The Pennsylvania State University
The Graduate School
Eberly College of Science

A NEW VARIABLE SCREENING PROCEDURE FOR COX'S MODEL

A Thesis in
Statistics
by
Ye YU

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December 2014

The thesis of Ye YU was reviewed and approved* by the following:

Runze Li
Distinguished Professor, Department of Statistics
Thesis Advisor

Aleksandra Slavkovic
Associate Professor, Department of Statistics
Associate Head for Graduate Studies

David Hunter
Professor, Department of Statistics
Department Head

*Signatures are on file in the Graduate School.

Abstract

Survival data with ultrahigh dimensional covariates such as genetic markers have been collected in medical studies and other fields. In this thesis, 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 also 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.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am very grateful to my thesis advisor, Dr. Runze Li, for his guidance in shedding light on the right track of my research. His encouragement helps me to become self-confident and active when I strike a sticky path. His mentorship is so treasurable in providing a well rounded experience consistent with my long-term career goals. Most importantly, his friendship during my graduate studies at Pennsylvania State University is priceless for me in my whole life.

I would like to thank the Department of Statistics, and my thesis reviewers, Dr. David Hunter and Dr. Aleksandra Slavkovic for their input, valuable discussions and accessibility. Without their knowledge and assistance this study would not have been successful. Additionally, I am very grateful for the friendship of all of the classmates, with whom I work together, puzzle over many problems, and enjoy the graduate life.

Finally, I would like to thank my parents, Mr. Yongjian Yu and Ms. Shaoping Hong for their support, encouragement, and unwavering love.

This dissertation research is supported by NIDA, NIH grant P50 DA10075. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA or the NIH.

Chapter 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, Peng, Zhang and Zhu, 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. One common issue with the aforementioned screening methods is that the ranking utility is a marginal one and therefore we need to iteratively apply screening procedures in order to enhance the finite sample performance. In other words, these screening procedures have a great chance to neglect important predictors that relates to responses jointly but not independently without iteration. To overcome this issue, Xu and Chen (2014) proposed a feature screening procedure by iterative hard-thresholding algorithm (IHT) for linear and generalized linear models and also establish the sure screening property for IHT.

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 (Tishirani, 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. However, both screening procedures have a great chance to neglect important predictors that relates to responses jointly but not independently without iteratively repeating the procedures.

In this thesis work, 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 work, we will show that the newly proposed procedure can outperform the sure independence screening procedure for the Cox model. This work makes the major contribution to the literature in that (a) we propose a sure joint screening (SJS) procedure for ultrahigh dimensional Cox model; (b) we further propose an effective algorithm to carry out the pro-

posed screening procedure, and demonstrate the ascent property of the proposed algorithm.

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 the thesis is organized as follows. In Chapter 2, we give a detailed review of the existing methods in literature. In Chapter 3, 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. Numerical comparisons and an empirical analysis of a real data example are then presented. Some discussion and conclusion remarks are given in Chapter 4.

Chapter 2

Literature Review

The review of literature is organized as follows. First, we introduce variable selection methods for linear model via penalized least squares. We then present a brief review of some basic concepts in survival data analysis together with some commonly-used survival models. Existing variable selection and feature screening procedures for survival data analysis are summarized.

2.1 Variable Selection in Linear Model

Variable selection is an important topic in linear regression analysis. In particular, a large number of predictors usually are introduced at the initial stage of modeling to reduce possible modeling biases. However, to get a parsimonious model with strong predictability, statisticians make great efforts on selecting significant variables. We first study the variable selection methods in linear model and then extend them to survival models. Considering the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1.1}$$

where $\mathbf{y}$ is an $n \times 1$ response vector, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)^T$ is an $n \times d$ predictor matrix, $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_d)^T$ denotes the coefficient vector, and $\boldsymbol{\varepsilon}$ is the $n \times 1$ independently identical distributed noise vector with mean zero.

2.1.1 Classical Variable Selection

Classical variable selection is to select an appropriate subset of variables that gives one good fitted or predicted values. For evaluating the performance of regression model, statisticians have developed a variety of measures of fit criteria, including the C_p statistic (Mallows,1973), the Akaike's Information Criterion (AIC, Akaike, 1973, 1974), the Bayesian Information Criterion (BIC, Schwarz, 1978), General information criterion (GIC, Nishii, 1984), and the generalized cross validation score (GCV, Craven and Wahba, 1979).

By fitting all possible models, we can easily find a model that seems best by whatever criterion we choose, which is so called "Best Subset Selection". However, with a d potential predictors, there are 2^d possible subsets, the computational cost is expensive and the exhaustive search is infeasible when d is large. In practice, forward selection, backward elimination and stepwise selection are used to search a good subset rather than best subset selection. Details are referred to Miller (2002).

2.1.2 Variable Selection via Penalized Least Squares

Subset selection provides interpretable models and gives us the most significant predictors, but it still has inherent drawbacks. It could be extremely variable because it is a discrete process – regressors are either retained or removed from the model. Small changes in the data could result in very different selected models. To reduce large variation and improve prediction accuracy, penalized least squares

(PLS) methods are proposed. Instead of minimizing the least squares function, we obtain the estimate by minimizing a penalized PLS function

$$Q(\boldsymbol{\beta}) = \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + n\sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{2.1.2}$$

where $p_\lambda(\cdot)$ is the penalty function with a tuning parameter λ, which controls the model complexity and can be selected by a data-driven method. For simplicity of presentation, we assume that the penalty functions for all coefficients are the same. In this subsection, we present several commonly used penalty functions and their specific properties.

The PLS with L_2 penalty

$$p_\lambda(|\theta|) = \frac{1}{2}\lambda|\theta|^2$$

yields the ridge regression (Hoerl and Kennard, 1970). Ridge regression estimate has an explicit form $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{y})$. It is a continuous process that shrinks coefficients and therefore is more stable. However, it cannot set any coefficient exactly to 0 and hence does not result in a sparse model.

The PLS with L_1 penalty

$$p_\lambda(|\theta|) = \lambda|\theta|$$

yields the LASSO estimate (Tibshirani 1996, 1997). The LASSO estimator equals to a soft-thresholding rule

$$\widehat{\theta}_j = \text{sgn}(z)(|z| - \lambda)_+$$

when the design matrix is orthogonal. The LASSO estimate shrinks the OLS

estimate and produces some coefficients that are exactly 0. Hence it enjoys some of the favorable features of both ridge regression and best subset selection. It produces interpretable model like subset selection and enjoys the stability of ridge regression. However, it would result in large bias for coefficients with large values and therefore is not selection consistent (Zou, 2006). To overcome the inconsistency of LASSO penalty, Zou (2006) proposed a new *Adaptive LASSO penalty* as

$$p_\lambda(|\theta|) = \lambda\widehat{\omega}|\theta|,$$

where $\widehat{\omega} = 1/|\widehat{\beta^0}|^\gamma$ with $\widehat{\beta^0}$ obtained from OLS estimates. The adaptive LASSO assigns different weights for different coefficients and it has been proved that it enjoys the oracle properties with appropriate λ. The adaptive LASSO is essentially an L_1 method and the estimates could be computed by the efficient LARS (Efron, et al., 2004) algorithms.

The PLS with L_q *penalty*

$$p_\lambda(|\theta|) = \lambda|\theta|^q \qquad (0 < q \leq 2)$$

leads to a bridge regression (Frank and Friedman, 1993). By definition, we could see that L_0, L_1 and L_2 penalty are the special cases of L_q penalty. The solution is continuous with respect to the OLS estimate only when $q \geq 1$. However, when $q > 1$, the L_q penalty can not produces a sparse solution.

L_0 (or Entropy penalty) $p_\lambda(|\theta|) = \frac{1}{2}\lambda^2 I(|\theta| \neq 0)$ results in best subset selection with specific value of λ. Compared with L_0 penalty, the hard-thresholding penalty

$$p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda)$$

also results in a hard-thresholding rule

$$\widehat{\theta}_j = zI(|z| > \lambda)$$

which also coincides with the best subset selection for orthonormal designs. However, the latter is more smoother and facilitates computational expedience in general settings.

We have been discussed several penalty functions so far, but what kind of penalty function should we apply to gain desirable properties? Fan and Li (2001) advocates that a good penalty function should result in an estimator with three properties.

- *Unbiasedness*. The resulting estimator is nearly unbiased when the true unknown parameter is large to avoid unnecessary modeling bias.
- *Sparsity*. The resulting estimator is a thresholding rule, which automatically sets small estimated coefficients to zero to reduce model complexity.
- *continuity*. The resulting estimator is continuous in data to avoid instability in model prediction.

Furthermore, Antoniadis and Fan (2001) shows three conditions to guarantee the above three properties.

- *Unbiasedness*. iff $p_\lambda^{'}(|\theta|) = 0$ for large $|\theta|$.
- *Sparsity*: if $\min_\theta\{p_\lambda^{'}(|\theta|) + |\theta|\} > 0$.
- *continuity*: if $\text{argmin}_\theta\{p_\lambda^{'}(|\theta|) + |\theta|\} = 0$.

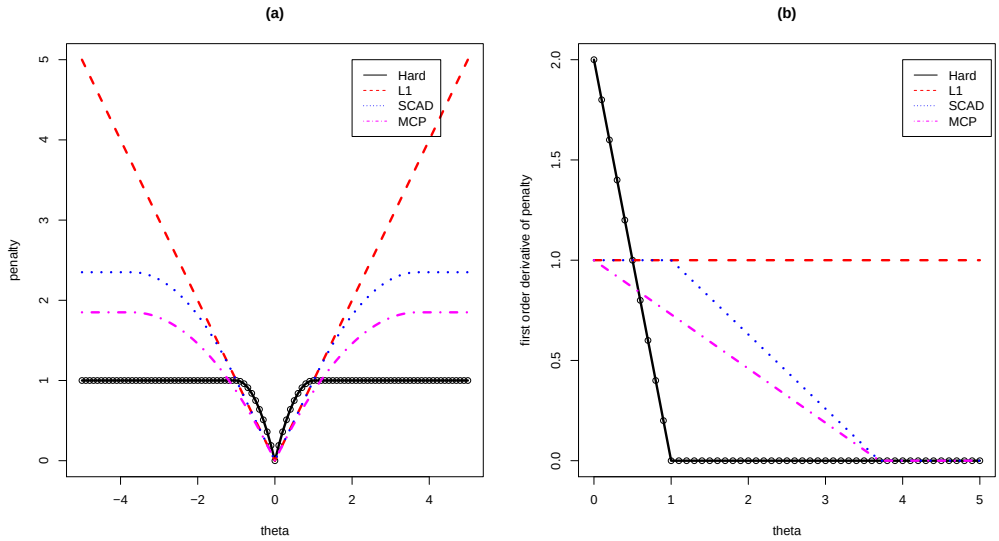

Figure 2.1. Example penalty functions and their corresponding first order deriveatives. (a) Plots the penalty functions of the Hard thresholding penalty, the Soft thresholding penalty, the SCAD penalty and the MCP. (b) illustrates the corresponding derivatives of each penalty function in (a).

It is obviously that all the penalty functions aforementioned can not simultaneously satisfy the mathematical conditions for unbiasedness, sparsity, and continuity.

Fan and Li (2001) proposed a continuous differentiable penalty function defined by

$$\begin{aligned} p_\lambda(\beta_j) &= \lambda \int_0^{|\beta_j|} \left\{ I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \right\} d\theta \qquad a > 2 \\ &= \lambda |\beta_j| I(|\beta_j| \le \lambda) - \frac{(|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2)}{2(a-1)} I(\lambda < |\beta_j| \le a\lambda) \\ &+ \frac{(a+1)\lambda^2}{2} I(|\beta_j| > a\lambda) \end{aligned}$$

This penalty function is called the smoothly clipped absolute deviation (SCAD) penalty, which retains the good mathematical properties of hard and soft thresholding penalty functions and hence is expected to perform the best. It corresponds to a quadratic (nonconcave) symmetric spline function singular at origin with knots

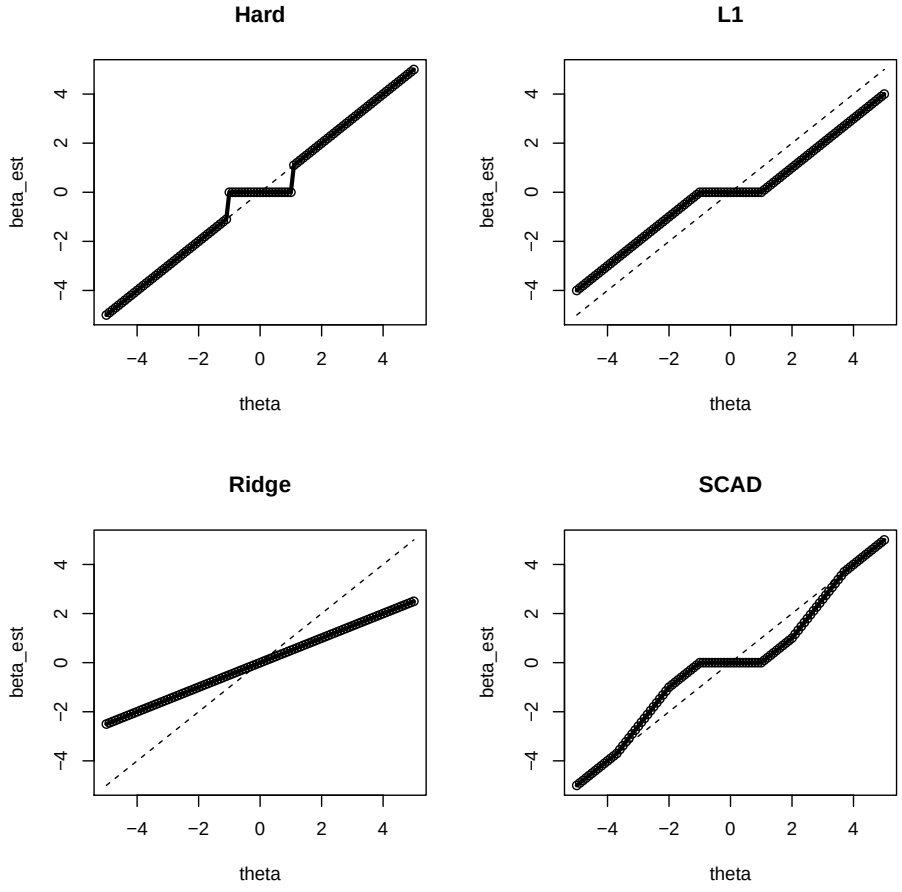

Figure 2.2. The relationship between PLS and OLS estimates when the design matrix is orthonormal.

at λ and $a\lambda$ and satisfies the condition of unbiasedness, sparsity and continuity. It involves two unknown parameters λ and a, where $a = 3.7$ is commonly used in practice.

Thus, the SCAD penalty results in sparse set of solution and approximately unbiased estimates for large coefficients. The SCAD thresholding rule can be given as

$$\widehat{\theta}_j = \begin{cases} \text{sgn}(z)(|z| - \lambda)_+ & \text{if } |z| \leq 2\lambda \\ \{(a-1)z - \text{sgn}(z)a\lambda\}/(a-2) & \text{if } 2\lambda < |z| \leq a\lambda \\ z & \text{if } |z| > a\lambda \end{cases}$$

Zhang (2010) proposed the minimax concave penalty (MCP) defined as

$$p_\lambda(\beta_j) = \lambda \int_0^{|\beta_j|} \left(1 - \frac{\theta}{a\lambda}\right)_+ d\theta \qquad a > 0$$
$$= (\lambda|\beta_j| - \frac{\beta_j^2}{2a})I(|\beta_j| \leq a\lambda) + \frac{a\lambda^2}{2} I(|\beta_j| > a\lambda)$$

The MCP shares the similar spirits with the SCAD penalty, including the 3 desirable properties and oracle property.

Figure 2.1 demonstrates some penalty functions (Hard thresholding, L_1, SCAD, and MCP) and their corresponding first order derivatives with $\lambda = 1$ and $a = 3.7$. Figure 2.2 illustrates the solutions for the Hard, the L_1, the ridge and the SCAD penalty functions with the same λ and a, which clear shows that only the SCAD penalty possess the three desirable properties, namely, unbiasedness, sparsity, and continuity.

2.2 Basic Concepts and Commonly Used Models in Survival Data Analysis

In this section, we first briefly introduce the basic definitions and concepts together with commonly used models in survival analysis. Variable selection and feature screening procedures are then discussed.

2.2.1 Definition and Notation

Survival analysis is introduced to analyze data in which the time until event is of interest. The response in survival data is often referred as a failure time, sur-

vival time, or event time, which are usually treated as continuous. However, the survival time may be incompletely determined for some subjects. For example, we sometimes are interested in how a risk factor or treatment affects time to disease or some other events. If we have study dropout, then for some subjects we know that the survival time is at least equal to some time t. Whereas, for other subjects, we would know their exact time of event. We consider these incompletely observed responses censored. Standard regression procedures could be applied without censoring, however, they may be inadequate because

(1) Time to event is restricted to be positive and has a skewed distribution.

(2) The probability of surviving past a certain point in time may be of more interest than their expected time of event.

(3) The hazard function, used for regression in survival analysis, can lend to more insight into the failure mechanism.

For the analytical methods discussed later to be valid, we assume that the censoring mechanism is noninformative throughout our discussion, namely, censoring is caused by something other than the impending event. Censoring might occur due to generally three reasons: (a) a subject does not experience the event before the study ends; (b) a subject is lost to follow-up during the study period; (c) a person withdraws from the study. All these examples are right-censoring, which commonly happens in real life and is also what of interest in the following discussion.

To record and represent the right-censored survival data, we introduce terminology as follows T_i denotes the survival or failure time for the ith subject; C_i denotes the censoring time for the ith subject; δ_i is the event indicator and defined

by $\delta_i = I(T_i \leq C_i)$. Hence $\delta_i = 1$ when events happen, otherwise, it is censored; Z_i is the observed response defined by Z_i=min(T_i, C_i).

Regarding the event time T a nonnegative continuous random variable, there are several equivalent ways to describe the probability distribution of T. Some of these are familiar, others are special to survival analysis. We will focus on the following quantities:

The *density function* $f(t)$ is defined as

$$f(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T \leq t + \Delta t)$$

The *cumulative distribution function* $F(t)$ denotes the probability of survival time $T \leq t$ and is defined by

$$F(t) = P(T \leq t).$$

However, in survival analysis, our interest tends to focus on the probability of surviving at a certain time t, and therefore we define the *survival function* $S(t)$ by

$$S(t) = P(T > t) = 1 - F(t),$$

which could be written as $S(t) = \bar{F}(t)$. As t ranges from 0 to ∞, $S(t)$ is non-increasing. At time $t = 0$, $S(t) = 1$ and at time $t = \infty$, $S(t) = 0$.

The *hazard function* $h(t)$ is the instantaneous rate at which events occur conditioning on zero previous events.

$$h(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t < T < t + \Delta t | T > t)$$

$$= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{P(t < T < t + \Delta t)}{P(T \geq t)} = \frac{f(t)}{S(t)}$$

The *cumulative hazard function* $H(t)$ describes the accumulated risk up to time t, and is defined by

$$H(t) = \int_0^t h(u)du.$$

The relationships between these descriptive functions $f(t), h(t), S(t)$, and $H(t)$ can be expressed as

$$f(t) = -\frac{d}{dt} S(t)$$

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\{\log S(t)\}$$

$$H(t) = \int_0^t h(u)du = \int_0^t -\frac{d}{du}\{\log S(u)\}du = -\log S(t)$$

$$S(t) = \exp\{-H(t)\}$$

Given a survival data, if we assume that every subject has the same survival curve, namely, there are no covariates or group differences, the survival or hazard function can then be estimated in two ways:

(1) by developing an empirical estimate, such as Kaplan-Meier estimator of the survival function, and Nelson-Aalen estimator for cumulative hazard. This approach requires few assumptions and gives robust estimates.

(2) by specifying a parametric model for hazard function $h(t)$ based on a particular density, such as exponential distribution, weibull distribution and gamma distribution. And use the maximum likelihood estimators(MLE) to estimate the unknown parameters of the parametric distributions. This approach may be too sure to draw inappropriate conclusions.

However, in real life, it is too rigid to assume that the event time of all subjects are governed by the same survival function, namely, the whole population is homogeneous. Therefore, another distinguishing characteristic of survival model, presented by a vector of covariates that may affect survival time, was introduced. The effects of the influential covariates are of such great interest that people developed several models based on different testable assumptions to study them.

2.2.2 Common Used Models for Survival Data

In this subsection, we concern survival model with the following three characteristics: (1) the response is the waiting time until the occurrence of an event; (2) observations are right-censored; and (3) there are covariates whose effects on the survival time are of interest.

Let $\mathbf{x}_i$ represent the set of covariates for ith subject. Note that $\mathbf{x}_i$ may be continuous, discrete and time-varying. The goal of survival analysis is to model the effects of significant covariates given a survival data $\{(\mathbf{x}_i, Z_i, \delta_i)\}_n$.

The most popular framework in analysis of right-censored survival data is the ***Cox's proportional hazards model*** (Cox, 1972), in which the hazard function $h(t|\mathbf{x})$ for a subject with covarties $\mathbf{x}$ is defined as

$$h(t|\mathbf{x}) = h_0(t)\exp(\mathbf{x}^T\boldsymbol{\beta}), \tag{2.2.3}$$

where $h_0(t)$ is the unspecific baseline hazard function who serves as a reference group. Cox's proportional hazards model contains non-parametric $h_0(t)$ and another parametric term, and hence is a semi-parametric model. Under the proportional hazard(PH) assumption, the relative hazard rate(risk) between two groups

only depends on their corresponding covariates, but not time t. In other words, for two subjects with fixed covariates, their relative risk is the same at all durations t. Returning to the relationships between the descriptive functions, we can integrate both sides of (2.2.3) from 0 to t to obtain the cumulative hazard function

$$H(t|\mathbf{x}) = H_0(t)\exp(\mathbf{x}^T\boldsymbol{\beta}),$$

which are also proportional. Applying the relationships again, the survival function $S(t|\mathbf{x})$ can be then determined uniquely by

$$S(t|\mathbf{x}) = S_0(t)^{\exp(\mathbf{x}^T\boldsymbol{\beta})}, \tag{2.2.4}$$

where $S_0(t) = \exp(-\int_0^t h_0(u)du)$ is the baseline survival function. Thus, the effects of covariates $\mathbf{x}$ on the survival function is to raise it to a power given by relative risk $\exp(\mathbf{x}^T\boldsymbol{\beta})$.

It is assumed for the Cox's proportional hazards model that the survival time of subjects are independent. However, this assumptions might be violated when the collected data are correlated. To deal with the dependence among the observations, ***Cox's frailty model*** (Vaupel et al., 1979) was introduced, in which the hazard rate for the jth subject in ith group is

$$h_{ij}(t|\mathbf{x}_{ij}) = h_0(t)u_i\exp(\mathbf{x}_{ij}^T\boldsymbol{\beta}), \tag{2.2.5}$$

where u_i is associated with frailties and follows a specific distribution, such as gamma frailty. It is frequently assumed that given the frailty u_i, the data in the ith group are independent. Frailty model is an extension of proportional hazards model

by considering the survival sample heterogeneous, namely, a mixture of individuals with different baseline hazards caused by unknown or unmeasured covariates.

The Cox PH assumption postulates that the covariates have a fixed multiplicative effect on the hazard, or the relative risk is the same at all durations t. In practice, it is not uncommon for the hazard functions based on two or more groups converge with time. Therefore, it is more reasonable to suppose that the effect of the covariates on the hazard disappears with time. One approach to model such behavior is to include time-varying covariates in the Cox's PH model. As an alternative, Bennett (1983) introduced the ***proportional odds model***, which was defined by

$$\frac{S(t|\mathbf{x})}{1-S(t|\mathbf{x})} = \frac{S_0(t|\mathbf{x})}{1-S_0(t|\mathbf{x})}\exp(\mathbf{x}^T\boldsymbol{\beta}), \tag{2.2.6}$$

where $S_0(t|\mathbf{x})$ is the baseline function of a unspecific form.

All the three models discussed above belong to the class of semi-parametric model with baseline functions of unknown form. In practice, researchers would like to consider some parametric models to gain efficiency in analysis or to obtain some estimates that could be used in future survival study. ***Accelerated failure time (AFT) model*** (Collett, 2003) is one of the parametric models who describes the logarithm of survival time using a conventional linear model. The AFT model can be expressed as

$$\log(T) = \mathbf{x}^T\boldsymbol{\beta} + \sigma\epsilon, \tag{2.2.7}$$

where ϵ is a random error term with a distribution to be specified, and σ is a scaler. This model specifies the distribution of log-survival as a simple shift of a baseline distribution represented by the error term. Furthermore, the conditional survival

function of T given $\mathbf{x}$ can be represented as

$$S(t|\mathbf{x}) = S_0(t\exp(-\mathbf{x}^T\boldsymbol{\beta})), \tag{2.2.8}$$

where $S_0(\cdot)$ is the baseline survival function determined by ϵ. In other words, the survival probability of a subject interested at time t would be exactly the same as the probability of the baseline subject at its time $t\exp(-\mathbf{x}^T\boldsymbol{\beta})$. Different choice of ϵ can result in different parametric survival model. For instance, if ϵ follows standard extreme value distribution, namely, e^ϵ follows a unit exponential distribution, the survival time T follows an exponential distribution with

$$h(t|\mathbf{x}) = h_0(t)\exp(-\mathbf{x}^T\boldsymbol{\beta}), \tag{2.2.9}$$

where $h_0(t) = \frac{1}{\sigma}t^{\frac{1}{\sigma}-1}$. Although model (2.2.9) has the same form as Cox's PH model, the hazard function here is parametric.

Estimation and inference of the parameters in the semi-parametric or parametric models aforementioned could be achieved by likelihood methods, which would be discussed later.

2.3 Variable Selection in Cox's Model

In this subsection, we introduce the techniques of variable selection via penalization to survival analysis setting with right-censored data. Given the i.i.d. observed data $\{(\mathbf{x}_i, Z_i, \delta_i)\}_n$ and the assumption that T and C are independent conditioning on

$\mathbf{x}$, a full likelihood of the data can be written as

$$L = \prod_u f(Z_i|\mathbf{x}_i) \prod_c \bar{F}(Z_i|\mathbf{x}_i) = \prod_u h(Z_i|\mathbf{x}_i) \prod_{i=1}^{n} \bar{F}(Z_i|\mathbf{x}_i), \tag{2.3.10}$$

where the subscript c and u denote the censored and uncensored data respectively, and $f(t|\mathbf{x}), \bar{F}(Z_i|\mathbf{x}_i)$ and $h(Z_i|\mathbf{x}_i)$ are the conditional density function , the conditional survival function and the conditional hazard function of T given $\mathbf{x}$. Furthermore, let $t_1 < \ldots < t_N$ denote the ordered observed failure times and j denote the label for item falling at t_j so that the covariates associated with the N failures are $\mathbf{x}_1, \ldots, \mathbf{x}_N$. Let R_j denote the risk set right before time t_j, namely, R_j=$\{i : Z_i \geq t_j\}$.

Based on the Cox's proportional hazards (PH) assumption,

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

where $h_0(t)$ and $\boldsymbol{\beta}$ are the baseline hazard function and the corresponding parameters. The likelihood in (2.3.10) becomes

$$L(h_0(t), \boldsymbol{\beta}) = \prod_u h_0(Z_i)\exp(\mathbf{x}_i^T\boldsymbol{\beta}) \prod_{i=1}^{n} \exp\{-H_0(Z_i)\exp(\mathbf{x}_i^T\boldsymbol{\beta})\}, \tag{2.3.11}$$

where $H_0(\cdot)$ is the cumulative baseline hazard function. What of interest is the estimate of $\boldsymbol{\beta}$ and therefore we treat $h_0(t)$ as a nuisance parameter. Following Breslow's idea, maximizing the profiled likelihood $L(h_0(t), \widehat{\boldsymbol{\beta}})$ with respect to $h_0(t)$ conditioning on the estimated $\boldsymbol{\beta}$ yields an estimate of $h_0(t)$. Substituting this profiled estimate $\widehat{h}_0(t)$ into (2.3.11), we get the resulting function that depends

only on $\boldsymbol{\beta}$ after dropping the constant terms

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{N} \frac{\exp(\mathbf{x}_j^T \boldsymbol{\beta})}{\sum_{i \in R_j} \exp(\mathbf{x}_i^T \boldsymbol{\beta})}. \tag{2.3.12}$$

The logarithm of (2.3.12) can be written as

$$\ell(\boldsymbol{\beta}) = \sum_{j}^{N} \left[\mathbf{x}_j^T \boldsymbol{\beta} - \log\Big\{ \sum_{i \in R_j} \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \Big\} \right], \tag{2.3.13}$$

which is the partial likelihood function (Cox, 1975). Similar to PLS, penalized maximum partial likelihood estimator based on (2.3.13) can be used to select significant covariates. The penalized partial likelihood can be defined by

$$Q(\boldsymbol{\beta}) = \sum_{j}^{N} \left[\mathbf{x}_j^T \boldsymbol{\beta} - \log\Big\{ \sum_{i \in R_j} \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \Big\} \right] - n \sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{2.3.14}$$

With proper choice of $p_\lambda(\cdot)$, many of the estimated coefficients will be zero and hence their corresponding covariates do not appear in the model. This achieves the objectives of variable selection.

The SCAD penalty functions was applied by Fan and Li (2002) to solve the variable selection problem for Cox's proportional hazards model while inheriting good properties, such as oracle property.

Fan and Li (2002) listed several mild conditions to guarantee the asymptotic normality of the maximum partial likelihood estimates, see Andersen and Gill (1982) and Murphy and van der Vaart (2000) for details. They then derived the theorems to show that SCAD thresholding penalized likelihood estimators, converging at rate of root-n, perform as well as the oracle procedure.

Denote by $\boldsymbol{\beta}_0$ the true value of $\boldsymbol{\beta}$, and let $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{20})^T$ where $\boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{20}$ are the corresponding zero and nonzero components of $\boldsymbol{\beta}$. Similarly, denote the SCAD estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2)^T$, where $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2$ are the nonvanishing and vanishing components of $\widehat{\boldsymbol{\beta}}$. Fan and Li (2002) showed that under the conditions that

$$\max\{|p''_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0\} \to 0,$$

$$\lambda \to 0,$$

$$\sqrt{n}\lambda \to \infty \text{ as } n \to \infty.$$

we can have, with probability tending to 1,

$$\widehat{\boldsymbol{\beta}}_2 = \mathbf{0},$$

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \to N\{\mathbf{0}, \mathrm{I}_1^{-1}(\boldsymbol{\beta}_{10}, \mathbf{0})\},$$

where $\mathrm{I}_1^{-1}(\boldsymbol{\beta}_{10}, \mathbf{0})$ is the Fisher information matrix of $\boldsymbol{\beta}_1$ when $\boldsymbol{\beta}_2 = \mathbf{0}$. Namely, they work as well as if the correct sub-model was known.

They adopted the local quadratic approximation (LQA, Fan and Li, 2001) algorithm to reduce the nonconvex optimization problem to a local quadratic one. maximizing $Q(\boldsymbol{\beta})$ is equivalent to minimizing

$$-\ell(\boldsymbol{\beta}) + n\sum_{i=1}^{d} p_\lambda(|\beta_j|).$$

They showed that penalty function can be locally approximated by a quadratic function respectively. Therefore, maximizing $Q(\boldsymbol{\beta})$ can be locally approximated

(except for the constant term) by minimizing

$$-\nabla\ell(\boldsymbol{\beta^0})^T(\boldsymbol{\beta^0}-\boldsymbol{\beta})+\frac{1}{2}(\boldsymbol{\beta^0}-\boldsymbol{\beta})^T\nabla^2\ell(\boldsymbol{\beta^0})(\boldsymbol{\beta^0}-\boldsymbol{\beta})+\frac{1}{2}n\boldsymbol{\beta}^T\Sigma_\lambda(\boldsymbol{\beta^0})\boldsymbol{\beta},$$

where $\boldsymbol{\beta}^0$ is the initial $\boldsymbol{\beta}$ estimate, $\Sigma_\lambda(\boldsymbol{\beta^0})$ =diag$\{p_\lambda^{'}(|\beta_1^0|)/|\beta_1^0|,\ldots,p_\lambda^{'}(|\beta_d^0|)/|\beta_d^0|\}$, β_j^0 are the components of initial value $\boldsymbol{\beta}^0$.

Hence Newton-Raphson algorithm could be applied here to estimate $\boldsymbol{\beta}$ by one-step procedure instead of fully iterative MLE as long as the initial estimator is good enough. Fivefold cross-validation and generalized cross validation are used to select the proper value of λ.

Zhang and Lu (2007) also considered Cox's model with noninformative censoring mechanism. To avoid the inconsistency of the LASSO and the numerical complexity of the SCAD, they applied the adaptive LASSO penalty, whose weights are determined by unpenalized maximum likelihood estimator $\widehat{\boldsymbol{\beta}}^{MLE}$. Namely,

$$Q(\boldsymbol{\beta})=-\ell(\boldsymbol{\beta})+n\lambda\sum_{i=1}^{d}|\beta_j|/w_j,$$

where $w_j=1/\widehat{\beta}_j^{MLE}$, since $\widehat{\boldsymbol{\beta}}_j^{MLE}$ is consistent, it can reflect the importance of the covariates. This method incorporates different important penalties for different coefficients: unimportant variables receive larger penalties than important ones, so that important covariates are more likely to be retained, whereas the unimportant ones are more likely to be dropped.

Zhang and Lu (2007) also showed that the adaptive LASSO estimator possess oracle estimator with proper choice of λ. They proved that as long as

$$\sqrt{n}\lambda\to 0,\ n\lambda\to\infty,$$

the adaptive LASSO estimator enjoys the oracle property. which is similar to the one described in Fan and Li (2002), with probability tending to 1, there exits a root-n consistent adaptive LASSO estimator $\widehat{\boldsymbol{\beta}}$ that works as well as if the correct sub-model was known. The modified Fu's shooting algorithm was used to solve the optimization problem. Similarly, they choose GCV statistic as their tuning parameter selector.

2.4 Feature screening in Ultra-High Dimensional Survival Data Analysis

The penalized variable selection methods work well with a moderate number of coveriates for Cox's PH model, however its usefulness is limited in the case of ultrahigh dimensional screening problems. Extent the idea of sure independence screening (SIS, Fan, Samworth, and Wu, 2009) and iterative independence screening (ISIS, Fan, Samworth, and Wu, 2009) proceduces in GLM, Fan, Feng and Wu (2010) employed maximum of the partial likelihood of the single covariate as a marginal utility measure. Following the same notation in section 2.4.2, define

$$\mu_k(\beta_k) = \max_{\beta_k} \left(\sum_{i=1}^{n} \delta_i x_{ik} \beta_k - \sum_{i=1}^{n} \delta_i \log\Big\{ \sum_{j \in R(z_i)} \exp(\mathbf{x}_{jk}^T \beta_k) \Big\} \right), \tag{2.4.15}$$

where $R(t) = \{i : Z_i \geq t\}$ and x_{ik} is the kth component of $\mathbf{x}_i$. So the marginal utility reflects how much information the corresponding covariate contains for the survival responses. Once all the marginal utilities are obtained, they can be ordered from the largest to smallest. Fan, Feng and Wu (2010) choose the top $d = [n/\log(n)]$ covariates as the screening features.In the rest of this dissertation,

we will refer to this procedure as Cox-SIS.

After Cox-SIS, the parameter dimensionality is reduced from the original p to $d < n$, where the refined variable selection technique can be applied. Cox-SIS can handle challenging cases, such as some covariates that are marginally independent but jointly dependent with the response variable by iterated Cox-SIS, which is actually the conditional feature ranking and iterative feature screening, similarly to ISIS in GLM. The only difference is that the likelihood function is now replaced by partial likelihood. Moreover, two variant of iterated Cox-SIS can be used to reduce false selected rates (FSR). However, two major problems remain unaddressed with Cox-SIS. First, the extension of sure screening property to Cox's model is difficult, because censoring is confounding between the covariates and the survival outcome. Second, Cox-SIS is a model-based method which only works well when the true underlying model is indeed a Cox's model. Consequently its power is very limited when Cox's model is not applicable.

The screening procedures require choosing a threshold to dictate how many variables to retain, but there are no principled methods for making such a choice, making the resulting screened models difficult to evaluate. Zhao and Li (2012) followed the spirit of Fan, Feng, and Wu (2010), but provided a new, principled method for choosing the number of covariates to retain based on specifying the desired false positive rate. They solve $\widehat{\beta}_k$ marginally by

$$\widehat{\beta}_k = \text{argmax}_{\beta_k} \left(\sum_{i=1}^{n} \delta_i x_{ik} \beta_k - \sum_{i=1}^{n} \delta_i \log\Big\{ \sum_{j \in R(z_i)} \exp(\mathbf{x}_{jk}^T \beta_k) \Big\} \right). \tag{2.4.16}$$

Let $I_k(\beta_k)$ denote the information matrix at $\widehat{\beta}_k$. They illustrated that by screening

the model with

$$\widehat{\mathcal{M}}_\gamma = \{1 \leq k \leq p,\ I_k^{\frac{1}{2}}(\beta_k)|\beta_k| > \gamma\},$$

one can control the expected false positive rate at $2(1 - \Phi(\gamma))$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. One sensible way to do this would be to first fix the number of false positives f that we are willing to tolerate. They are conservative by letting $\gamma = \Phi^{-1}(1-\frac{f}{2p})$, so that the expected false positive rate is $2(1 - \Phi(\gamma))$= f/p, which is smaller than the desirable false positive rate. Term this method a principled Cox sure independence screening procedure (PSIS), as the cutoff γ is selected to control the false positive rate. Specifically, PSIS is implemented as follows:

(1) Fit a marginal Cox's model for each of the covariates according to (2.4.38) to get parameter estimates $\widehat{\beta}_k$ and their corresponding variance estimates $I_k^{-1}(\widehat{\beta}_k)$.

(2) Fix the false positive rate as f/p and let $\gamma = \Phi^{-1}(1 - \frac{f}{2p})$.

(3) Retain covariates by $\widehat{\mathcal{M}}_\gamma = \{1 \leq k \leq p,\ I_k^{\frac{1}{2}}(\beta_k)|\beta_k| > \gamma\}$.

They also gave the first theoretical justifications of the sure independence screening procedure for censored data. Zhao and Li (2012) first shown that $\widehat{\beta}_k$ are consistent for β_{k0}, the real value of β_k, and they also suggested that in order to detect covariate $j \in \widehat{\mathcal{M}}$, $|\beta_{k0}|$ has to be at least $O(n^{-\frac{1}{2}})$. Under the asymptotic framework where the number of covariates can grow with the sample size, Zhao and Li (2012) proved the sure screening propoerty (SSP) for PSIS.

They showed that by choosing $\gamma = \Phi^{-1}(1-\frac{f}{2p})$, under the condition that $\kappa < \frac{1}{2}$

and $\log(p) = O(n^{\frac{1}{2}-\kappa})$, with probability going to 1,

$$P(\mathcal{M} \subseteq \widehat{\mathcal{M}}) \geq 1 - s \exp(-cn^{1-2\kappa}),$$

where $c > 0$, s is the number of nonzero components in $\boldsymbol{\beta}_0$. Hence, their screening procedure PSIS will select all of the important variables with a false positive rate close to f/p. Therefore, it is reasonable to work on $\widehat{\mathcal{M}}$ to further select significant variables and identify the underlying model structure by penalized variable selection tools.

Chapter 3

Feature Screening in Ultrahigh Dimensional Cox's Model

Survival data with ultrahigh dimensional covariates such as genetic markers have been collected in medical studies and other fields. Feature screening plays 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. However, both screening procedures have a great chance to neglect important predictors that relates to responses jointly but not independently without iteratively repeating the procedures.

In this chapter, 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 also develop a computationally effective algorithm to carry out the proposed procedure and establish the ascent property of the proposed algorithm. The proposed procedure is further shown to possess the sure screening property.

3.1 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:

$$h(t|\mathbf{x}) = h_0(t)\exp(\mathbf{x}^T\boldsymbol{\beta}), \tag{3.1.1}$$

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 (3.1.1). 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

$$\ell_p(\boldsymbol{\beta}) = \sum_{j=1}^{N} [\mathbf{x}_{(j)}^T \boldsymbol{\beta} - \log\{\sum_{i \in R_j} \exp(\mathbf{x}_i^T \boldsymbol{\beta})\}]. \tag{3.1.2}$$

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

$$\widehat{\boldsymbol{\beta}}_m = \arg\max_{\boldsymbol{\beta}} \ell_p(\boldsymbol{\beta}) \ \text{ subject to } \ \|\boldsymbol{\beta}\|_0 \leq m \tag{3.1.3}$$

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 (3.1.3) 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 $\ell_p''(\boldsymbol{\gamma})$

$$g(\boldsymbol{\gamma}|\boldsymbol{\beta}) = \ell_p(\boldsymbol{\beta}) + (\boldsymbol{\gamma}-\boldsymbol{\beta})^T\ell_p'(\boldsymbol{\beta}) - \frac{u}{2}(\boldsymbol{\gamma}-\boldsymbol{\beta})^T W(\boldsymbol{\gamma}-\boldsymbol{\beta}), \tag{3.1.4}$$

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 (3.1.4) 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, $g(\boldsymbol{\gamma}|\boldsymbol{\beta}) \leq \ell_p(\boldsymbol{\beta})$ 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

$$\max_{\boldsymbol{\gamma}} g(\boldsymbol{\gamma}|\boldsymbol{\beta}) \qquad \text{subject to} \quad \|\boldsymbol{\gamma}\|_0 \leq m \tag{3.1.5}$$

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 the order statistics of $\tilde{\gamma}_j$ by $|\tilde{\gamma}_{(1)}| \geq |\tilde{\gamma}_{(2)}| \geq \cdots \geq |\tilde{\gamma}_{(p)}|$. The solution of maximization problem (3.1.5) is the hard-thresholding rule defined below

$$\widehat{\gamma}_j = \tilde{\gamma}_j I\{|\tilde{\gamma}_j| > |\tilde{\gamma}_{(m+1)}|\} \widehat{=} H(\tilde{\gamma}_j; m). \tag{3.1.6}$$

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

$$\tilde{\boldsymbol{\beta}}^{(t)} = (H(\tilde{\gamma}_1^{(t)}; m), \cdots, H(\tilde{\gamma}_p^{(t)}; m))^T \widehat{=} \mathbf{H}(\tilde{\boldsymbol{\gamma}}^{(t)}; m). \tag{3.1.7}$$

Set $S_t = \{j : \tilde{\beta}_j^{(t)} \neq 0\}$, 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 updated $\boldsymbol{\beta}$ 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). Theorem 1 below offers convergence behavior of the proposed algorithm.

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

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

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. One also 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, Tishirani, 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.

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 $\boldsymbol{\beta}$.

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

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

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

Thus it follows that $\ell_p(\boldsymbol{\gamma}) \geq g(\boldsymbol{\gamma}|\boldsymbol{\beta})$ 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

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

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 $\|\boldsymbol{\gamma}\|_0 \leq m$. By definition of $\boldsymbol{\beta}^{(t+1)}$, $\ell_p(\boldsymbol{\beta}^{(t+1)}) \geq \ell_p(\boldsymbol{\beta}_*^{(t+1)})$ and $\|\boldsymbol{\beta}^{(t+1)}\|_0 = m$. This proves Theorem 1.

3.2 Monte Carlo Simulations

In this section, we evaluate the finite sample performance of the proposed feature screening procedure via Monte Carlo simulations. All simulations were conducted by using R codes.

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=1000 and 2000. 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 3.2 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) with sample size equal to 100. 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

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. The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ with $\Sigma = (1-\rho)I + \rho\mathbf{1}\mathbf{1}^T$ (n=100)

			Cox-SIS					SJS				
			$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
p	ρ	$\boldsymbol{\beta}$	X_1	X_2	X_3	X_4	ALL	X_1	X_2	X_3	X_4	ALL
1000	0.25	b1	0.995	0.993	0.997	0	0	1	1	1	1	1
		b2	0.892	0.885	0.882	0.882	0.603	1	1	1	1	1
1000	0.5	b1	0.967	0.972	0.966	0	0	0.987	0.989	0.992	1	0.986
		b2	0.813	0.792	0.795	0.814	0.384	0.998	0.997	0.998	0.998	0.992
1000	0.75	b1	0.854	0.868	0.860	0.007	0.006	0.996	0.993	0.991	0.987	0.976
		b2	0.699	0.717	0.684	0.713	0.201	0.967	0.968	0.957	0.966	0.891
2000	0.25	b1	0.993	0.992	0.988	0	0	0.999	1	0.999	1	0.998
		b2	0.837	0.848	0.842	0.827	0.469	0.998	0.999	1.000	0.998	0.997
2000	0.5	b1	0.942	0.95	0.946	0	0	0.973	0.975	0.977	1	0.966
		b2	0.730	0.729	0.718	0.727	0.259	0.989	0.986	0.984	0.984	0.962
2000	0.75	b1	0.801	0.774	0.782	0.006	0.002	0.979	0.980	0.975	0.982	0.952
		b2	0.610	0.613	0.609	0.621	0.117	0.926	0.911	0.906	0.916	0.760

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 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

Table 3.3. The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ with $\Sigma = (1-\rho)I + \rho\mathbf{1}\mathbf{1}^T$ (n=200)

			Cox-SIS					SJS				
			$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
p	ρ	$\boldsymbol{\beta}$	X_1	X_2	X_3	X_4	ALL	X_1	X_2	X_3	X_4	ALL
1000	0.25	b1	1	1	1	0	0	1	1	1	1	1
		b2	0.983	0.982	0.989	0.988	0.942	1	1	1	1	1
1000	0.5	b1	1	1	1	0	0	1	1	1	1	1
		b2	0.961	0.974	0.965	0.964	0.871	1	1	1	1	1
1000	0.75	b1	0.996	0.997	0.993	0	0	1	1	1	1	1
		b2	0.916	0.944	0.919	0.925	0.721	1	1	0.999	1	0.999
2000	0.25	b1	1	1	1	0	0	1	1	1	1	1
		b2	0.972	0.974	0.979	0.973	0.903	1	1	1	1	1
2000	0.5	b1	1	1	0.999	0	0	1	1	1	1	1
		b2	0.940	0.943	0.938	0.942	0.779	1	1	1	1	1
2000	0.75	b1	0.993	0.996	0.988	0	0	1	1	1	0.998	0.998
		b2	0.886	0.891	0.873	0.877	0.591	0.999	1	0.999	0.999	0.997

to be the one in (b1). There is no doubt that SJS outperforms Cox-SIS easily in this setting. Increasing our sample size from 100 to 200 as shown in Table 3.3, we have similar results to Table 3.2 with even better screening performance due to the larger sample size.

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. Tables 3.2 and 3.3 clearly show 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.

Tables 3.2 and 3.3 also indicate 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 Tables 3.2 and 3.3, SJS outperforms Cox-SIS no matter whether there presents marginally independent but jointly dependent predictors or not.

Table 3.4. The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ with $\Sigma = (\rho^{|i-j|})$ (n=100)

			Cox-SIS					SJS				
			$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
p	ρ	$\boldsymbol{\beta}$	X_1	X_2	X_3	X_4	ALL	X_1	X_2	X_3	X_4	ALL
1000	0.25	b1	0.999	1	1	0.32	0.32	1	1	1	1	1
		b2	0.995	1	1	0.994	0.989	1	1	1	1	1
1000	0.5	b1	1.000	1	0.971	0.575	0.546	0.999	1	0.994	1.000	0.994
		b2	0.999	1	1	1	0.999	1	1	1	1	1
1000	0.75	b1	1.000	1.000	0.643	0.423	0.140	0.998	0.992	0.885	0.990	0.879
		b2	1	1	1	1	1	1	1	1	1	1
2000	0.25	b1	1	1	0.998	0.217	0.216	1	1	1	0.999	0.999
		b2	0.985	1	0.999	0.986	0.970	1	1	1	1	1
2000	0.5	b1	1	1	0.939	0.436	0.382	1	0.996	0.981	0.996	0.979
		b2	0.999	1	1	1	0.999	1	1	1	1	1
2000	0.75	b1	1	1	0.579	0.344	0.074	0.984	0.965	0.751	0.952	0.733
		b2	1	1	1	1	1	1	1	1	1	1

Tables 3.4 and 3.5 depict the simulation results for the AR covariance structure (S2) with sample size equal to 100 or 200. It is worth to note that with the AR

Table 3.5. The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ with $\Sigma = (\rho^{|i-j|})$ (n=200)

			Cox-SIS					SJS				
			$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
p	ρ	$\boldsymbol{\beta}$	X_1	X_2	X_3	X_4	ALL	X_1	X_2	X_3	X_4	ALL
1000	0.25	b1	1	1	1	0.699	0.699	1	1	1	1	1
		b2	1	1	1	1	1	1	1	1	1	1
1000	0.5	b1	1	1	1	0.929	0.929	1	1	1	1	1
		b2	1	1	1	1	1	1	1	1	1	1
1000	0.75	b1	1	1	0.953	0.849	0.802	1	1	1	1	1
		b2	1	1	1	1	1	1	1	1	1	1
2000	0.25	b1	1	1	1	0.595	0.595	1	1	1	1	1
		b2	0.999	1	1	1	0.999	1	1	1	1	1
2000	0.5	b1	1	1	1	0.871	0.871	1	1	1	1	1
		b2	1	1	1	1	1	1	1	1	1	1
2000	0.75	b1	1	1	0.92	0.773	0.693	1	1	1	1	1
		b2	1	1	1	1	1	1	1	1	1	1

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, it is expected that the Cox-SIS works well for both cases (b1) and (b2). Table 3.4 indicates that both Cox-SIS and SJS perform very well when $\boldsymbol{\beta}$ is set to be the one in (b2). On the other hand, the Cox-SIS has very low $\mathcal{P}_a$ when $n = 100$ and $\boldsymbol{\beta}$ is set to be the one in (b1), while $\mathcal{P}_a$ becomes much higher when the sample size increases from 100 to 200 as shown in Table 3.5. In summary, SJS outperform Cox-SIS in all cases considered in Table 3.4 and Table 3.5, in particular, when $\boldsymbol{\beta}$ is set to be the one in (b1).

We next compare SJS with the iterative Cox-SIS. Table 3.2 and Table 3.3 indicate 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 Tables 3.6 and 3.7. 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. Tables 3.6 and 3.7 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 $p = 1000$, and is comparable in computing time to Cox-ISIS when $p = 2000$.

Tables 3.6 and 3.7 with $\tau = 1$ together with Tables 3.2 and 3.3 indicate that Cox-ISIS outperforms Cox-SIS in the presence of predictors that are marginally independent of, but jointly dependent of the survival time, although SJS still outperforms Cox-ISIS. An important question is: does Cox-ISIS always perform better than Cox-SIS?

To address this question, we conduct simulations to directly compare the per-

Table 3.6. Comparison with Cox-ISIS (p=1000)

		Cox-ISIS						SJS					
		$\mathcal{P}_s$				$\mathcal{P}_a$	Time	$\mathcal{P}_s$				$\mathcal{P}_a$	Time
τ	ρ	X_1	X_2	X_3	X_4	ALL	(second)	X_1	X_2	X_3	X_4	ALL	(second)
1	0.25	1	1	1	0.999	0.999	13.13	1	1	1	1	1	3.91
	0.5	0.931	0.935	0.945	1	0.824	13.18	0.990	0.986	0.992	1	0.986	4.40
	0.75	0.775	0.782	0.739	1	0.425	11.77	0.998	0.996	0.998	0.991	0.987	4.37
0.75	0.25	1	0.999	0.999	0.999	0.997	7.69	1	1	1	1	1	2.30
	0.5	0.934	0.937	0.933	1	0.812	10.58	0.993	0.996	0.993	1	0.991	3.65
	0.75	0.777	0.782	0.769	1.000	0.439	7.72	0.993	0.990	0.994	0.984	0.976	2.76
0.50	0.25	0.998	0.999	0.999	0.999	0.995	14.65	1	1	1	1	1	4.48
	0.5	0.922	0.931	0.942	1	0.805	13.95	0.986	0.989	0.994	1	0.986	5.19
	0.75	0.761	0.744	0.760	1	0.402	14.44	0.978	0.983	0.975	0.98	0.943	5.28
0.25	0.25	0.988	0.992	0.988	0.994	0.965	11.17	0.996	0.995	0.992	0.985	0.969	2.99
	0.5	0.869	0.884	0.873	1	0.65	7.62	0.933	0.925	0.940	0.999	0.902	2.68
	0.75	0.647	0.637	0.643	1	0.245	7.59	0.776	0.758	0.755	0.947	0.479	2.43

Table 3.7. Comparison with Cox-ISIS (p=2000)

		Cox-ISIS						SJS					
		$\mathcal{P}_s$				$\mathcal{P}_a$	Time	$\mathcal{P}_s$				$\mathcal{P}_a$	Time
τ	ρ	X_1	X_2	X_3	X_4	ALL	(second)	X_1	X_2	X_3	X_4	ALL	(second)
1	0.25	1	0.998	1.000	0.998	0.996	13.64	1	0.998	0.999	0.999	0.997	10.94
	0.5	0.893	0.895	0.912	1	0.714	13.52	0.979	0.980	0.983	1	0.975	12.86
	0.75	0.720	0.675	0.70	1	0.315	13.39	0.971	0.969	0.97	0.991	0.952	13.96
0.75	0.25	0.999	0.999	1	0.997	0.995	13.90	1	0.999	0.999	1	0.998	11.56
	0.5	0.884	0.888	0.897	1	0.689	14.05	0.961	0.950	0.958	1	0.942	13.66
	0.75	0.681	0.720	0.686	1	0.319	13.75	0.973	0.969	0.966	0.989	0.945	14.41
0.50	.25	1	0.997	0.997	0.997	0.992	13.97	1	1	0.999	0.999	0.998	11.31
	0.5	0.903	0.909	0.884	1	0.718	13.84	0.961	0.966	0.963	0.999	0.947	14.41
	0.75	0.689	0.682	0.692	1	0.307	13.49	0.941	0.948	0.954	0.973	0.879	13.71
0.25	0.25	0.973	0.971	0.972	0.986	0.91	13.83	0.979	0.983	0.983	0.963	0.92	9.79
	0.5	0.818	0.831	0.843	1	0.555	13.74	0.857	0.849	0.855	0.996	0.779	13.03
	0.75	0.548	0.516	0.535	1	0.128	13.64	0.624	0.634	0.616	0.942	0.303	12.51

Table 3.8. Comparison among Cox-SIS, Cox-ISIS and SJS ($\Sigma = S1$, n=200)

			Cox-SIS		Cox-ISIS		SJS	
τ	p	ρ	$\mathcal{P}_a$	Time (s)	$\mathcal{P}_a$	Time (s)	$\mathcal{P}_a$	Time (s)
1	1000	0.25	0.940	4.43	1	13.94	1	2.59
		0.50	0.855	3.72	1	12.34	1	2.29
		0.75	0.742	4.47	0.999	15.77	1	2.91
1	2000	0.25	0.915	6.86	1	21.15	1	10.07
		0.50	0.779	6.95	1	22.96	1	10.67
		0.75	0.613	6.96	0.994	23.98	0.999	11.49
0.5	1000	0.25	0.929	4.99	1	14.67	1	2.71
		0.50	0.796	4.99	0.996	14.78	1	2.90
		0.75	0.660	5.90	0.856	17.38	0.928	3.91
0.5	2000	0.25	0.879	7.08	0.999	20.43	1	9.35
		0.50	0.725	7.07	0.989	20.55	0.996	9.97
		0.75	0.546	6.94	0.706	21.13	0.851	11.17

formance of Cox-SIS, Cox-ISIS and SJS, with the setting of $\beta_j = \tau(-1)^U(a + |V_j|)$ for $j = 1, 2, 3$ and 4, where $a = 4\log n/\sqrt{n}$, $U \sim$ Bernoulli(0.4) and $V_j \sim \mathcal{N}(0, 1)$ with $\tau = 1$ and 0.5. To save space, Table 3.8 depicts simulation results for only the cases of $n = 200$ with the covariance matrix set to be S1, and the case of $n = 100$ with the covariance matrix set to be S2 are summarized in Table 3.9. The values of $\mathcal{P}_a$ and computing time are reported in Tables 3.8 and 3.9 from which, it can be seen that Cox-ISIS outperforms Cox-SIS in terms of $\mathcal{P}_a$ in all scenarios included this table. While Cox-SIS needs less computing time than Cox-ISIS. Tables 3.8 and 3.9 also indicate that SJS outperforms both Cox-SIS and Cox-ISIS in terms

Table 3.9. Comparison among Cox-SIS, Cox-ISIS and SJS ($\Sigma = S2$, n=100)

			Cox-SIS		Cox-ISIS		SJS	
τ	p	ρ	$\mathcal{P}_a$	Time (s)	$\mathcal{P}_a$	Time (s)	$\mathcal{P}_a$	Time (s)
1	1000	0.25	0.981	3.48	1	8.49	1	2.06
		0.50	1	3.73	1	9.32	1	2.08
		0.75	1	3.47	1	9.17	0.999	1.81
1	2000	0.25	0.961	5.75	1	13.64	1	9.47
		0.50	1	5.86	1	14.50	1	9.05
		0.75	1	5.85	1	14.94	1	8.80
0.5	1000	0.25	0.974	3.75	1	8.79	1	2.28
		0.50	1	3.16	1	7.37	0.999	1.87
		0.75	1	3.14	1	7.32	0.989	1.82
0.5	2000	0.25	0.945	5.76	1	13.33	0.999	9.65
		0.50	0.998	5.92	1	13.76	0.999	9.75
		0.75	1	5.88	1	13.68	0.982	9.56

of $\mathcal{P}_a$. Furthermore, SJS needs less computing time than Cox-ISIS for all cases, and needs less computing times than Cox-SIS when $p = 1000$, but needs more computing time than Cox-SIS when $p = 2000$.

3.3 An Application: DLBCL Data Study

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.

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 3.10. The maximum of partial likelihood function of three corresponding models obtained by SJS, Cox-ISIS and Cox-SIS procedures are -556.9332, -561.0333, and -600.0885, respectively. This implies that both SJS and Cox-ISIS performs much better than Cox-SIS with SJS performing the best.

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 3.11. The likelihood, the degree of freedom (df), the BIC score and the AIC

Table 3.10. Fourty-three gene IDs selected by Cox-SJS, Cox-ISIS and Cox-SIS

	SJS			Cox-ISIS			Cox-SIS		
Gene	66	3813	5476	427	2108	4548	1072	1841	5027
IDs	773	3819	5668	655	2109	4721	1188	2437	5054
	1112	3820	5953	1188	2244	4723	1439	2579	5055
	1662	3824	6125	1456	2246	5034	1456	2672	5297
	1664	3825	6598	1579	2361	5055	1660	3799	5301
	1680	3826	6607	1662	2579	5301	1662	3810	5614
	1681	4317	6860	1671	3799	5614	1663	3811	5950
	1753	4531	7175	1681	3811	5649	1664	3812	5953
	1825	4573	7301	1682	3813	5950	1671	3813	6365
	3361	4574	7302	1825	3822	6956	1672	3820	6519
	3362	5025	7307	1878	3824	7098	1678	3821	7096
	3497	5172	7343	1996	3825	7343	1680	3822	7343
	3595	5214	7357	2064	4131	7357	1681	3824	7357
	3799	5297		2106	4317		1682	3825	
	3810	5362		2107	4448		1825	4131	

score of the resulting models are listed in Table 3.12, 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 20.6646 with df=21. This leads to the P-value of this partial likelihood ratio test to be 0.480. 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 3.13, which indicates

Table 3.11. IDs of selected genes by SCAD and Lasso

	Gene IDs										
SJS-SCAD	1112	1664	1680	1681	1825	3497	3810	3813	3819	3820	3825
	3826	4317	4531	4574	5668	5953	6498	6607	7175	7307	7357
SJS-Lasso	66	773	1112	1662	1664	1681	1825	3362	3497	3595	3813
	3820	3825	4317	4531	4573	4574	5172	5362	5476	5668	5953
	6125	6498	6607	6860	7175	7302	7307	7343	7357		
ISIS-SCAD	1188	1456	1662	1681	1682	1825	1878	1996	2108	3799	3822
	3824	3825	4317	4448	4548	4723	5043	5055	5301	5649	5950
	6956	7098	7343	7357							
ISIS-Lasso	427	655	1188	1456	1579	1662	1671	1681	1825	1878	1996
	2106	2107	2108	2579	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 3.12. Likelihood, df, AIC and BIC of Resulting Models.

	Likelihood	df	BIC	AIC
SJS-SCAD	-567.2655	22	1254.642	1178.531
SJS-Lasso	-565.1238	31	1299.495	1192.248
ISIS-SCAD	-572.1241	26	1286.197	1196.248
ISIS-Lasso	-565.0359	33	1310.238	1196.072
SIS-SCAD	-622.5386	9	1294.213	1263.077
SIS-Lasso	-610.6605	14	1297.755	1249.321

Table 3.13. Estimates and standard errors (SE) based on SJS-SCAD

Gene ID	Estimate(SE)	P-value
1112	-0.3339(0.1024)	5.54e-04
1664	0.55814(0.1339)	1.53e-05
1680	-0.3597(0.3762)	0.1695
1681	0.5299(0.3893)	0.0867
1825	0.8251(0.1149)	3.49e-13
3497	0.3283(0.1001)	5.21e-04
3810	0.8080(0.3606)	0.0125
3813	-0.9893(0.3639)	3.28e-03
3819	0.4356(0.2469)	0.0389
3820	-0.9841(0.3493)	2.42e-03
3825	-0.4024(0.2850)	7.90e-02
3826	0.6696(0.3519)	0.0285
4317	0.5435(0.1151)	1.16e-06
4531	-0.1489(0.0545)	3.14e-03
4574	0.3054(0.0968)	8.09e-04
5668	-0.6550(0.1321)	3.57e-07
5953	0.4244(0.1185)	1.72e-04
6498	0.0582(0.0237)	7.11e-03
6607	0.5278(0.1067)	3.82e-07
7175	-0.0796(0.0274)	1.82e-03
7307	-0.4131(0.1246)	4.60e-04
7357	-0.4718(0.1022)	1.97e-06

that most selected genes have significant impact on the survival time. We further compare Tables 3.10 and 3.13, and find that Gene 4317 was selected by both SJS and Cox-ISIS, but not by Cox-SIS. From Tables 3.11, 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 3.14 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 3.14. Likelihood, AIC and BIC of Models with and without Gene 4317.

	SJS	SJS-SCAD	SJS-Lasso	ISIS	ISIS-SCAD	ISIS-Lasso
LKHD with Gene4317	-556.9332	-567.2655	-565.1238	-561.0333	-572.1241	-565.0359
LKHD w/o Gene4317	-569.451	-575.6539	-576.577	-567.9259	-576.527	-571.4108
df	1	1	1	1	1	1
BIC w/o Gene4317	1368.205	1265.959	1316.942	1365.154	1289.544	1317.528
AIC w/o Gene4317	1222.902	1193.308	1213.154	1219.852	1203.054	1206.822
p-value of LRT	5.63e-07	4.20e-05	1.70e-06	0.0002	0.0030	0.0004

For the DLBCL data, we also conduct LRT to compare the full models obtained by screening procedures in Table 3.10 and their reduced models based on SCAD and Lasso in Table 3.11. All p-values there are greater than 0.05, which indicates that the penalized sub-models work as well as their corresponding full models.

Moreover, the overlapped gene IDs among the fourty-three potential gene IDs are reported in Table 3.15. The overlapped selected variables based on SJS/ISIS/SIS-SCAD and SJS/ISIS/SIS-Lasso are given in Tables 3.17 and 3.18.

From Table 3.15, we notice that gene 4317 is included in both SJS and ISIS models, but not in models based on SIS. And the smallest screening model obtained

Table 3.15. Overlapped features by three different screening procedures

SJS ∩ ISIS		SJS ∩ SIS			ISIS∩ SIS			SJS∩ISIS∩ SIS	
1662	4317	1662	3813	7357	1188	2579	4131	1662	7343
1681	7343	1664	3820		1456	3799	5055	1681	7357
1825	7357	1680	3824		1662	3811	5301	1825	
3799		1681	3825		1671	3813	5614	3799	
3813		1825	5297		1681	3822	5950	3813	
3824		3799	5953		1682	3824	7343	3824	
3825		3810	7343		1825	3825	7357	3825	
# of genes = 10		15			21			9	

by (SJS∩ISIS∩SIS) only misses gene 4317 compared with model (SJS∩ISIS). We also can conclude that gene 4137 is significant since p-value for the corresponding likelihood ration test is smaller than 0.05 and the likelihood/BIC/AIC criteria of model (SJS∩ISIS) are all better than those of model (SJS∩ISIS∩SIS). Table 3.16 lists the results of the likelihood ratio test (LRT) for model (SJS∩ISIS) and model (SJS∩ISIS∩SIS).

Table 3.16. Likelihood ratio tests for gene 4317

Model	(SJS∩ISIS)	(SJS∩ISIS∩SIS)
Likelihood	-621.0064	-627.0382
BIC	1296.609	1303.213
AIC	1262.013	1272.076
p-value	0.000514	-

Table 3.17. Overlapped genes obtained from SCAD penalty based on three different screening procedures

	SJS ∩ ISIS	SJS ∩ SIS	ISIS ∩ SIS	SJS ∩ ISIS ∩ SIS
Intersect	1681 4317 1825 7357 3825	1825 3810 7357	1825 3824 3799 7357 3822	1825 7357
Counts of genes	5	3	5	2

Table 3.18. Overlapped genes obtained from LASSO penalty based on three different screening procedures

Intersect	SJS ∩ ISIS	SJS ∩ SIS	ISIS ∩ SIS	SJS ∩ ISIS ∩ SIS
Intersect	1662 3825 1681 4317 1825 7343 3813 7357	1664 1825 7343 7357	1188 4131 1456 7343 1671 7357 1825	1825 7343 7357
Counts of genes	8	4	7	3

Likelihood ratio test (LRT) is also applied to compare the overlapped models with their corresponding full models. The results of LRT are summarized in Table 3.19. From Table 3.19, we can first conclude that the post-screening model based on SJS results in largest partial likelihood, and hence is considered the best in terms of likelihood. Moreover, the p-values for SJS models compared with its overlapped models nested in SIS and ISIS models are 3.53e-14 and 3.80e-13, which indicates that the SJS models is significantly different from the overlapped models. In other words, We should not neglect the effects of the other variables in SJS model besides the overlapped ones. Similarly, ISIS model is also significantly from its overlapped

models nested in other procedures. For SIS model, the p-value for SIS/(SIS ∩ SJS) is around 0.09, which considers (SIS ∩ SJS) a reduced but explanatory model for SIS. In other words, sub-SJS model carries the spirit of SIS model indeed. However, sub-ISIS model would miss some information from SIS model since the p-value for SIS/(SIS ∩ ISIS) is around 0.027, smaller than 0.05.

Table 3.19. Likelihood ratio tests for models based on different procedures

	SJS		Cox-ISIS		SIS	
Likelihood	-556.9332		-561.0333		-600.0885	
AIC	1199.866		1208.067		1286.177	
BIC	1348.629		1356.829		1434.939	
Comparison	SJS ∩ SIS	SJS ∩ ISIS	ISIS ∩SIS	ISIS ∩SJS	SIS ∩ ISIS	SIS ∩ SJS
Likelihood	-619.3188	-621.0064	-618.3533	-621.0064	-618.3533	-619.3188
BIC	1320.531	1296.609	1351.358	1296.609	1351.358	1320.531
AIC	1268.638	1262.013	1278.707	1262.013	1278.707	1268.638
p-value	3.53e-14	3.80e-13	1.62e-14	8.39e-12	0.027	0.090

Chapter 4

Conclusion and Future Research

Survival data with ultrahigh dimensional covariates such as genetic markers is of great interest in medical studies and other fields. In this thesis, we propose 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 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.

In this thesis, we focus the empirical performance and application of the proposed methods. We examine the finite sample performance of the proposed pro-

cedure and compare it with existing procedures by Monte Carlo simulations. It is of great interest to investigate the theoretical property of the proposed procedure. This can be an excellent topic for future research.

Bibliography

[1] Akaike, H. (1973). "Information Theory as An Extension of the Maximum Likelihood Principle," Pages 267-281 in B. N. Petrov, and F. Csaki, (Eds.) *Second International Symposium on Information Theory,* Akademiai Kiado, Budapest.

[2] Akaike, H. (1974), "A new look at the statistical model identification," *IEEE Trans. on Automatic Control*, **19**, 716–723.

[3] Andersen, P. K. and Gill, R. D. (1982), "Cox's Regression Model for Counting Processes: A Large Sample Study," *The Annals of Statistics*, **1**0, 1033–1311.

[4] Antoniadis, A. and Fan J. (2001), "Regularization of Wavelets Approximations"(with discussion), *Journal of the American Statistical Association*, **96**, 939–967.

[5] Bennett, S. (1983), "Analysis of Survival Data by the Proportional Odds Model," *Statistics in Medicine,* **2**, 273–277.

[6] Bradic, J., Fan, J. and Jiang, J. (2011),“ Regularization for Cox's Proportional Hazards Model with NP-dimensionality,” *The Annals of Statistics*, **3**9, 3092–3120.

[7] Collett, D. (2003), “Modelling Survival Data in Medical Research (2nd ed.),” *CRC press*.

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

[9] Cox, D. R. (1975), “Partial Likelihood,” *Biometrika,* **62**, 269–276.

[10] Craven, P. and Wahba, G. (1979), “Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-validation,” *Numerische Mathematik,* **31**, 377–403.

[11] Efron, B., Hastie, T., and Tibshirani, R. (2004), “Least angle regression,” *The Annals of Statistics*, **32**, 407–499.

[12] 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.

[13] 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.

[14] Fan, J., and Li, R. (2001), “Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties,” *Journal of the American Statistical Association*, **96**, 1348–1360.

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

[16] 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.

[17] 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*. To appear.

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

[19] Frank, I. E. and Friedman, J. H. (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, **35(2)**, 109–148.

[20] Hoerl, A. E. and Kennard, R. W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, **12(1)**, 55–67.

[21] 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*, **4**1, 1142–1165.

[22] Li, G., Peng, H., Zhang, J. and Zhu, L.-X. (2012), "Robust rank correlation based screening," *The Annals of Statistics*, **4**0, 1846–1877.

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

[24] Mallows, C. (1973), "Some comments on Cp," *Technometrics*, **15**, 661–675.

[25] Miller, A. (2002), "Subset selection in regression, 2nd edition," *New York: Chapman and HALL/CRC*.

[26] Murphy, S. A. and van der Vaart, A. W. (2000), "On Profile Likelihood," *Journal of American Statistical Association*, **95**, 449C-465.

[27] Nishii, R. (1984)," Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression," *The Annals of Statistics,* **12**, 758–765.

[28] Rosenwald A. et al. (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.

[29] Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics,* **19**, 461–464.

[30] Tibshirani, R. (1996), "Regression Shrinkage and Selection via LASSO," *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

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

[32] Vaupel, J.W., Manton, K.G., and Stallard, E. (1979), "The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality," *Demography,* **16**, 439–454.

[33] Wang, H. (2009), "Forward Regression for Ultra-high Dimensional Variable Screening," *Journal of the American Statistical Association*, **1**04, 1512–1524.

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

[35] Xu, C. and Chen, J. (2014), "The Sparse-MLE for Variable Screening in Ultra-High-Dimensional Feature Space," *Journal of the American Statistical Association*, to appear.

[36] Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, **38**, 894–942.

[37] Zhang, H. and Lu, W. (2007), "Adaptive-LASSO for Cox's Proportional Hazards Model," *Biometrika*, **9**4, 1–13.

sparse estimation problems," *Statistical Science*, **27**, 576–593.

[38] 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.

[39] Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, **101**, 1418–1429.

[40] Zou, H. (2008), "A Note on Path-based Variable Selection in The Penalized Proportional Hazards Model," *Biometrika*, **9**5, 241–247.