active predictors. We conducted Monte Carlo simulation to evaluate the finite sample performance of the proposed procedure and compare it with existing SIS procedures. The proposed methodology is also demonstrated through an empirical analysis of a data example.

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

1. Introduction

Modeling high-dimensional data has become most important research topic. Variable selection is fundamental in analysis of high-dimensional data. Feature screening procedures that can effectively reduce ultrahigh dimensionality become indispensable 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 marginal screening procedures may possesses a sure screening property under certain conditions. Such a 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)). It 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 settings in which the sample consists of independently and identically distributed observations from a population.

In many studies, survival data have primary outcomes or responses subject to censoring. 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. 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 large number of covariates such as genetic markers. 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. While, in practice, it may suffer from algorithm instability, statistical inaccuracy, and high computational cost when the dimension of covariate vector is much greater than the sample size. Feature screening can 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 its marginal partial likelihood estimate and selects a cutoff to control the false discovery rate.

We propose a new feature screening procedure for ultrahigh dimensional Cox models. It 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 potentially important features rather than the marginal partial likelihood of individual features. Non-marginal screening procedures have been demonstrated to have 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. Wang (2009) and Xu and Chen (2014) demonstrated that their approaches can perform significantly better than the SIS procedures under some scenarios. But their methods are for linear

and generalized linear models. In this paper, we show that our procedure can outperform the sure independence screening procedures for the Cox model.

We establish the screening property for the sure joint screening (SJS) procedure. Despite the fact that 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 a concentration inequality for the score function of partial likelihood.

We have conducted Monte Carlo simulation studies to assess the finite sample performance of the proposed procedure and compared 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 data example.

The rest of this paper is organized as follows. In Section 2, we propose a feature screening for the Cox model, and demonstrate the ascent property of our proposed algorithm to carry it out. 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 data example. Some discussion and concluding remarks are in Section 4. Technical proofs are 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, we consider the 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. Suppose the survival time is censored by the censoring time C, and write 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 in that, given $\mathbf{x}$, T and C are conditionally independent.

Suppose that $\{(\mathbf{x}_i, Z_i, \delta_i) : i = 1, \ldots, n\}$ is an independently and identically distributed random sample from (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 $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(N)}$. Denote the risk set right before the time t_j^0 by $R_j = \{i : Z_i \geq t_j^0\}$. 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 the effect of $\mathbf{x}$ is sparse, and the true value of $\boldsymbol{\beta}$ is $\boldsymbol{\beta}^*$. Sparsity implies that $\|\boldsymbol{\beta}^*\|_0$ is small, where $\|\mathbf{a}\|_0$ is the number of nonzero elements of $\mathbf{a}$. In the presence of ultrahigh dimensional covariates, one may consider reducing the ultrahigh dimensionality of $\boldsymbol{\beta}$ 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

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

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

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

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 large p and small n, $\ell_p''(\boldsymbol{\beta})$ is not invertible. Low computational costs are always desirable here. To deal with singularity of the Hessian matrix and save computational costs, we use an 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, we use $W = \operatorname{diag}\{-\ell_p''(\boldsymbol{\beta})\}$, the matrix consisting of the diagonal elements of $-\ell_p''(\boldsymbol{\beta})$. Thus we approximate $\ell_p''(\boldsymbol{\beta})$ by $u \operatorname{diag}\{\ell_p''(\boldsymbol{\beta})\}$.

Remark. Xu and Chen (2014) proposed a feature screening procedure by an 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 that in (2.4) if one set $W = I_p$, the $p \times p$ identity matrix, but our motivation indeed is different from theirs, and the working matrix W is not set to I_p.

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

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

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

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

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

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

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

Unlike the screening procedures based on marginal partial likelihood methods, our procedure iteratively updates $\boldsymbol{\beta}$ at Step 2. This enables the procedure to incorporate correlation information among the predictors through updating $\ell_p'(\boldsymbol{\beta})$ and $\ell_p''(\boldsymbol{\beta})$. Thus, our procedure should perform better than the marginal screening procedures when there are 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, it can be carried out at low computational cost. Based on our simulations, our procedures can be implemented with less computing time than the marginal screening procedure studied in Fan, Feng, and Wu (2010) and Zhao and Li (2012) in some scenarios (see Tables 2 and 3 for details).

Theorem 1. *Suppose that Conditions* (D1)−(D4) *in the Appendix hold, and let*

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

where $\lambda_{\max}(A)$ *stands for the maximal eigenvalue of a matrix* A. *If* $u_t \ge \rho^{(t)}$, *then* $\ell_p(\boldsymbol{\beta}^{(t+1)}) \ge \ell_p(\boldsymbol{\beta}^{(t)})$, *where* $\boldsymbol{\beta}^{(t+1)}$ *is defined in Step* 2b *of the algorithm.*

Theorem 1 asserts the ascent property of the proposed algorithm if u_t is appropriately chosen. That is, the algorithm can improve the current estimate within the feasible region, $\|\beta\|_0 \leq m$, and the resulting estimate in the current step can serve as a refinement of the last step. The result also provides some insight about the choice of u_t in practice settings. In our numerical studies, the algorithm typically converged within six iterations. Still, the algorithm is not guaranteed to converge to the global optimizer.

2.2. Sure screening property

For convenience of presentation, s denotes an arbitrary subset of $\{1, \ldots, p\}$, thus a submodel with covariates $\mathbf{x}_s = \{x_j, j \in s\}$ and associated coefficients $\boldsymbol{\beta}_s = \{\beta_j, j \in s\}$. We use $\tau(s)$ to indicate the size of model s, and 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 $\hat{s}$ such that $s^* \subset \hat{s}$ with a very high probability.

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

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

We need some additional notation. 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 $\tau(s^*) = q < m$. We take $S_+^m = \{s : s^* \subset s; \|s\|_0 \leq m\}$ and $S_-^m = \{s : s^* \not\subset s; \|s\|_0 \leq m\}$ as the collections of the over-fitted models and the under-fitted models. We investigate the asymptotic properties of $\hat{\boldsymbol{\beta}}_m$ under the scenario in which p, q, m, and $\boldsymbol{\beta}^*$ are allowed to depend on the sample size n. We impose the following conditions, some of which are technical and only serve to facilitate 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 $\tau_1 + \tau_2 < 1/2$ with $\min_{j \in s^*} |\beta_j^*| \geq w_1 n^{-\tau_1}$ and $q < m \leq w_2 n^{\tau_2}$.

(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) dictates the sparsity of $\boldsymbol{\beta}^*$ which makes the sure screening possible with $\tau(\hat{s}) = m > q$; it requires that the minimal component in $\boldsymbol{\beta}^*$ does not degenerate too fast, so that the signal is detectable in the asymptotic sequence.

Together with (C3), it sets an appropriate order of m that guarantees the identifiability of s^* over s for $\tau(s) \leq m$. Condition (C2) has p diverge with n at up to an exponential rate.

Theorem 2. *Suppose that Conditions* (C1)−(C3) *and Conditions* (D1)−(D7) *in the Appendix hold. If* $\hat{s}$ *is the model obtained by* (2.3) *as size* m*,*

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

The proof is given in the Appendix. One has to specify the value of m in practical implementation. In the literature, it is typical to set $m = [n/\log(n)]$ (Fan and Lv (2008)). Although an ad hoc choice, this 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. 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 the sure joint screening (SJS) procedure.

3. Numerical Studies

We evaluated the finite sample performance of the proposed feature screening procedure via Monte Carlo simulations, and applied it to a data set. All simulations were conducted by using R codes.

3.1. Simulation studies

We compared 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 procedure. In simulations, the predictor variable $\mathbf{x}$ was generated from a p-dimensional normal distribution with mean zero and covariance matrix $\Sigma = (\sigma_{ij})$. Two commonly-used covariance structures were adopted.

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

(S2) Σ has the autoregressive structure. $\sigma_{ij} = \rho^{|i-j|}$. Again with $\rho = 0.25$, 0.5, and 0.75.

We generated the censoring time from an exponential distribution with mean 10, and the survival time from the Cox model with $h_0(t) = 10$ and (b1) $\beta_1 = \beta_2 = \beta_3 = 5$, $\beta_4 = -15\rho$, and other β_js equal 0, or (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 $V_j \sim \mathcal{N}(0, 1)$.

Table 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

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

In our simulation, we considered the sample sizes $n = 100$ and 200, and the dimensions p=2,000 and 5,000. For each combination, we conducted 1,000 replicates of Monte Carlo simulation. We compared the performance of feature screening procedures using P_s: the proportion of times an individual active predictor was selected for a given model size m in the 1,000 replications, and P_a: the proportion of times all active predictors were selected for a given model size m in the 1,000 replications. We can expect P_s and P_a both be close to one when the estimated model size m is sufficiently large. We chose $m = [n/\log n]$ throughout our simulations.

It is expected that the performance of SJS depends on: the structure of the covariance matrix, the values of $\boldsymbol{\beta}$, the dimension of all candidate features and the sample size n. For survival data analysis, performance depends also on the censoring rate. Table 1 gives the censoring rates for the 12 combinations of covariance structure, the values of ρ and $\boldsymbol{\beta}$.

Table 2 reports $\mathcal{P}_s$ for the active predictors and $\mathcal{P}_a$ under S1. Table 2 also gives the average computing time for each replication. Under S1 and (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. From Table 2, Cox-SIS fails to identify X_4 as an active predictor completely under (b1). This is expected. The SJS procedure, on the other hand, includes X_4 with nearly always. In addition, SJS has $\mathcal{P}_a$ very close to one for every case under (b1). Thus, SJS outperforms Cox-SIS easily in this setting.

We next consider the performance under S1 and (b2). In this setting, there is no predictor that is marginally independent of, but jointly dependent with the response. Table 2 clearly shows how performances are affected by sample size, dimension of predictors, and ρ. Overall, SJS outperforms Cox-SIS in all cases in terms of $\mathcal{P}_s$ and $\mathcal{P}_a$. The margin is quite significant when the sample size is small ($n = 100$) or when $\rho = 0.75$. The performance of SJS becomes better as the sample size increases.

Table 2. The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ with $\Sigma = (1-\rho)I + \rho\mathbf{1}\mathbf{1}^T$.

		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 $p = 2,000$											
0.25	b1	0.984	0.991	0.991	0	0	13.07	0.999	0.995	0.997	0.981	0.975	7.54
	b2	0.826	0.817	0.826	0.842	0.437	12.94	0.993	0.992	0.993	0.997	0.984	7.81
0.50	b1	0.951	0.948	0.937	0.001	0.001	13.07	0.961	0.962	0.962	0.983	0.937	8.31
	b2	0.73	0.707	0.707	0.734	0.236	12.95	0.981	0.976	0.977	0.976	0.936	8.47
0.75	b1	0.761	0.783	0.775	0.008	0.005	12.09	0.954	0.943	0.942	0.987	0.898	8.36
	b2	0.611	0.638	0.619	0.620	0.134	9.22	0.887	0.891	0.900	0.898	0.717	6.07
		$n = 100$ and $p = 5,000$											
0.25	b1	0.977	0.975	0.981	0	0	32.00	0.988	0.981	0.984	0.925	0.912	26.00
	b2	0.739	0.788	0.763	0.769	0.317	27.76	0.972	0.974	0.978	0.975	0.938	54.98
0.50	b1	0.892	0.900	0.894	0	0	42.82	0.871	0.861	0.862	0.948	0.805	31.89
	b2	0.636	0.619	0.643	0.629	0.127	28.25	0.919	0.922	0.934	0.923	0.812	59.68
0.75	b1	0.701	0.696	0.659	0.008	0.002	30.94	0.829	0.838	0.828	0.988	0.724	36.73
	b2	0.501	0.501	0.488	0.472	0.045	25.90	0.780	0.799	0.784	0.783	0.486	49.65
		$n = 200$ and $p = 2,000$											
0.25	b1	1	1	1	0	0	15.90	1	1	1	1	1	16.32
	b2	0.977	0.971	0.979	0.964	0.897	6.99	1	1	1	1	1	5.94
0.50	b1	0.999	1	1	0	0	12.20	1	1	1	1	1	12.54
	b2	0.950	0.946	0.932	0.942	0.786	16.29	1	1	1	1	1	16.46
0.75	b1	0.989	0.990	0.994	0.001	0.001	15.79	1	1	1	1	1	17.70
	b2	0.887	0.873	0.883	0.909	0.597	18.34	1	0.998	1	1	0.998	20.33
		$n = 200$ and $p = 5,000$											
0.25	b1	1	1	1	0	0	34.32	1	1	1	1	1	160.33
	b2	0.952	0.962	0.949	0.958	0.825	42.47	1	1	1	1	1	211.99
0.50	b1	0.999	0.998	1	0	0	32.71	1	1	1	1	1	181.90
	b2	0.904	0.903	0.892	0.885	0.637	30.38	1	1	1	1	1	152.62
0.75	b1	0.978	0.976	0.985	0.004	0.004	34.83	1	1	1	0.999	0.999	218.22
	b2	0.823	0.832	0.832	0.812	0.431	28.40	0.998	0.999	0.997	0.999	0.993	146.69

Table 2 also has the performance of Cox-SIS better as the sample size increases, the feature dimension decreases or ρ decreases. Still, these factors have less impact on the performance of SJS. In terms of computing time, SJS and Cox-SIS are comparable. For $p = 2,000$, SJS needs slightly less computing time than Cox-SIS, while SJS needs more for $p = 5,000$.

Table 3 gives the simulation results for S2. Here with (b1) or (b2), none of the active predictors $X_1, \ldots, X_4$ is marginally independent of the survival time. Thus, one expects the Cox-SIS to work well for (b1) and (b2). Table 3 indicates that both Cox-SIS and SJS perform well under (b2). On the other hand, the Cox-SIS has low $\mathcal{P}_a$ when $n = 100$ under (b1), although $\mathcal{P}_a$ is much higher

Table 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 $p = 2,000$											
0.25	b1	1	1	0.997	0.183	0.182	10.46	1	1	1	0.989	0.989	5.84
	b2	0.989	1	0.999	0.983	0.971	10.60	1	1	1	1	1	5.55
0.50	b1	1	1	0.941	0.446	0.394	10.61	0.998	0.997	0.936	0.970	0.931	5.91
	b2	1	1	1	0.999	0.999	12.29	1	1	1	1	1	6.31
0.75	b1	1	1	0.525	0.364	0.048	6.57	0.985	0.927	0.641	0.907	0.615	3.77
	b2	1	1	1	1	1	10.71	1	1	1	1	1	5.47
		$n = 100$ and $p = 5,000$											
0.25	b1	1	1	0.991	0.135	0.131	32.23	1	1	1	0.965	0.965	59.62
	b2	0.981	0.999	1	0.975	0.955	40.31	0.999	1	1	0.999	0.999	74.80
0.50	b1	1	1	0.888	0.296	0.214	38.82	0.992	0.981	0.821	0.896	0.811	70.76
	b2	0.999	1	1	0.999	0.998	42.13	1	1	1	1	1	71.58
0.75	b1	1	1	0.439	0.23	0.019	29.09	0.959	0.82	0.449	0.783	0.415	53.55
	b2	1	1	1	1	1	31.05	1	1	1	1	1	52.37
		$n = 200$ and $p = 2,000$											
0.25	b1	1	1	1	0.592	0.592	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	0.999	0.869	0.868	12.96	1	1	1	1	1	10.47
	b2	1	1	1	1	1	12.78	1	1	1	1	1	11.39
0.75	b1	1	1	0.921	0.757	0.678	12.91	1	1	0.999	0.999	0.998	11.17
	b2	1	1	1	1	1	14.26	1	1	1	1	1	12.39
		$n = 200$ and $p = 5,000$											
0.25	b1	1	1	1	0.450	0.450	37.59	1	1	1	1	1	192.79
	b2	1	1	1	1	1	35.63	1	1	1	1	1	166.09
0.50	b1	1	1	1	0.790	0.790	38.47	1	1	1	1	1	166.29
	b2	1	1	1	1	1	27.90	1	1	1	1	1	132.96
0.75	b1	1	1	0.880	0.674	0.554	47.62	1	1	0.993	0.997	0.991	235.95
	b2	1	1	1	1	1	34.52	1	1	1	1	1	163.85

$n = 200$. In summary, SJS outperforms Cox-SIS throughout Table 3. In terms of computing time, the pattern is similar to that in Table 2.

We compared SJS with the iterative Cox-SIS. Table 2 indicates that Cox-SIS fails to identify the active predictor X_4 under S1 and (b1) because this setting has X_4 jointly dependent but marginally independent of the survival time. Fan, Feng, and Wu (2010) proposed iterative SIS for the Cox model (abbreviated as Cox-ISIS), and we compared our procedure with the Cox-ISIS. We did simulations under S1, (b1) with $n = 100$. We also investigate the impact of signal strength on the performance of our procedure by considering $\beta_1 = \beta_2 = \beta_3 = 5\tau$, $\beta_4 = -15\tau\rho$, and all other $\beta_j = 0$. We took $\tau = 1$, 0.75, 0.5, and 0.25. To make a fair

Table 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)
		$\tau = 1$											
2,000	0.25	0.998	0.998	0.999	1	0.996	23.34	0.999	0.996	0.995	0.979	0.975	5.75
	0.5	0.898	0.894	0.897	1	0.708	21.47	0.970	0.968	0.975	0.983	0.952	6.05
	0.75	0.697	0.696	0.694	1	0.303	19.03	0.952	0.949	0.953	0.993	0.903	5.72
5,000	0.25	0.998	0.994	0.999	0.992	0.983	36.47	0.988	0.981	0.984	0.925	0.912	26.00
	0.5	0.819	0.833	0.853	1	0.562	37.81	0.871	0.861	0.862	0.948	0.805	31.89
	0.75	0.579	0.583	0.611	1	0.177	38.81	0.829	0.838	0.828	0.988	0.724	36.73
		$\tau = 0.75$											
2,000	0.25	1	0.997	1	0.999	0.996	14.19	0.999	0.998	1	0.980	0.978	3.85
	0.5	0.896	0.899	0.904	1	0.712	14.10	0.970	0.969	0.970	0.987	0.952	4.47
	0.75	0.709	0.687	0.724	1	0.334	22.99	0.936	0.938	0.942	0.990	0.882	7.33
5,000	0.25	0.991	0.996	0.990	0.990	0.972	42.64	0.983	0.985	0.988	0.931	0.914	52.50
	0.5	0.840	0.823	0.844	1	0.563	44.85	0.895	0.89	0.896	0.956	0.848	43.96
	0.75	0.566	0.584	0.555	1	0.167	50.80	0.832	0.819	0.836	0.985	0.7	55.27
		$\tau = 0.5$											
2,000	0.25	0.997	0.997	0.999	1	0.994	14.45	1	0.997	0.998	0.981	0.978	3.99
	0.5	0.891	0.888	0.899	1	0.702	26.78	0.957	0.962	0.963	0.987	0.943	8.81
	0.75	0.672	0.678	0.665	1	0.273	13.95	0.883	0.889	0.889	0.990	0.772	4.79
5,000	0.25	0.993	0.995	0.990	0.993	0.975	41.41	0.977	0.983	0.989	0.912	0.897	34.82
	0.5	0.806	0.847	0.805	1	0.527	56.10	0.874	0.867	0.855	0.946	0.803	57.31
	0.75	0.560	0.574	0.544	1	0.161	40.54	0.738	0.761	0.746	0.975	0.564	61.49
		$\tau = 0.25$											
2,000	0.25	0.970	0.972	0.976	0.973	0.902	14.40	0.971	0.971	0.981	0.853	0.824	3.72
	0.5	0.822	0.806	0.819	1	0.534	14.45	0.866	0.845	0.833	0.966	0.748	5.00
	0.75	0.528	0.536	0.526	1	0.126	14.48	0.552	0.566	0.564	0.952	0.238	4.72
5,000	0.25	0.941	0.936	0.934	0.949	0.805	43.85	0.901	0.914	0.897	0.675	0.592	59.46
	0.5	0.731	0.736	0.709	0.999	0.366	45.25	0.664	0.671	0.645	0.860	0.475	50.66
	0.75	0.466	0.432	0.419	1	0.067	49.79	0.427	0.389	0.372	0.958	0.1	118.30

comparison, the Cox-ISIS was implemented by iterating Cox-SIS twice (each with the size $m/2$) so that the number of included predictors was $m = [n/\log(n)] = 22$ for both Cox-SIS and the SJS.

The simulation results are summarized in Table 4, in which we also report the computing time consumed the procedures. Table 4 indicates that when $\rho = 0.25$ is small, both Cox-ISIS and SJS work quite well, while SJS takes less time. When $\rho = 0.5$ and 0.75, SJS significantly outperforms Cox-ISIS. SJS has less computing time than Cox-ISIS when $p = 2,000$, and is comparable in computing time to Cox-ISIS when $p = 5,000$.

Table 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	

3.2. An application

We applied the proposed feature screening procedure in an 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, with a survival rate of only about 35 to 40 percent after standard chemotherapy, there is interest in understanding the genetic markers that may have impacts on survival.

The data set consists of the survival time of $n = 240$ DLBCL patients after chemotherapy, with $p = 7{,}399$ cDNA microarray expressions of each individual patient as predictors. Given such a large number of predictors and the small sample size, feature screening is a necessary initial step for a statistical modeling procedure that cannot deal with high dimensional survival data. All predictors were standardized so that they had mean zero and variance one.

Five patients had survival times being close to 0. After removing them, our analysis in this example is based on the sample of 235 patients. Cox-SIS, Cox-ISIS, and SJS were all applied to the data to obtain a reduced model with $[235/\log(235)] = 43$ genes. The IDs of genes selected by the three screening procedures are listed in Table 5. The maximum of partial likelihood function of the three corresponding models obtained by SJS, Cox-ISIS, and Cox-SIS procedures were -536.9838, -561.8795, and -600.0885, respectively. This has both SJS and Cox-ISIS performing much better than Cox-SIS, with SJS performing the best.

We applied penalized partial likelihood with the L_1 penalty (Tibshirani (1997)) and with SCAD (Fan and Li (2002)) for the models obtained from the screening stage, Lasso and SCAD for short. The tuning parameters in the SCAD

Table 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 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

and the Lasso were 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 6. The likelihood, the degree of freedom (df), the BIC score, and the AIC score of the resulting models are listed in Table 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 was 18.41286 with df=13. The P-value of this partial likelihood ratio test was 0.142. This favors the model selected by SJS-SCAD, compared with the one obtained in the screening stage. The resulting estimates and standard errors of the model selected by SJS-SCAD are in Table 8, which indicates that most selected genes have significant impact on the survival time. Comparing Tables 5 and 8, we find that Gene 4317 was selected by both SJS and Cox-ISIS, but not by Cox-SIS. From Tables 6, this gene was also included in models selected by SJS-SCAD, SJS-Lasso, Cox-ISIS-SCAD and Cox-

Table 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

Table 9. Likelihood, AIC and BIC of Models with and without Gene 4,317.

	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

ISIS-Lasso, suggesting investigation of this variable. Table 5 presents likelihoods and AIC/BIC scores for models with and without Gene 4,317. The P-values of the likelihood ratio tests indicate that Gene 4,317 should be included in the models, while Cox-SIS fails to identify it.

4. Discussions

We have proposed a sure joint screening (SJS) procedure for feature 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 the 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 an ascent property. We study the sampling property of SJS, and establish the sure screening property for SJS.

Theorem 1 ensures the ascent property of the proposed algorithm under certain conditions, but 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. We have simply set $m = [n/\log(n)]$ in our numerical studies. It is of interest to derive a data-driven method to determine m and reduce false positive rate in the screening stage.

Acknowledgement

Yang's research was supported by the National Nature Science Foundation of China grant 11471086, the Fundamental Research Funds for the Central Universities 15JNQM019 and 21615452, the National Statistical Scientific Research Center Projects 2015LD02 and the China Scholarship Council 201506785010. Li is the corresponding author and his research was supported by NIDA, NIH grants P50 DA 10075, P50 DA039838, and P50 DA036107, and a NSF grant DMS 1512422. 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, the NIH or the NSF. The authors are indebted to the referees, an associate editor and the Co-editor for their valuable comments, that have significantly improved the paper.

Appendix

Most of our notation is adapted from Andersen and Gill (1982), where counting processes were introduced for the Cox model, and consistency and asymptotic normality of the partial likelihood estimate were established. Let $\overline{N}_i(t) = I\{T_i \le t, T_i \le C_i\}$ and $R_i(t) = \{T_i \ge t, C_i \ge t\}$. Assume no two component processes $N_i(t)$ jump at the same time. For simplicity, we work on the interval $[0, \tau]$. In Cox's model, such 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 $(\Omega, \mathcal{F}, \mathcal{P})$; $\mathcal{F}_t$ to encompass everything that happens up to time t. Take $\Lambda_0(t) = \int_0^t h_0(u)\, du$.

By stating that $\overline{N}_i(t)$ has intensity process $h_i(t)\hat{=}h(t|\mathbf{x}_i)$, 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 $[0, \tau]$.

Let

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

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

where $\mathbf{x}_i^{\otimes 0} = 1$, $\mathbf{x}_i^{\otimes 1} = \mathbf{x}_i$, and $\mathbf{x}_i^{\otimes 2} = \mathbf{x}_i\mathbf{x}_i^T$. Here $\mathbf{A}^{(0)}(\boldsymbol{\beta}, t)$ 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 $E[Q_j|\mathcal{F}_{j-1}] = Q_{j-1}$ or $E[Q_j - Q_{j-1}|\mathcal{F}_{j-1}] = 0$. If $b_j = Q_j - Q_{j-1}$, then $(b_j)_{j=1,2,\ldots}$ is a sequence of bounded martingale differences on $(\Omega, \mathcal{F}, P)$, so 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 take $e = \mathbf{a}^{(1)}/\mathbf{a}^{(0)}$ 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 $\mathbf{A}$ is positive definite with

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

(D5) The functions $\mathbf{A}^{(0)}(\boldsymbol{\beta}^*, t)$ 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,

$$E\{\exp(tb_j) \mid \mathcal{F}_{j-1}\} \le \exp\left(\frac{c_j^2 t^2}{2}\right) \quad a.s. \quad (j = 1, 2, \ldots, N).$$

For each j, the minimum of those c_j is denoted by $\eta(b_j)$. $|b_j| \le K_j$ a.s. for $j = 1, \ldots, N$ and $E\{b_{j_1}, b_{j_2}, \ldots, b_{j_k}\} = 0$ for $b_{j_1} < b_{j_2} < \cdots < b_{j_k}$;$k = 1, 2, \ldots$.

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 $\mathbf{A}$ is automatically positive semidefinite. Furthermore the interval $[0, \tau]$ in the conditions may everywhere be replaced by the set $\{t : h_0(t) > 0\}$.

Conditions (D1)—(D5) are standard for the proportional hazards model (Anderson and Gill (1982)); they are weaker than those required by Bradic, Fan, and Jiang (2011) and $\mathbf{A}^{(k)}(\boldsymbol{\beta}_0, t)$ converges uniformly to $\mathbf{a}^{(k)}(\boldsymbol{\beta}_0, t)$. Condition (D6) is routine, needed to apply the concentration inequality for general empirical processes. 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.

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

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

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

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

Thus if $u > \lambda_{\max}[W^{-1/2}(\boldsymbol{\beta})\{-\ell_p''(\tilde{\boldsymbol{\beta}})\}W^{-1/2}(\boldsymbol{\beta})]$, non-negative 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 $\hat{\boldsymbol{\beta}}_s$ be the partial likelihood estimate of $\boldsymbol{\beta}_s$ based on model s. The theorem follows if $Pr\{\hat{s} \in S_+^m\} \to 1$. Thus, it suffices to show that, as $n \to \infty$,

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

For any $s \in S_-^m$, define $s' = s \cup s^* \in S_+^{2m}$. Under (C1), 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 (C3) becomes applicable. Thus, it follows by (C3) and the Cauchy-Schwarz inequality that

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

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

$$\begin{aligned} Pr\{\ell_p(\boldsymbol{\beta}_{s'}) - \ell_p(\boldsymbol{\beta}^*_{s'}) \ge 0\} &\le Pr\Big\{ \|\ell'_p(\boldsymbol{\beta}^*_{s'})\|_2 \ge \Big(\frac{c_1 w_1}{2}\Big) n^{1-\tau_1} \Big\} \\ &= Pr\Big\{ \sum_{j \in s'} [\ell'_j(\boldsymbol{\beta}^*_{s'})]^2 \ge \Big(\frac{c_1 w_1}{2}\Big)^2 n^{2-2\tau_1} \Big\} \\ &\le \sum_{j \in s'} Pr\Big\{ [\ell'_j(\boldsymbol{\beta}^*_{s'})]^2 \ge (2m)^{-1} \Big(\frac{c_1 w_1}{2}\Big)^2 n^{2-2\tau_1} \Big\}. \end{aligned}$$

Also, by (C1), we have $m \le w_2 n^{\tau_2}$, and

$$\begin{aligned} &Pr\Big\{ \ell'_j(\boldsymbol{\beta}^*_{s'}) \ge (2m)^{-1/2} \Big(\frac{c_1 w_1}{2}\Big) n^{1-\tau_1} \Big\} \\ &\quad \le Pr\Big\{ \ell'_j(\boldsymbol{\beta}^*_{s'}) \ge (2 w_2 n^{\tau_2})^{-1/2} \Big(\frac{c_1 w_1}{2}\Big) n^{1-\tau_1} \Big\} \\ &\quad = Pr\left\{ \ell'_j(\boldsymbol{\beta}^*_{s'}) \ge c n^{1-\tau_1-0.5\tau_2} \right\} \\ &\quad = Pr\left\{ \ell'_j(\boldsymbol{\beta}^*_{s'}) \ge n c n^{-\tau_1-0.5\tau_2} \right\}, \end{aligned} \tag{A.2}$$

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

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

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, Fan, and Jiang (2011) and Huang et al. (2013) are not directly applicable. The martingale process associated with $\overline{N}_i(t)$ is $M_i(t) = \overline{N}_i(t) - \int_0^t R_i(s) \exp(\mathbf{x}^T \boldsymbol{\beta}^*) d\Lambda_0(u)$.

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 $t_0 = 0$. The 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 $|b_i(u)| \le 1$.

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

$$Pr(|Q_N| > nC_0x) \le 2\exp\Big\{-\frac{n^2C_0^2x^2}{2N}\Big\} \le 2\exp\Big(-\frac{nx^2}{2}\Big). \tag{A.5}$$

By (A.4), $Q_N = n\ell_p'(\boldsymbol{\beta})$ if and only if $\sum_{i=1}^n \int_0^\infty R_i(t)d\overline{N}_i(t) \le N$. Thus, the left-hand side of (3.15) in Lemma 3.3 of Huang et al. (2013) is no greater than $Pr(|Q_N| > nC_0x) \le 2\exp(-nx^2/2)$.

Now (A.2) can be rewritten as.

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

By the same arguments, we have

$$Pr\Big\{\ell_j'(\boldsymbol{\beta}_{s'}^*) \le -m^{-1/2}\Big(\frac{c_1w_1}{2}\Big)n^{1-\tau_1}\Big\} \le \exp\{-0.5n^{1-2\tau_1-\tau_2}\}. \tag{A.7}$$

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

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

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

$$\begin{aligned}
Pr\left\{\max_{s\in S_-^m} \ell_p(\boldsymbol{\beta}_{s'}) \ge \ell_p(\boldsymbol{\beta}_{s'}^*)\right\} &\le \sum_{s\in S_-^m} Pr\{\ell_p(\boldsymbol{\beta}_{s'}) \ge \ell_p(\boldsymbol{\beta}_{s'}^*)\} \\
&\le 4mp^m\exp\{-0.5n^{1-2\tau_1-\tau_2}\} \\
&= 4\exp\{\log m + m\log p - 0.5n^{1-2\tau_1-\tau_2}\} \\
&\le 4\exp\{\log w_2 + \tau_2\log n + w_2n^{\tau_2}\tilde{c}n^{\kappa} - 0.5n^{1-2\tau_1-\tau_2}\} \\
&= 4w_2\exp\{\tau_2\log n + w_2\tilde{c}n^{\tau_2+\kappa} - 0.5n^{1-2\tau_1-\tau_2}\} \\
&= a_1\exp\{\tau_2\log n + a_2n^{\tau_2+\kappa} - 0.5n^{1-2\tau_1-\tau_2}\} \\
&= o(1) \quad \text{as} \quad n\to\infty
\end{aligned} \tag{A.8}$$

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 $\|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s'}^*\| \ge w_1n^{-\tau_1}$.

For any $s \in S_-^m$, let $\breve{\boldsymbol{\beta}}_{s'}$ be $\hat{\boldsymbol{\beta}}_s$ augmented with zeros corresponding to the elements in s'/s^*, $s' = \{s\cup(s^*/s)\}\cup(s'/s^*)$. By (C1), $\|\breve{\boldsymbol{\beta}}_{s'} - \boldsymbol{\beta}_{s'}^*\|_2 = \|\breve{\boldsymbol{\beta}}_{s^*\cup(s'/s^*)} -$

$\boldsymbol{\beta}^*_{s^*\cup(s'/s^*)}\|_2 = \|\breve{\boldsymbol{\beta}}_{s^*\cup(s'/s^*)} - \boldsymbol{\beta}^*_{s^*}\|_2 \geq \|\boldsymbol{\beta}^*_{s^*\cup(s'/s^*)} - \boldsymbol{\beta}^*_{s^*}\|_2 \geq \|\boldsymbol{\beta}^*_{s'/s^*}\|_2 \geq w_1 n^{-\tau_1}$.
Consequently,

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

Theorem is proved.

References

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10**, 1033-1311.

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Math. J.* **19**, 357-367.

Bradic, J., Fan, J. and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann. Statist.* **39**, 3092-3120.

Cox, D. R. (1972). Regression models and life tables (with Discussion). *J. Roy. Statist. Soc. Ser. 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. *J. Amer. Statist. Assoc.* **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. *Ann. Statist.* **30**, 74-99.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. Roy. Statist. Soc. Ser. B* **70**, 849-911.

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

Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *J. Machine Learn. Res.* **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. *Ann. Statist.* **41**, 1142-1165.

Li, G., Peng, H., Zhang, J. and Zhu, L.-X. (2012). Robust rank correlation based screening. *Ann. Statist.* **40**, 1846-1877.

Liu, J., Li, R. and Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh dimensional covariates. *J. Amer. Statist. Assoc.* **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 J. Medicine* **346**, 1937-1947.

Tibshirani, R. (1997). The Lasso method for variable selection in the Cox Model. *Statist. Medicine* **16**, 385-395.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *J. Amer. Statist. Assoc.* **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. *J. Amer. Statist. Assoc.* **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. *J. Multivariate Anal.* **105**, 397-411.

Zou, H. (2008). A note on path-based variable selection in the penalized proportional hazards model. *Biometrika* **95**, 241-247.

School of Economics, Jinan University, Guangzhou, P. R. China.

E-mail: tygr@jnu.edu.cn

Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA.

E-mail: ywy5092@psu.edu

Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802, USA.

E-mail: rzli@psu.edu

School of Nursing, University of Michigan, Ann Arbor, MI 48109, USA.

E-mail: buu@umich.edu

(Received May 2014; accepted August 2015)