The Pennsylvania State University
The Graduate School
Eberly College of Science

NEW PROCEDURES FOR COX'S MODEL WITH HIGH DIMENSIONAL PREDICTORS

A Dissertation in
Statistics
by
Ye Yu

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2015

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

Runze Li
Verne M. Willaman Professor, Department of Statistics
Dissertation Advisor, Chair of Committee

Matthew Reimherr
Assistant Professor, Department of Statistics

Lingzhou Xue
Assistant Professor, Department of Statistics

Lan Kong
Associate Professor, Department of Public Health Sciences

Aleksandra Slavkovic
Professor, Department of Statistics
Chair for Graduate Studies of the Department of Statistics

*Signatures are on file in the Graduate School.

Abstract

This thesis studies feature screening and variable selection procedures for ultrahigh dimensional Cox's models and the asymptotic behaviors of tuning parameter selectors, such as the AIC, BIC, and GCV, for penalized partial likelihood.

Survival data with ultrahigh dimensional covariates such as genetic markers have been collected in medical studies and other fields. In our first project, 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 an effective algorithm to carry out the proposed procedure and establish the ascent property of the proposed algorithm. We further prove that the proposed procedure possesses the sure screening property. That is, with the probability tending to one, the selected variable set includes the actual active predictors. We conduct Monte Carlo simulation to evaluate the finite sample performance of the proposed procedure and further compare the proposed procedure and existing SIS procedures. The proposed methodology is also demonstrated through an empirical analysis of a real data example.

Due to the need of studying the theoretical property of variable selection procedure for Cox's model, we study the asymptotic behavior of partial likelihood for the Cox model in our second project. We find that the partial likelihood does not behave like an ordinary likelihood, whose sample average typically tends to its expected value, a finite number, in probability. Under some mild conditions, we prove that the sample average of partial likelihood tends to infinity at the rate of logarithm of the sample size in probability. This is an interesting and surprising

results because the maximum partial likelihood estimate has the same asymptotical behavior as the ordinal maximum likelihood estimate. We further apply the asymptotic results on the partial likelihood to study tuning parameter selection for penalized partial likelihood. Our finding indicates that the penalized partial likelihood with the generalized cross-validation (GCV) tuning parameter proposed in Fan and Li (2002) enjoys the model selection consistency property. This is another surprising result because it is well known that the GCV, AIC and C_p are all equivalent in the context for linear regression models, and are not model selection consistent. Our empirical studies via Monte Carlo simulation and real data example confirms our theoretical finding.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to gratefully and sincerely thank my thesis advisor, Dr. Runze Li, for his guidance in casting light onto 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 members of my doctoral committee, Dr. Lan Kong, Dr. Matthew Reimherr, and Dr. Lingzhou Xue for their input, valuable discussions and accessibility. Without their knowledge and assistance, this study would not have been successful. In addition, I appreciate my home department, the Statistics Department at Penn State University for not only providing me with the great opportunity for graduate study, but also offering a free and friendly atmosphere for me to learn, to communicate and to develop. Moreover, 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.

Last but not least, I would like to thank my parents, Mr. Yongjian Yu and Ms. Shaoping Hong for their support, encouragement, and unwavering love.

Chapter 1

Introduction

1.1 Ultra-high Dimensional Survival Data Analysis

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's 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's 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 dissertation work, we propose a new feature screening procedure for ultrahigh dimensional Cox's 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 dissertation, we will show that the newly proposed procedure can outperform the sure independence screening procedure for the Cox model. This work makes the following major contribution to the literature.

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

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

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

1.2 Asymptotic Behavior of Regularization Parameter Selectors

The Cox model (Cox, 1972) has been the most popular model in the survival data analysis during the past decades. In practice, many risk factors and covariates are available for the initial analysis, thus an important and challenging task is to identify the significant risk factors and covariates which impact the hazard function. Therefore, variable selection is a fundamental problem in the analysis of survival data in the presence of many covariates. Several methods for such a problem may be or have been considered. For instance, variable selection criteria, such as the AIC and BIC, for linear regression models can be naturally extended for the Cox model by replacing the log-likelihood in the AIC (Akaike, 1973) and BIC (Schwarz, 1978) by log-partial likelihood to deal with censored survival data analysis. Tibshirani (1997) proposed LASSO for the Cox model; Faraggi and Simon (1998) proposed Bayesian variable selection methods for censored survival data by extending the idea of Lindley (1968). In particular, Fan and Li (2002) proposed the SCAD procedure for the Cox model by using nonconcave penalized partial likelihood, including the SCAD penalized partial likelihood, which enjoys the oracle property. Zhang and Lu (2007) and Zou (2008) further proposed adaptive LASSO for the Cox model to improve the SCAD procedure in terms of computational efficiency, while retaining the oracle property. However, the oracle property depends on the choice of the regularization parameter in these proposed procedures.

It is well known that the regularization parameter controls the model complexity of the selected models, thus it plays a crucial role in the proposed procedures above mentioned. To our best knowledge, the issue of regularization parameter selection for penalized partial likelihood has not been systematically studied, though the asymptotic properties of the partial likelihood estimator have extensively studied (Tsiatis, 1981; Andersen and Gill, 1982; Takemi and Toshinari, 1984). Under mild regularity condition, the maximum partial likelihood estimator behaves the same as the ordinary maximum likelihood estimator of independent and identically distributed random samples in terms of asymptotic consistency, asymptotic normality and asymptotic efficiency. See, for example, Murphy and van der Vaart (2000). In this dissertation, we study the asymptotic behavior of the partial likeli-

hood, and prove that the '*sample average*' of partial likelihood diverges to infinity at a rate of logarithm of the sample size. This clearly indicates that the Cox's partial likelihood does not behave like an ordinary likelihood in that under mild regularity conditions, the sample average of the ordinary likelihood function converges to its expectation (a finite value) in probability as the sample size tends to infinity.

We then study the asymptotic property of tuning parameter selectors, such as the AIC, BIC, and GCV of penalized partial likelihood for Cox's model. The method of the *generalized cross-validation* (GCV) was used by Fan and Li (2001) to select the regularization parameter in their nonconvex penalized least squares and nonconcave penalized likelihood. Wang, Li and Tsai (2007) studied the selection of regularization parameter in the SCAD penalized least squares for linear regression models and partially linear model. They showed that with a positive probability, the GCV selector yields an over-fitted model, thus the SCAD penalized least squares with the GCV selector does not enjoy the oracle property. However, in this dissertation work, we will prove the consistency of GCV selector in the nonconcave penalized partial likelihood for the Cox model. This work makes the following major contribution to the literature.

(a) We study the asymptotic properties of Cox's partial likelihood, which is needed for the study of the asymptotic behavior of the regularization parameter selector. We find that Cox's partial likelihood does not behave like an ordinary likelihood in the sense that the '*sample average*' of partial likelihood function diverges to infinity, which is in contrast to the well-known fact that under mild regularity conditions, the sample average of the ordinary likelihood function converges to its expectation (a finite value) in probability as the sample size $n \to \infty$. This is an interesting and surprising result because it has been shown that in most usual senses, Cox's partial likelihood behaves asymptotically like an ordinary likelihood; see Murphy and van der Vaart (2000).

(b) We study the issue of regularization parameter selection of penalized partial likelihood for the Cox model. We find that the GCV selector proposed in Fan and Li (2002) enjoys the model selection consistency for the penalized

partial likelihood in the Cox model, which is in contrast to its model selection inconsistency in the least squares setting as demonstrated in Wang, Li and Tsai (2007). This is another surprising result also because the GCV is equivalent to the AIC and the C_p in the context for linear regression models, and it is well known that both AIC and C_p yields an overfitted model with a positive probability, thus are not model selection consistent.

Monte Carlo simulation is further conducted to assess the finite sample performance of the GCV selecetor and compare its performance with AIC and BIC for the Cox model. Our numerical results indicate that GCV selector performs similar to BIC selector, and is consistent in model selection for the Cox model. An empirical real data example is demonstrated in the end.

1.3 Organization of Dissertation

The rest of the dissertation is organized as follows. In Chapter 2, we give a detailed review of the existing methods in literature about three foregoing topics: the traditional variable selection methods and the penalized variable selection methods in high dimensional data analysis, the feature screening procedures for ultra-high dimensional data analysis, and the high dimensional Cox's survival data analysis. 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. The technical proofs are also provided in details for the aforementioned properties of the new procedure. Numerical comparisons and an empirical analysis of a real data example are then presented. In Chapter 4, We prove that Cox's partial likelihood does not behave like an ordinary likelihood in the sense that the '*sample average*' of partial likelihood function diverges to infinity. We further show that the asymptotic behavior of the GCV selectors for penalized partial likelihood under the Cox model setting would be different from that under the linear or generalized linear settings. We also present our numerical studies, which have shown that GCV is a consistent selector for penalized partial likelihood. Conclusion remarks and several future work are listed in Chapter 5.

Chapter 2

Literature Review

The review of literature is organized as follows. First, we introduce several variable selection methods for linear model, including classical variable selection and penalized regression approach. We then extend the penalized methods to generalized linear models. Feature screening methods and techniques for ultrahigh-dimensional data analysis are reviewed in the second part. We then present a brief review of some basic concetps 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 generalized linear and 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 penalized fit criteria. They are measures of fit with a penalty for each additional parameter. The measures of fit could be defined as

$$RSS = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2,$$

where $\widehat{y}_i$ is the fitted value for y_i.

Let RSS_p denote the RSS with p variables ($1 \leq p \leq d$) in the model and RSS_0 is the RSS when a constant being fitted. Based on RSS, there are several commonly used penalized fit criteria for variable selection.

The *adjusted* R^2 statistic with p variables is defined as

$$A_p = 1 - (1 - R_p^2)\frac{n-1}{n-p},$$

where $R_p^2 = 1 - \frac{RSS_p}{RSS_0}$.

The C_p statistic (Mallows,1973) is defined as

$$C_p = \frac{RSS_p}{\sigma^2} - (n - 2p),$$

where σ^2 is usually be replaced by its unbiased estimate under the full model in practice.

$$\widehat{\sigma}^2 = \frac{RSS_d}{n-d}.$$

Consider the Akaike's Information Criterion (AIC, Akaike, 1973, 1974) in the context of linear model. The AIC_p statistic for linear model with p variables is defined as

$$AIC_p = RSS_p + 2p\sigma^2,$$

where σ^2 is still could be estimated by $\widehat{\sigma}^2$. Hence, minimization of AIC_p with respect to p is equivalent to minimization of C_p.

Similar extension of Bayesian Information Criterion (BIC, Schwarz, 1978) yields BIC_p statistic in linear model, which is defined as

$$BIC_p = RSS_p + \log(n)p\sigma^2.$$

Compared with AIC, BIC penalizes higher for more complicated models with larger p. Hence the BIC tends to favor smaller models than AIC.

Moreover, Risk Inflation Criterion (RIC, Foster and George, 1994) for linear model is defined as

$$RIC_p = RSS_p + 2\log(d)p\sigma^2.$$

General information criterion (GIC, Nishii, 1984) for linear model is defined as

$$GIC_{p,\kappa_n} = \log\widehat{\sigma}^2 + \frac{1}{n}\kappa_n p,$$

where κ_n is a positive number that controls the properties of variable selection. Note that when $\kappa_n = 2$, GIC becomes AIC, while $\kappa_n = \log(n)$ leads GIC to be BIC. We can see that the larger κ_n is, the larger the penalty is for models with more variables.

A good regression model should also give predictions as accurate as possible, some criteria based on measurements of predictability are also investigated by statisticians.

The prediction sum of squares (PRESS, Allen, 1974) is defined as

$$PRESS_p = \sum_{i=1}^{n}(y_i - \widehat{y}_{ip})^2,$$

where $\widehat{y}_{ip}$ is the predicted value for y_i based on a subset of p-variables. In practise, the cross validation (CV) approach, including leave-one-out cross validation and K-fold cross validation, may be used to estimate the PRESS statistic. Cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subsets (called the validation set or testing set). To reduce variability, cross-validation are performed repeatedly with respect to different partitions, and the validation results are averaged. It can be shown that under mild

conditions, PRESS statistic for a linear model can be asymptotically equivalent to generalized cross validation score (GCV, Craven and Wahba, 1979), which would be covered in details later.

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 can not 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.

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.

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.

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 $\text{argmin}_\theta\{p_\lambda^{'}(|\theta|) + |\theta|\} = 0$.

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

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

$$= \lambda|\beta_j| I(|\beta_j| \leq \lambda) - \frac{(|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2)}{2(a-1)} I(\lambda < |\beta_j| \leq a\lambda)$$
$$+ \frac{(a+1)\lambda^2}{2} I(|\beta_j| > a\lambda)$$

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(noncancave) symmetric spline function singular at origin with knots 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.

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, Soft thresholding, 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 Soft, 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.

2.1.3 Computational Algorithms

Among the penalty functions aforementioned, two mainstream terms are the LASSO (L_1) and the nonconvex penalization such as the SCAD and the MCP. A strong irrepresentable condition is necessary for the LASSO to be selection consistent (Zou, 2006), while nonconcave penalty achieves the variable consistency with correcting the bias of LASSO (Fan and Li, 2001; Zhang, 2010). The LASSO owns its popularity largely to its computational properties. Tibshirani (1996) proposed

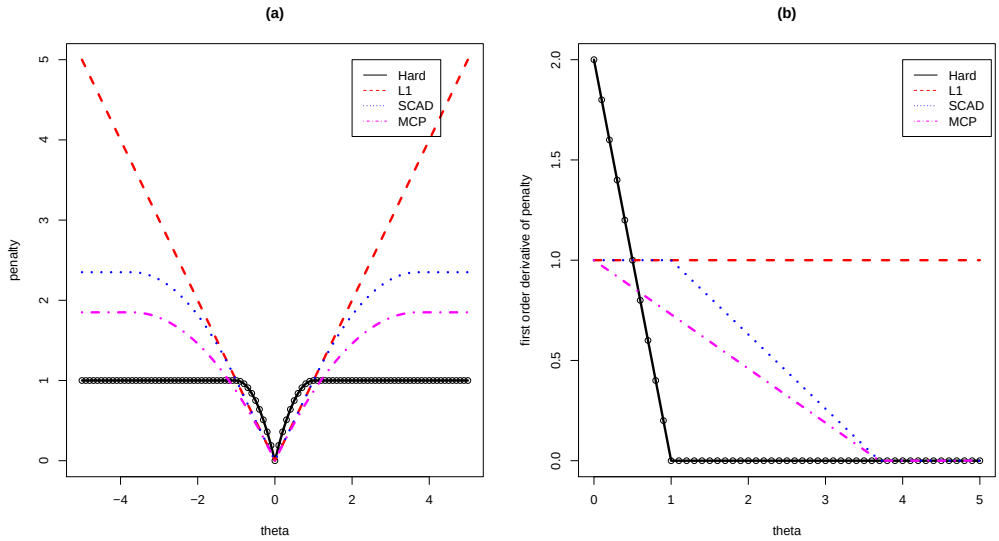

Figure 2.1. Example penalty functions and their corresponding first order deriveatives. (a) Plots the penalty functions of the Hard thresholding penalty, the Soft thresholding penalty, the SCAD penalty and the MCP. (b) illustrates the corresponding derivatives of each penalty function in (a).

an algorithm for solving contrained least squares problems of LASSO, whereas Fu (1998) provided a "shooting algorithm". The "Least Angel Regression" ("LARS") was suggested by Efron et al.,(2004), which could also be applied to solve the optimization problem with the adaptive LASSO penalty. Furthermore, the coordinate descent algorithm has been shown to be very useful and efficient for a more general class of L_1 penalization problems (Friedman et al.,2010).

The nonconcave penalty functions have recently attracted much interest since they can eliminate the estimation bias for parameters with large values and attain oracle properties with more refined rates of convergence. The computation for nonconcave penalization is much more complicated since the resulting optimization problem is usually nonconcave and has multiple local minimizers. In this subsection, we would discuss several main algorithms for minimization problems of PLS.

Local Quadratic Approximation (LQA) algorithm was proposed by Fan and Li (2001). Suppose we are given an initial value $\boldsymbol{\beta}^0$ that is close to the minimizer of (2.1.2). If β_j^0 is very close to 0, then set $\widehat{\beta}_j = 0$. Otherwise they can be locally

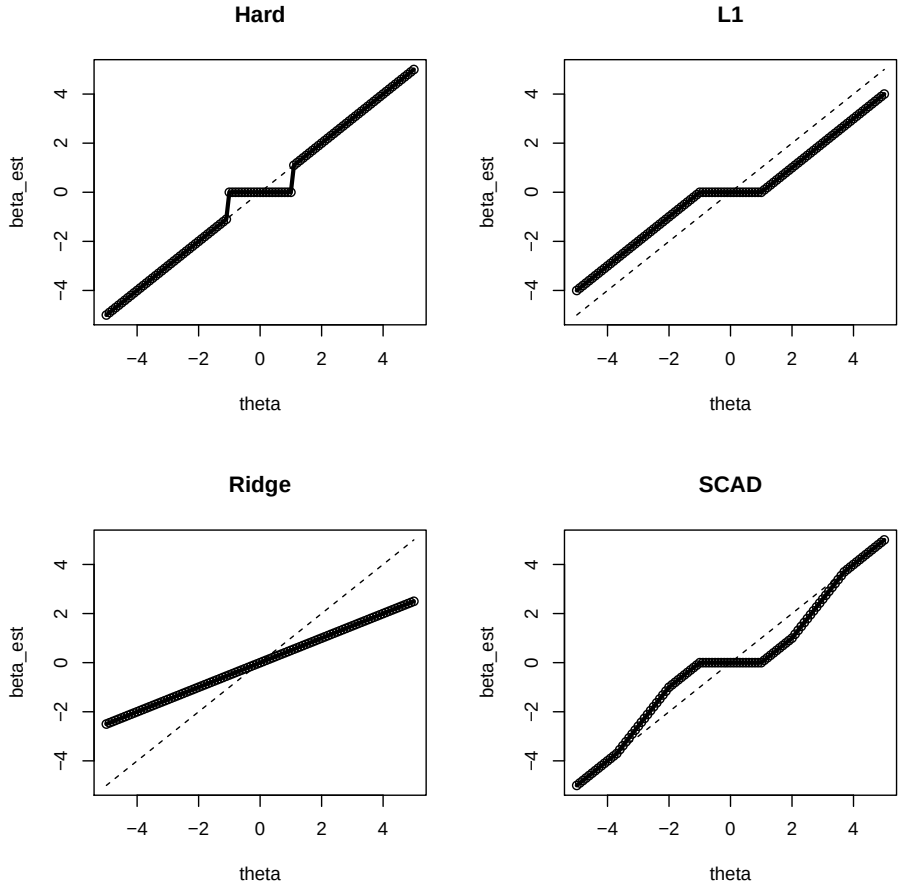

Figure 2.2. The relationship between PLS and OLS estimates when the design matrix is orthonormal.

approximated by a quadratic functions as

$$[p_\lambda(|\beta_j|)]' = p'_\lambda(|\beta_j|)\text{sgn}(\beta_j) \approx \{p'_\lambda(|\beta_j|)/|\beta_j|\}\beta_j,$$

where $\beta_j \neq 0$. In other words,

$$p_\lambda(|\beta_j|) = p_\lambda(|\beta_j^0|) + \frac{1}{2}\{p'_\lambda(|\beta_j^0|)/|\beta_j^0|\}(\beta_j^2 - (\beta_j^0)^2). \tag{2.1.3}$$

Note that the first term in (2.1.2) is convex with respect to $\boldsymbol{\beta}$, therefore (2.1.2) could be locally approximated (except for a constant term) by

$$Q(\boldsymbol{\beta}) = \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \frac{1}{2}n\boldsymbol{\beta}^T\Sigma_\lambda(\boldsymbol{\beta}^0)\boldsymbol{\beta}, \tag{2.1.4}$$

where $\Sigma_\lambda(\boldsymbol{\beta}_0)$ =diag$\{p'_\lambda(|\beta_1^0|)/|\beta_1^0|, \ldots, p'_\lambda(|\beta_d^0|)/|\beta_d^0|\}$, β_j^0 are the jth components of initial value $\boldsymbol{\beta}_0$. The solution for PLS (2.1.4) can be found by iteratively computing the ridge regression

$$\boldsymbol{\beta}^{(k+1)} = \{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\boldsymbol{\beta}^{(k)})\}^{-1}\mathbf{X}^T\mathbf{y} \qquad k = 0, 1, \cdots .$$

However, as pointed out in Fan and Li (2001) and Hunter and Li (2005), a drawback of LQA algorithm is that once a covariate is deleted at any step in the LQA algorithm, it would necessarily be excluded from the final selected model. Hunter and Li (2005) addressed this issue by optimizing a slightly perturbed version of LQA, which alleviates the aforementioned drawback, but it is difficult to choose the size of perturbation. To overcome the computational difficulty, instead of using the fully iterative method, Fan and Li (2001) suggests using the estimator obtained by LQA algorithm with a few iterations based on a good initial, which could be the unpenalized estimate.

To eliminate the weakness of the LQA, Zou and Li (2008) constructed an efficient one-step sparse estimation procedure based on *Local Linear Approximation (LLA)*, which enjoys three significant advantages over the LQA algorithm. First, LLA does not delete any small coefficient and hence avoid the numerical instability. Secondly, they demonstrated that LLA is the best minorization-maximization (MM) algorithm. Thirdly, the LLA naturally produces a sparse estimates via continuous penalization. According to LLA, the penalty function could be locally approximated by

$$p_\lambda(|\beta_j|) = p_\lambda(|\beta_j^0|) + p'_\lambda(|\beta_j^0|)(|\beta_j| - |\beta_j^0|), \qquad \text{for} \quad \beta_j \approx \beta_j^0. \tag{2.1.5}$$

Similar to the LQA algorithm, the minimization of the PLS can be carried out as follows. Set the initial value $\boldsymbol{\beta}^0$ be the OLS estimates and then repeatedly solve the local linear approximation function for $k = 0, 1, \cdots$

$$\boldsymbol{\beta}^{(k+1)} = \text{argmin}_{\boldsymbol{\beta}} \left\{ \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + n\sum_{j=1}^{d} p'_\lambda(|\beta_j^k|)|\beta_j| \right\}. \tag{2.1.6}$$

From (2.1.6), we can see that the LLA algorithm inherits the good features of

weighted LASSO, therefore the minimization could be solved by the exiting efficient algorithms, such as "LARS" (Efron et al., 2004). Moreover, LLA automatically adopts a sparse estimator. The one-step LLA estimator ($k = 0$ for (2.1.6)) is as efficient as the fully iterative one given a good initial value, which avoids expensive computational cost.

Both LQA and LLA algorithms are related to the MM principle (Hunter and Li, 2005). There are several other useful algorithms developed for computing the nonconcave penalized estimators than LQA and LLA. Zhang (2010) proposed a PLUS algorithm for solving the PLS with the MCP and the SCAD. Breheny and Huang (2011) investigate the application of coordinate descent algorithms to the SCAD and the MCP penalization problems. Fan and Peng (2004) showed that the SCAD estimator possesses the oracle property even with diverging number of parameters, namely, $p \propto n^{\alpha}$ ($\alpha < \frac{1}{2}$). The Dantzig Selector (Candes and Tao, 2007) was proposed for PLS with L_1 penalty when $d >> n$. Kim et al. (2008) applied the CCCP algorithm (An and Tao 1997; Yuille and Rangarajan 2003) for solving the SCAD problem after decomposing the PLS as the sum of the convex and concave function under high dimensional case. More recently, Zhang and Zhang (2012) provided statistical guarantees concerning global optima of PLS with various nonconcave penalties and proposed that gradiant descent initialized at a LASSO optimum could be used to obtained specific local minima with well behaviors. In the same spirit, Fan et.al (2014) showed that if the LLA algorithm is initialized at a LASSO optimum that satisfies certain properties, then the two-stage procedure produces an oracle solution for various nonconcave penalties. Loh and Wainwright (2014) proved that any local optimum of the composite objective function lies within statistical precision of the true parameter vector as long as the loss function satisfies restricted strong convexity (RSC) and the penalty function satisfies suitable regularity conditions. Wang, Liu, and Zhang (2014) proposed an algorithm based on LLA and proximal-gradient method (Nesterov, 2007) to obtain the local solution for PLS that attain a global geometric rate of convergence for calculating the entire regularization path without relying on the quality of the initial lasso solution. Wang, Kim, and Li (2013) investigated a calibrating CCCP algorithm to produce a consistent solution path which contains the oracle estimator with probability approaching to one based on the tuning parameter selected by

high-dimensional BIC criterion.

2.1.4 Tuning Parameters Selection

To implement the methods described in the last two subsections, we need to determine the values of tuning parameters in the penalty function. In this section, we mainly focus on the selection of tuning parameter λ. For the SCAD penalty, another parameter a is usually fixed at its empirical value 3.7, and for the MCP, the choice of a is largely determined by the available computational resources as long as the method reaches a sufficient small λ. We refer our readers to Fan and Li (2001) and Zhang (2010) for details.

For each penalty function, given λ, PLS estimate $\widehat{\boldsymbol{\beta}}_\lambda$ could be obtained by applying algorithms discussed above. Model selection critera introduced in section 2.1.1 then could be written as a function of $\widehat{\boldsymbol{\beta}}_\lambda$, namely, λ. We tend to favor λ that results in relatively small value of these model selectors.

AIC, BIC, and K-fold cross validation (CV) are the three popular model selectors applied in practise, which could be expressed as

$$AIC(\lambda) = ||\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_\lambda||^2 + 2df_\lambda \widehat{\sigma}^2, \tag{2.1.7}$$

$$BIC(\lambda) = ||\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_\lambda||^2 + \log(\mathrm{n}) df_\lambda \widehat{\sigma}^2, \tag{2.1.8}$$

$$CV(\lambda) = \sum_{\nu=1}^{K} ||\mathbf{y}_\nu - \mathbf{X}_\nu^T \widehat{\boldsymbol{\beta}}_\lambda^\nu||^2, \tag{2.1.9}$$

where df_λ in (2.1.7) and (2.1.8), is the estimated degrees freedom in selected model who depends on the value of λ. And $\mathbf{y}_\nu, \mathbf{X}_\nu^T \widehat{\boldsymbol{\beta}}_\lambda^\nu$ in (2.1.9) are corresponding to the testing set and the estimated coefficients obtained from the training set.

Unlike AIC and BIC, K-fold cross validation leads to heavy computational burden with larger K because K models should be built for each specific λ. To save the computational cost, generalized cross validation (GCV, Craven and Wahba, 1979) was introduced, which is asymptotically equivalent to the CV for linear estimator. GCV is defined by

$$GCV(\lambda) = \frac{1}{n} \frac{||\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_\lambda||^2}{\{1 - df_\lambda/n\}^2}. \tag{2.1.10}$$

For PLS, df_λ in all those selectors is defined by

$$df_\lambda = \text{trace}\{\mathbf{X}_\lambda(\mathbf{X}_\lambda^T\mathbf{X}_\lambda + n\Sigma_\lambda(\widehat{\beta}_\lambda))^{-1}\mathbf{X}_\lambda^T\},$$

where $\mathbf{X}_\lambda$ and $\widehat{\beta}_\lambda$ are the design matrix and the PLS estimate with a given λ. One may simply define $df_\lambda = p$, the number of significant predictors in the final selected model.

In summary, we may select the optimal tuning parameter λ by

$$\widehat{\lambda} = \text{argmin}_\lambda\{\text{Selector}(\lambda)\}. \tag{2.1.11}$$

The selection criteria are usually classified into two categories: consistent ones and efficient ones. A consistent criterion identifies the true model with a probability that approaches 1 as $n \to \infty$ when a set of candidate models contains the true model. An efficient criterion selects the model so that its average squared error is asymptotically equivalent to the minimum offered by the candidate models when the true model is approximated by a family of candidate models. Detailed discussion on efficiency and consistency can be found in Shibata (1981, 1984), Li (1987) and Shao (1997) and McQuarrie and Tsai (1998).

In the linear regression models, Wang et al. (2007) demonstrated that Fan and Li's (2001) GCV-selector for the PLS with SCAD penalty can not select the tuning parameter consistently. They found that GCV, behaves similarly to AIC, tends to result in an overfitting selected model. They further proposed a BIC-type tuning parameter selector as $BIC_\lambda^\star = \log(\widehat{\sigma}_\lambda^2) + \frac{1}{n}\log(n)df_\lambda$, where $\widehat{\sigma}_\lambda^2$ is the corresponding mean of residual errors based on the selected model. This BIC-type tuning parameter selector is proved to possess the desirable oracle property and be able to identify the finite-dimensional true linear models consistently.

2.2 Variable Selection via Penalized Likelihood

The methodology and algorithms discussed in PLS can be applied directly to many other statistical contexts. In this section, we would mainly consider the likelihood-based generalized linear model. For generalized linear model (GLM, McCullagh and Nelder, 1989), statistical inferences are based on underlying likelihood func-

tions. The penalized maximum likelihood estimator can be used to select significant variables.

Assume that the data $\{\mathbf{x}_i, y_i\}$ are collected independently. Conditioning on $\mathbf{x}_i, y_i$ has a density $f_i(g(\mathbf{x}_i^T\boldsymbol{\beta}), y_i)$, where g is a known link function. Let ℓ_i=logf_i denote the conditional log-likelihood of y_i. The penalized likelihood can be expressed as

$$\sum_{i=1}^{n} \ell_i(g(\mathbf{x}_i^T\boldsymbol{\beta}), y_i) - n\sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{2.2.12}$$

Similar to PLS, we obtained the penalized maximum likelihood estimator by maximizing (2.2.12) with respect to $\boldsymbol{\beta}$ for some tuning parameter λ. We mainly focus on LQA and LLA algorithm in the following discussion.

Define our goal function $Q(\boldsymbol{\beta})$ by

$$Q(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - n\sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{2.2.13}$$

where $\ell(\boldsymbol{\beta})$ denotes the corresponding log likelihood function. All the commonly used penalty functions shown in Figure 2.1(a) are singular at the origin, and they do not have continuous second order derivatives. Therefore, penalized partial likelihood is an object function that is nonsmooth and nonconvex, which poses challenges in searching for maximizer of (2.2.13) . However, $p_\lambda(\cdot)$ can be locally approximated by a convex function and hence local maximum of (2.2.13) could be obtained.

Now assume the likelihood function is smooth with respect to $\boldsymbol{\beta}$ so that its first two partial derivatives are continuous. Adopting LQA algorithm, (2.2.13) could be locally approximated (except for a constant term) by

$$\ell(\boldsymbol{\beta}^0)+\nabla\ell(\boldsymbol{\beta}^0)^T(\boldsymbol{\beta}-\boldsymbol{\beta}^0)+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^0)^T\nabla^2\ell(\boldsymbol{\beta}^0)(\boldsymbol{\beta}-\boldsymbol{\beta}^0)-\frac{1}{2}n\boldsymbol{\beta}^T\Sigma_\lambda(\boldsymbol{\beta}^0)\boldsymbol{\beta}, \tag{2.2.14}$$

where $\Sigma_\lambda(\boldsymbol{\beta}_0)$ =diag$\{p'_\lambda(|\beta_1^0|)/|\beta_1^0|, \ldots, p'_\lambda(|\beta_d^0|)/|\beta_d^0|\}$, β_j^0 are the jth components of initial value $\boldsymbol{\beta}^0$. $\nabla\ell(\boldsymbol{\beta}^0)$ and $\nabla^2\ell(\boldsymbol{\beta}^0)$ are the first two partial derivatives of the likelihood function in (2.2.13).

Adopting the Newton-Raphson algorithm, the quadratic maximization problem

(2.2.14) yields the solution

$$\boldsymbol{\beta}^1 = \boldsymbol{\beta}^0 - \{\nabla^2\ell(\boldsymbol{\beta}^0) - n\Sigma_\lambda(\boldsymbol{\beta}_0)\}^{-1}\{\nabla\ell(\boldsymbol{\beta}^0) - \mathbf{U}_\lambda(\boldsymbol{\beta}^0)\}, \tag{2.2.15}$$

where $\mathbf{U}_\lambda(\boldsymbol{\beta}^0) = \Sigma_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\beta}^0$. The partial likelihood estimator can be obtained when the algorithm converges. Please note that for each iteration, we delete the corresponding covariate when $|\widehat{\beta}_j| < \eta$. η is commonly set to be the estimated standard error of $\widehat{\beta}_j$, which can be obtained by sandwich formula (Fan and Li, 2001). Since the one-step procedure is as efficient as the fully iterative method with a good initial value, we do not have to iterate the algorithm until converges. The unpenalized MLE are usually considered as a good initial value.

Instead of LQA, LLA approximates (without the constant term) (2.2.13) by

$$\ell(\boldsymbol{\beta}^0) + \nabla\ell(\boldsymbol{\beta}^0)^T(\boldsymbol{\beta} - \boldsymbol{\beta}^0) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^T\nabla^2\ell(\boldsymbol{\beta}^0)(\boldsymbol{\beta} - \boldsymbol{\beta}^0) - n\sum_{j=1}^{d} p'_\lambda(|\beta_j^0|)|\beta_j|,$$

which could be expressed as

$$\ell(\boldsymbol{\beta}^0) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^T\nabla^2\ell(\boldsymbol{\beta}^0)(\boldsymbol{\beta} - \boldsymbol{\beta}^0) - n\sum_{j=1}^{d} p'_\lambda(|\beta_j^0|)|\beta_j|, \tag{2.2.16}$$

with $\boldsymbol{\beta}^0 = \widehat{\boldsymbol{\beta}}_{MLE}$.

Express $-\nabla^2\ell(\boldsymbol{\beta}^0) = \mathbf{X}^T\mathbf{D}\mathbf{X}$, for example, D= diag$\{\widehat{p}_i(1-\widehat{p}_i)\}_n$ in logistic regression. Hence maximizing (2.2.16) is equivalent to minimizing

$$\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^T\mathbf{X}^T\mathbf{D}\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^0) + n\sum_{j=1}^{d} p'_\lambda(|\beta_j^0|)|\beta_j|, \tag{2.2.17}$$

which can be expressed as

$$\frac{1}{2}(\mathbf{y}^\star - \mathbf{X}^\star\boldsymbol{\beta})^T(\mathbf{y}^\star - \mathbf{X}^\star\boldsymbol{\beta}) + n\sum_{j=1}^{d} p'_\lambda(|\beta_j^0|)|\beta_j|, \tag{2.2.18}$$

with $\mathbf{y}^\star = \mathbf{D}^{\frac{1}{2}}\mathbf{X}\boldsymbol{\beta}^0, \mathbf{X}^\star = \mathbf{D}^{\frac{1}{2}}\mathbf{X}$. It is obvious that the penalty term in (2.2.18) is actually a form of adaptive LASSO, and therefore its local minimum could be

searched by LARS algorithm with specific λ.

Tuning Parameter Selection The algorithm for tuning parameter selection is described as follows:

Step 1. Choose a grid point set for λ, say $(\lambda_1, \cdots, \lambda_s)$, and let $i = 1$.

Step 2. With λ_i, compute the $\widehat{\boldsymbol{\beta}}$ using the iterative LQA or one-step LLA algorithm.

Step 3. Compute the variable selection criteria, such as BIC and GCV, with $\lambda = \lambda_i$. Let $i = i + 1$.

Step 4. Repeat **Step 2** and **3** until all s grid points are exhausted.

Step 5. The final estimator for $\boldsymbol{\beta}$ is the one yields the lowest selector value.

Under the likelihood setting, the definition of AIC, BIC and GCV statistic can be defined as

$$AIC(\lambda) = -2\ell(\widehat{\boldsymbol{\beta}}) + 2df_\lambda,$$

$$BIC(\lambda) = -2\ell(\widehat{\boldsymbol{\beta}}) + \log(\mathrm{n})df_\lambda,$$

$$GCV(\lambda) = \frac{1}{n}\frac{-\ell(\widehat{\boldsymbol{\beta}})}{\{1 - df_\lambda/n\}^2},$$

where $df_\lambda = trace\left[\left\{\nabla^2\ell(\widehat{\boldsymbol{\beta}}) + \Sigma_\lambda(\widehat{\boldsymbol{\beta}})\right\}^{-1}\nabla^2\ell(\widehat{\boldsymbol{\beta}})\right]$.

Zhang, Li, and Tsai (2010) studied the issues of tuning parameter selection for more general case with a nonconcave penalty. They proposed a more general GIC-type tuning parameter selector as

$$GIC_{p,\kappa_n} = \frac{1}{n}\{G(\mathbf{y}, \widehat{\boldsymbol{\beta}}_p) + \kappa_n p\}, \tag{2.2.19}$$

where $G(\mathbf{y}, \widehat{\boldsymbol{\beta}}_\lambda)$ measures the fitting of model. This general GIC provides more flexibility for practitioners to employ their own favored selector by plugging different values of κ_n. They derived the consistent results with the features of AIC, BIC and GCV studied in Wang et al. (2007) as summarized in section 2.1.4. They found that for a partial likelihood with the nonconvex penalty, BIC-type selector enables us to identify the true model consistently when the true model is among

set of candidate models, whereas the AIC-type elector is asymptotically loss efficient if the true model is approximated by a set of candidate models with the linear structure, and GCV-type selector tends to yield an overfitted model with a positive probability and therefore does not possess the oracle property.

2.3 Feature Screening for Ultrahigh Dimensional Feature Space

The ultra-high dimensional data analysis is of great importance in diverse frontiers of research, ranging from computational biology and health studies to financial engineering and risk management. In these cases, the dimension p can grow much faster than the sample size n such that many models are not even identifiable, and such a demand from application brings a lot of challenge to statistical inference. By ultrahigh dimensionality, we mean $p = O(\exp(an))$ for some $a > 0$. And from this section on, the letter p no longer stands for the selected-model size from the full d-dimensional model as in the previous sections, it becomes the ultrahigh dimension defined above. And d is used to denote the moderate dimensionality, which usually has $d = o(n)$.

When dimension p is high or ultrahigh, it is often assumed that only a small number of predictors contribute to the response, which leads to the sparsity of the parameter vector. Therefore, variable selection plays a prominent role in high or ultra-high statistical modeling. In section 2.1 and 2.2, we have reviewed many variable selection techniques in high dimensional settings. Most of them are based on the penalized likelihood approach, such as LASSO (Tibashirani, 1996), the SCAD and other folded-concave penalty (Fan and Li, 2001; Fan and Peng, 2004; Kim et al., 2008), the Dantzig selector (Candes and Tao, 2007) and their related methods (Zou, 2006; Zou and Li, 2008). However, in ultrahigh statistical learning problem, these methods may not perform well due to the simultaneous challenges of computational expediency, statistical accuracy and algorithmic stability (Fan, Samworth and Wu, 2009). To address these issues, a two-stage approach is applied. Namely, we first conduct feature screening procedure to reduce the dimensionality from ultrahigh to below sample size and then the refined variable selection methods

could be applied to select the important variables. In this section, we focus on the first step of the two-stage approach, where independent screening procedures are used to achieve dimension reduction by ranking features of predictors according to their marginal utilities.

2.3.1 Sure Independence Screening for Linear Models

Fan and Lv (2008) proposed a sure screening method based on a correlation learning, called the Sure Independence Screening (SIS), and developed its sure screening property in linear model settings.

Consider the linear model (2.1.1), but now the dimension of $\mathbf{X}$ becomes $n \times p$, where $p = O(\exp(an))$ for some $a > 0$. The goal of SIS is to reduce dimensionality p from a huge scale to a relatively large scale $d = o(n)$ that is below sample size n by a fast and efficient method described below.

Let $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_p)^T$ be the p-vector obtained by the componentwise regression

$$\boldsymbol{\omega} = \mathbf{X}^T\mathbf{y}. \tag{2.3.20}$$

where the design matrix $\mathbf{X}$ and $\mathbf{y}$ are first standardized for simplicity to gain insight of this procedure. Hence $\boldsymbol{\omega}$ is indeed a vector proportional to the marginal correlations between predictors with the response variable. For any given $\gamma \in (0, 1)$, the p componentwise magnitudes of the vector $\boldsymbol{\omega}$ could be sorted in a decreasing order, which defines a submodel by

$$\mathcal{M}_\gamma = \{1 \leq i \leq p, |\omega_i| \text{ is among the first } [\gamma n] \text{ largest of all}\}, \tag{2.3.21}$$

where the $d = [\gamma n] < n$ denotes the integer part of γn. The full model is shrunken down to the submodel $\mathcal{M}_\gamma$ with sample size d after this procedure, which is so called SIS. In practice, the value of d is usually determined by $n - 1$ or $n/\log n$ to be conservative.

Let $\mathcal{M}_\star$ represent the true sparse model, Fan and Lv (2008) establish the sure screening property (SSP) for SIS, i.e.,

$$P(\mathcal{M}_\star \subset \mathcal{M}_\gamma) \to 1 \text{ as n} \to \infty.$$

for some given γ, where $\gamma \to 0$ as n tends to infinity. *Theorem* 1 in Fan and Lv (2008) shows how to choose the sequence of γ_n under some regularity conditions. Namely, $\gamma \sim cn^{\theta}$, where $c > 0, \theta < 1 - 2\kappa - \tau$ with $2\kappa - \tau < 1, \kappa \geq 0$, and $\tau \geq 0$. Therefore, it is reasonable to work on $\mathcal{M}_{\gamma}$ to further select significant variables and identify the underlying model structure in the second step of the two-stage approach such as SIS-SCAD and SIS-LASSO.

The iteratively thresholded ridge regression screener (ITRRS) is introduced in Fan and Lv (2008) to better understand the rationale of SIS. Define the ridge estimator

$$\boldsymbol{\omega}^{\lambda} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y},$$

where $\lambda > 0$. It is obvious that $\boldsymbol{\omega}^{\lambda} \to \widehat{\boldsymbol{\beta}}_{LS}$ as $\lambda \to 0$ and $\lambda\boldsymbol{\omega}^{\lambda} \to \boldsymbol{\omega}$ as $\lambda \to \infty$. Here $\widehat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}^T\mathbf{X})^{+}\mathbf{X}^T\mathbf{y}$, where $(\mathbf{X}^T\mathbf{X})^{+}$ denotes the generalized inverse of $\mathbf{X}^T\mathbf{X}$. Therefore the componentwise regression estimator is a specific case of the ridge regression and the ranking of $|\boldsymbol{\omega}^{\lambda}|$ is the same as $|\lambda\boldsymbol{\omega}^{\lambda}|$. Given $\delta \in (0, 1)$, define

$$\mathcal{M}^1_{\delta,\lambda} = \{1 \leq i \leq p : |\omega_i^{\lambda}| \text{ is among the first } [\delta p] \text{ largest of all}\}. \tag{2.3.22}$$

The ITRRS can be described as follows:

Step 1. Carry out the procedure in (2.3.22) to the full model to get a submodel $\mathcal{M}^1_{\delta,\lambda}$ with size $[\delta p]$.

Step 2. Apply a similar procedure to the model $\mathcal{M}^1_{\delta,\lambda}$ and obtain a submodel $\mathcal{M}^2_{\delta,\lambda} \subset \mathcal{M}^1_{\delta,\lambda}$ with size $[\delta^2 p]$, and so on.

Step 3. Repeat **Step 1** and **2** until final model $\mathcal{M}^k_{\delta,\lambda}$ with size $d = [\delta^k p] < n$ was obtained, where $[\delta^{k-1} p] \geq n$.

The ITRRS was shown to have SSP provided that λ and δ are chosen appropriately.

Although SIS possess SSP, there are three potential issues might arise with this approach: (1) some unimportant predictors that are highly correlated with important predictors might have high priority to be selected by SIS than other important predictors that are relatively weakly related to the response; (2) an important predictor that is marginally uncorrelated but jointly correlated with the

response can not be picked by SIS; (3) issue of collinearity. To address all the above issues, iterated SIS (ISIS) was proposed as follows:

(1) Select a subset of k_1 variables $\mathcal{A}_1$ using SIS based model selection methods such as SIS-SCAD or SIS-LASSO. Calculate the residuals from regressing response $\mathbf{y}$ over $\mathcal{A}_1$.

(2) Treat those residuals obtained in step 1 as new response and apply the same method as in the previous step to the remaning $(p - k_1)$ variables, which results in a subset of k_2 variables $\mathcal{A}_2$.

(3) Repeat step 1 and 2 until we get l disjoint subsets $\mathcal{A}_1, \cdots, \mathcal{A}_l$ whose union $\mathcal{A} = \cup_{i=1}^{l} \mathcal{A}_i$ has a size $d < n$.

Therefore, the SIS based model selection methods can be extended to ISIS based model selection methods.

2.3.2 Sure Independence Screening for Generalized Linear Models

The SIS provides a novel way for dimension reduction through fast and efficient procedures in linear models. However, SIS can not be applied directly when the distribution of $\mathbf{y}$ is not normal, where a generalized linear model (GLM) is appropriate. Extension of SIS to GLM is of great interest. A likelihood ratio ranking method was proposed by Fan, Samworth, and Wu (2009) and a maximum marginal likelihood estimator (MMLE) screening procedure was illustrated by Fan and Song (2010). Both of these two methods are the extension of SIS into GLM.

Consider the GLM with the canonical link. The conditional density fucntion is given by

$$f(y|\mathbf{x}) = \exp\{y\theta(\mathbf{x}) - b(\theta(\mathbf{x}) + c(y))\},$$

where $b(\cdot)$ and $c(\cdot)$ are known functions and $\theta(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}$. The $p + 1$ dimensional predictor $\mathbf{x}$ contains an intercept in the first column, and all the other columns are standardized without loss of generality. We take the dispersion parameter equal to 1 for simplicity, since we are only interested in the conditional mean function of y, namely, $E(y|\mathbf{x}) = b'(\mathbf{x}^T\boldsymbol{\beta})$.

Fan, Samworth, and Wu (2009) defined the marginal utility of the jth feature by

$$L_j = \min_{\beta_0,\beta_j} \frac{1}{n} \sum_{i=1}^{n} L(\beta_0 + x_{ij}\beta_j, y_i),$$

where $L(\cdot)$ is the negative likelihood function based on dataset. Therefore, we can compute the vector of marginal utilities $\mathbf{L} = (L_1, \ldots, L_p)$ and rank them based on $\mathbf{L}$. The smaller L_j is, the more important jth covariate is. The jth feature would be selected if L_j is among the d smallest components of $\mathbf{L}$. Typically, d can be determined by $[n/\log(n)]$. SSP was established for this method under some mild conditions. Recall the ISIS discussed in linear models, similar technique could be applied in GLM. Note that working residuals are easy to compute for linear models but not obvious for GLM. Instead of computing residuals, Fan, Samworth, and Wu (2009) defined the conditional marginal utilities in this framework for the kth iteration by

$$L_j^{k+1} = \min_{\beta_0,\boldsymbol{\beta}_{\mathcal{M}_k},\beta_j} \frac{1}{n} \sum_{i=1}^{n} L(\beta_0 + x_{ij}\beta_j),$$

where $j \in \mathcal{M}_k^c$. Repeat SIS-SCAD or SIS-LASSO procedure, the final model with size $d < n$ could be obtained.

Fan and Song (2010) proposed the maximum marginal likelihood estimators (MMLE) as another marginal utilities for the purpose of screening. MMLE can be expressed as

$$\widehat{\beta}_j^M = \text{argmin}_{\beta_0,\beta_j} \frac{1}{n} \sum_{i=1}^{n} L(\beta_0 + x_{ij}\beta_j, y_i).$$

Therefore, ranking the vector of marginal utilities $\widehat{\boldsymbol{\beta}}^M = (\widehat{\beta}_1^M, \ldots, \widehat{\beta}_j^M)$ in a decreasing order. The bigger $\widehat{\beta}_j^M$ is, the more important the jth covariate is. The jth feature would be selected if $\widehat{\beta}_j^M > \gamma$ and hence among the d largest components of $\widehat{\boldsymbol{\beta}}^M$. When $\gamma = c_3 n^{-\kappa}$, where $c_3 > 0$ and $0 < \kappa < 1/2$, SSP was established for this method under some mild conditions. MMLE and Likelihood Ratio Ranking Method are shown to be equivalent in Fan and Song (2010).

2.3.3 Nonparametric Independence Screening for Additive Models

In practice, there is often little prior information indicating that the effects of the covariates take a linear form or belong to any other finite-dimensional parametric family. Even if the linear model holds in the joint regression, the marginal regression can be highly nonlinear. Motivated by this concern, Fan et al. (2011) extend independent screening method to a more flexible class of nonparametric models, namely, additive models.

Consider an additive model as

$$y = \sum_{j=1}^{p} m_j(X_j) + \varepsilon, \tag{2.3.23}$$

where X_js are standardized and ε is the random error with conditional mean equal to 0. First, fit p marginal nonparametric regression problems by

$$\min_{f_j \in L_2(P)} E(y - f_j(X_j))^2, \tag{2.3.24}$$

where $L_2(P)$ is the class of square integrable function. The minimizer of (2.3.24) is $f_j = E(y|X_j)$, where the B-spline method with polynomial degree $d_n = o(n^{1/3})$ could be applied. To identify important variables, the marginal utilities can be defined and ranked, which then can yield to a small group of covariates by thresholding. Two marginal utilities were introduced.

The first marginal utilities used is $\|\widehat{f}_{nj}\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}_{nj}(x_{ij})^2$. The resulting selected model could be expressed as $\mathcal{M}_{\nu_n} = \{1 \leq i \leq p, \|\widehat{f}_{nj}\|_n^2 > \nu_n\}$, where ν_n is a predefined threshold value. This screening procedure also can be viewed as ranking by magnitude of the correlation of the marginal nonparametric estimate $\{\widehat{f}_{nj}\}_{i=1}^{n}$ with the response because $\|\widehat{f}_{nj}\|_n^2 = \|\mathbf{y}^T \widehat{f}_{nj}\|_n$. Hence, the nonparametric independent screening (NIS) can be related to SIS by Fan and Lv (2008). Another approach is to rank according to the descent order of the residual sum of squares of the componentwise nonparametric regressions, where the selected model is $\mathcal{N}_{\gamma_n} = \{1 \leq i \leq p, u_j < \gamma_n\}$, with $u_j = \min_{\beta_j} E(y - \widehat{f}_{nj}^2)^2$. These two methods can be shown equivalent. Fan et al. (2011) also investigated the SSP of NIS when

$\nu_n = c_5 d_n n^{-2\kappa}$, where $c_5 > 0$ and $0 < \kappa < d/(2d+1)$ under certain regularity conditions. Adopting the same idea of ISIS, to cope with the inherent issues of marginal-utility-based feature screening, iterative NIS (INIS) could be applied in nonparametric additive models.

2.3.4 Model-Free Independence Screening

Zhu et al. (2011) proposed a feature screening approach called sure independent ranking and screening (SIRS) for ultrahigh dimensional data which only imposes a very general model framework instead of a specific model structure. Therefore, this approach is more suitable for ultrahigh dimensional regression where usually little prior information is known of the true model. Furthermore, this method is more robust to model misspecification. Let $\mathbf{x} = (X_1, \ldots, X_p)^T$ denote the standardized covariate vector. Define the active and inactive index sets $\mathcal{A}$ and $\mathcal{I}$ through conditional cumulative distribution function $F(y|\mathbf{x})$:

$$\mathcal{A} = \{k : F(y|\mathbf{x}) \text{ functionally depends on } X_k \text{ for some y}\},$$

$$\mathcal{I} = \{k : F(y|\mathbf{x}) \text{ does not functionally depends on } X_k \text{ for any y}\}.$$

Define $\mathbf{\Omega}(y) = E\{\mathbf{x}F(y|\mathbf{x})\}$. It can be shown that $\mathbf{\Omega}(y) = Cov\{\mathbf{x}, \mathbf{I}(Y < y)\}$. Let $\Omega_k(y)$ be the kth element of $\mathbf{\Omega}(y)$, and define $\omega_k = E\{\Omega_k^2(Y)\}$ with $k = 1, \ldots, p$. If X_k and Y are independent, $\Omega_k = \omega_k$=0; and if X_k and Y are correlated, $\Omega_k \neq 0$ while $\omega_k > 0$. Therefore, the sample estimate of ω_k was employed as the marginal utility measures for ranking and screening.

$$\widehat{\omega}_k = \frac{n^3}{n(n-1)(n-2)}\tilde{\omega}_k,$$

where $\tilde{\omega}_k = \frac{1}{n}\sum_{j=1}^{n}\left\{\frac{1}{n}\sum_{i=1}^{n} X_{ik}\mathbf{I}(Y_i < Y_j)\right\}^2$ for $k = 1, \ldots, p$. To obtain the significant features, we rank all the candidate predictors according to $\widehat{\omega}_k$ from the largest to smallest and then select the top ones as the active predictors. The thresholding rule for obtaining the cutoff value that separate the active and inactive predictors were proposed by a combination of hard and soft thresholding strategies.

The soft thresholding strategy introduces auxiliary variables. Generate d auxil-

iary variables $\boldsymbol{z} \sim N(\mathbf{0}, \mathbf{I}_d)$ such that $\boldsymbol{z}$ is independent of both $\mathbf{x}$ and $\mathbf{y}$, where $d = p$ empirically. Regard the $(p+d)$ dimensional vector $(\mathbf{x}^T, \boldsymbol{z}^T)^T$, we can calculate $\widehat{\omega}_k$ for $k = 1, \ldots, p+d$. Define C_d=$\max_{l=1,\ldots,d} \widehat{\omega}_{p+l}$, therefore, the kth covariate would be selected if $\widehat{\omega}_k > C_d$. In addition to soft thresholding strategy, hard thresholding strategy can also be applied, which retains a fixed number of predictors with the largest N values of $\widehat{\omega}_k$'s, where N is usually chosen to be $n/\log n$. Soft thresholding strategy is usually applied when there are many active predictors while hard thresholding strategy is better used when the signal space is sparse. Consistency in ranking (CIR) property was shown to hold for SIRS under some mild conditions, which implies that in probability, the proposed marginal utility measures always ranks an active predictor above an inactive one, and so guarantees a clear separation between the active and inactive predictors. Iterated SIRS is motivated by solving the inherent issues of feature screening based on marginal untilty measures.

In addition to SIRS, Li, Zhong, and Zhu (2012) proposed a feature screening procedure via distance correlation (DC), where DC was advocated in Szekely et al. (2007). DC has two remarkable properties, the first one is the relationship between the DC of two univariate normal random variables is a strictly increasing function of the absolute value of the Pearson correlation coefficient of these two variables. The second one is the DC equals to zero if and only if these two random variables are independent. Therefore DC based feature screening procedure is equivalent to marginal Pearson correlation learning for linear regression with normally distributed predictors and random error, and it is more effective than Pearson correlation learning for nonlinear marginal relationship between the predictors and response. Sample componentwise DC between the covariates and the response can be calculated and we select the top ones according to some prespecified threshold values, which was shown by Li et al. (2012). The SSP holds for the DS based screening procedure under milder conditions than that for the SIS (Fan and Lv, 2008) in that we do not require the linear regression function. Moreover, DC based screening procedure can handle grouped predictors and multivariate responses compared with SIRS, which is a very appealing feature.

2.4 High Dimensional 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.

2.4.1 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, survival 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.

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t < T < t + \Delta t | T > t) \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{P(t < T < t + \Delta t)}{P(T \geq t)} = \frac{f(t)}{S(t)} \end{aligned}$$

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.

2.4.2 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

$$h(t|\mathbf{x}) = h_0(t)\exp(\mathbf{x}^T\boldsymbol{\beta}), \tag{2.4.25}$$

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 (2.4.25) 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

$$S(t|\mathbf{x}) = S_0(t)^{\exp(\mathbf{X}^T\boldsymbol{\beta})}, \tag{2.4.26}$$

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 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

$$h_{ij}(t|\mathbf{x}_{ij}) = h_0(t)u_i\exp(\mathbf{x}_{ij}^T\boldsymbol{\beta}), \tag{2.4.27}$$

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 PH model. As an alternative, Bennett (1983) introduced the ***proportional odds model***, which was defined by

$$\frac{S(t|\mathbf{x})}{1-S(t|\mathbf{x})} = \frac{S_0(t|\mathbf{x})}{1-S_0(t|\mathbf{x})}\exp(\mathbf{x}^T\boldsymbol{\beta}), \tag{2.4.28}$$

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

$$\log(T) = \mathbf{x}^T\boldsymbol{\beta} + \sigma\epsilon, \tag{2.4.29}$$

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

$$S(t|\mathbf{x}) = S_0(t\exp(-\mathbf{x}^T\boldsymbol{\beta})), \tag{2.4.30}$$

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

$$h(t|\mathbf{x}) = h_0(t)\exp(-\mathbf{x}^T\boldsymbol{\beta}), \tag{2.4.31}$$

where $h_0(t) = \frac{1}{\sigma}t^{\frac{1}{\sigma}-1}$. Although model (2.4.31) has the same form as Cox 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.

2.4.3 Variable Selection in Cox's Survival Data Analysis

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

$$L = \prod_u f(Z_i|\mathbf{x}_i)\prod_c \bar{F}(Z_i|\mathbf{x}_i) = \prod_u h(Z_i|\mathbf{x}_i)\prod_{i=1}^n \bar{F}(Z_i|\mathbf{x}_i), \tag{2.4.32}$$

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 (2.4.32) becomes

$$L(h_0(t), \boldsymbol{\beta}) = \prod_u h_0(Z_i)\exp(\mathbf{x}_i^T\boldsymbol{\beta}) \prod_{i=1}^{n} \exp\{-H_0(Z_i)\exp(\mathbf{x}_i^T\boldsymbol{\beta})\}, \tag{2.4.33}$$

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 (2.4.33), we get the resulting function that depends only on $\boldsymbol{\beta}$ after dropping the constant terms

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{N} \frac{\exp(\mathbf{x}_{(j)}^T\boldsymbol{\beta})}{\sum_{i\in R_j}\exp(\mathbf{x}_i^T\boldsymbol{\beta})}. \tag{2.4.34}$$

The logarithm of (2.4.34) can be written as

$$\ell(\boldsymbol{\beta}) = \sum_j^N \left[\mathbf{x}_{(j)}^T\boldsymbol{\beta} - \log\Big\{\sum_{i\in R_j}\exp(\mathbf{x}_i^T\boldsymbol{\beta})\Big\}\right], \tag{2.4.35}$$

which is the partial likelihood function (Cox, 1975). Based on (2.4.35), penalized maximum partial likelihood estimator can be used to select significant covariates. The penalized partial likelihood (PPLKD) can be defined by

$$Q(\boldsymbol{\beta}) = \sum_j^N \left[\mathbf{x}_{(j)}^T\boldsymbol{\beta} - \log\Big\{\sum_{i\in R_j}\exp(\mathbf{x}_i^T\boldsymbol{\beta})\Big\}\right] - n\sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{2.4.36}$$

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 while inheriting good properties, such as oracle property. They adopted the LQA algorithm to reduce the nonconvex optimization problem to a local quadratic one. The modified Newton-Raphson algorithm based on a good initial yields an efficient penalized partial likelihood estimator, and a sandwich formula was given for statistical inference. Fan and Li (2002) estimate λ via minimizing an approximate generalized cross validation (GCV) statistic, which following the same notation as that in GLM setting.

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 estimator. 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. They shown that the proposed estimator possess oracle estimator with proper choice of λ and modified Fu's shooting algorithm was used to solve the optimization problem. Similarly, they choose GCV statistic as their tuning parameter selector.

2.4.4 Feature screening in Ultra-High Dimensional Survival Data Analysis

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

$$\mu_k(\beta_k) = \max_{\beta_k} \left(\sum_{i=1}^{n} \delta_i x_{ik} \beta_k - \sum_{i=1}^{n} \delta_i \log\Big\{ \sum_{j \in R(z_i)} \exp(x_{jk}\beta_k) \Big\} \right), \qquad (2.4.37)$$

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 to ensure the sure screening property. That is, the selected model $\mathcal{M}$ includes the true model $\mathcal{M}^\star$ with probability tending to one as n goes to infinity. In the rest of this dissertation, we will refer to this procedure as Cox-SIS. After Cox-SIS, the parameter dimensionality is reduced 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

$$\widehat{\beta}_k = \text{argmax}_{\beta_k} \left(\sum_{i=1}^{n} \delta_i x_{ik} \beta_k - \sum_{i=1}^{n} \delta_i \log\Big\{ \sum_{j \in R(z_i)} \exp(x_{jk}\beta_k) \Big\} \right). \tag{2.4.38}$$

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

They also gave the first theoretical justifications of the sure independence screening procedure for censored data. Under the asymptotic framework where the number of covariates can grow with the sample size, we showed that with probability going to 1, their screening procedure will select all of the important variables with a false positive rate close to the prespecified level.

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's 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 Methodology

Let T and $\mathbf{x}$ be the survival time and its p-dimensional covariate vector, respectively. Throughout this chapter, we consider the following Cox's 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 censoring 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

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

3.1.1 A New Feature Screening Procedure

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

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

where u is a scaling constant to be specified and W is a diagonal matrix. Throughout this chapter, 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 $h(\boldsymbol{\gamma}|\boldsymbol{\beta})$ 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 chapter.

It can be seen that $h(\boldsymbol{\beta}|\boldsymbol{\beta}) = \ell_p(\boldsymbol{\beta})$, and under some conditions, $h(\boldsymbol{\gamma}|\boldsymbol{\beta}) \leq \ell_p(\boldsymbol{\beta})$ for all $\boldsymbol{\gamma}$. This ensures the ascent property. See Theorem 1 below for more details.

Since W is a diagonal matrix, $h(\boldsymbol{\gamma}|\boldsymbol{\beta})$ 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

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

for given $\boldsymbol{\beta}$ and m.

Define $\tilde{\boldsymbol{\gamma}}_j = \beta_j + u^{-1}W^{-1}\partial\ell_p(\boldsymbol{\beta})/\partial\beta_j$ for $j = 1, \cdots, p$, and $\tilde{\boldsymbol{\gamma}} = \boldsymbol{\beta} + u^{-1}W^{-1}(\boldsymbol{\beta})\ell_p'(\boldsymbol{\beta})$ is the maximizer of $h(\boldsymbol{\gamma}|\boldsymbol{\beta})$. Denote $g_j = \omega_j\tilde{\gamma}_j^2$ for $j = 1, \cdots, p$, and sort g_j so that $g_{(1)} \geq g_{(2)} \geq \cdots \geq g_{(p)}$, where $W = \text{diag}\{\omega_1, \cdots, \omega_p\}$. The solution of maximization problem (3.1.5) is the hard-thresholding rule defined below

$$\widehat{\gamma}_j = \tilde{\gamma}_j I\{g_j > g_{(m+1)}\}$$

This enable 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

$$\tilde{\boldsymbol{\gamma}}_j^{(t)} = \beta_j^{(t)} + u^{-1}W^{-1}\partial\ell_p(\boldsymbol{\beta}^{(t)})/\partial\beta_j,$$

and

$$g_j^{(t)} = \omega_j\tilde{\gamma}_j^2.$$

Let $g_{(1)}^{(t)} \geq g_{(2)}^{(t)} \geq \cdots \geq g_{(p)}^{(t)}$, the order statistics of $g_j^{(t)}$s. Set $S_t = \{j : \|g_j^{(t)} \geq g_{(m+1)}^{(t)}\}$, the nonzero index set.

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 section 3.2 hold. Denote*

$$\rho^{(t)} = \sup\{\lambda_{\max}\{W^{-1/2}(\boldsymbol{\beta}^{(t)})\{-\ell_p''(\tilde{\boldsymbol{\beta}})\}W^{-1/2}(\boldsymbol{\beta}^{(t)})\} : \tilde{\boldsymbol{\beta}} = \alpha\boldsymbol{\beta}^{(t+1)}+(1-\alpha)\boldsymbol{\beta}^{(t)}, 0 \le \alpha \le 1\},$$

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

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

where $\boldsymbol{\beta}^{(t+1)}$ is defined in Step 2b 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. $\|\beta\|_0 \le m$), 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.

3.1.2 Sure Screening Property

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

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

$$Pr(s^* \subset \widehat{s}) \longrightarrow 1, \quad \text{as} \quad n \to \infty. \tag{3.1.6}$$

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

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

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

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

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

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

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

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

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

Condition (C1) states a few requirements for establishing the sure screening property of the proposed procedure. The first one is the sparsity of $\boldsymbol{\beta}^*$ which makes the sure screening possible with $\tau(\widehat{s}) = m > q$. Also, it requires that the minimal

component in $\boldsymbol{\beta}^*$ does not degenerate too fast, so that the signal is detectable in the asymptotic sequence. Meanwhile, together with (C3), it confines an appropriate order of m that guarantees the identifiability of s^* over s for $\tau(s) \leq m$. Condition (C2) assumes that p diverges with n at up to an exponential rate; it implies that the number of covariates can be substantially larger than the sample size. We establish the sure screening property of the quasi-likelihood estimation by the following theorem.

Theorem 2. *Suppose that Conditions (C1)—(C3) and Conditions (D1)—(D7) in section 3.2 hold. Let $\widehat{s}$ be the model obtained by the (3.1.3) of size m. We have*

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

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

3.2 Proof of Theoretical Properties of SJS

We need the following notation to present the regularity conditions for the partial likelihood and the Cox model. Most notations are adapted from Andersen and Gill (1982), in which counting processes were introduced for the Cox model and the consistency and asymptotic normality of the partial likelihood estimate were established. Denote $\overline{N}_i(t) = I\{T_i \leq t, T_i \leq C_i\}$ and $R_i(t) = I\{T_i \geq t, C_i \geq t\}$. Assume that there are no two component processes $N_i(t)$ jumping at the same time. For simplicity, we shall work on the finite interval $[0, \tau]$. In Cox's model,

properties of stochastic processes, such as being a local martingale or a predictable process, are relative to a right-continuous nondecreasing family $(\mathcal{F}_t : t \in [0,\tau])$ of sub σ-algebras on a sample space $(\Omega, \mathcal{F}, \mathcal{P})$; $\mathcal{F}_t$ represents everything that happens up to time t. Throughout this section, we define $\Lambda_0(t) = \int_0^t h_0(u)\,du$.

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

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

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

Define

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

and

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

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

Define

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

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

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

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

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

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

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

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

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

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

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

is positive definite.

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

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

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

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

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

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

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

Note that the partial derivative conditions on $\mathbf{a}^{(0)}$,$\mathbf{a}^{(1)}$ and $\mathbf{a}^{(2)}$ are satisfied by $\mathbf{A}^{(0)}$, $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$; and that $\mathbf{A}$ is automatically positive semidefinite. Furthermore the

interval $[0, \tau]$ in the conditions may everywhere be replaced by the set $\{t : h_0(t) > 0\}$.

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

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

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

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

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

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

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

The second inequality is due to the fact that $\|\boldsymbol{\beta}_*^{(t+1)}\|_0 = \|\boldsymbol{\beta}^{(t)}\|_0 = m$, and $\boldsymbol{\beta}_*^{(t+1)} = \arg\max_{\boldsymbol{\gamma}} h(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(t)})$ 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.

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

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

as $n \to \infty$.

For any $s \in S_-^m$, define $s' = s \cup s^* \in S_+^{2m}$. Under (C1) condition, we consider $\boldsymbol{\beta}_{s'}$ close to $\boldsymbol{\beta}^*_{s'}$ such that $\|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}\| = w_1 n^{-\tau_1}$ for some $w_1, \tau_1 > 0$. Clearly, when n is sufficiently large, $\boldsymbol{\beta}_{s'}$ falls into a small neighborhood of $\boldsymbol{\beta}^*_{s'}$, so that Condition (C3) becomes applicable. Thus, it follows Condition (C3) and the Cauchy-Schwarz inequality that

$$\begin{aligned}
\ell_p(\boldsymbol{\beta}_{s'}) - \ell_p(\boldsymbol{\beta}^*_{s'}) &= [\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}]^T \ell_p'(\boldsymbol{\beta}^*_{s'}) + (1/2)[\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}]^T \ell_p''(\tilde{\boldsymbol{\beta}}_{s'})[\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}] \\
&\leq [\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}]^T \ell_p'(\boldsymbol{\beta}^*_{s'}) - (c_1/2) n \|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}\|_2^2 \\
&\leq w_1 n^{-\tau_1} \|\ell_p'(\boldsymbol{\beta}^*_{s'})\|_2 - (c_1/2) w_1^2 n^{1-2\tau_1}, \qquad (3.2.7)
\end{aligned}$$

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

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

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

$$\begin{aligned}
Pr\{\ell_j'(\boldsymbol{\beta}^*_{s'}) \geq (2m)^{-1/2} (c_1 w_1/2) n^{1-\tau_1}\} &\leq Pr\{\ell_j'(\boldsymbol{\beta}^*_{s'}) \geq (2w_2 n^{\tau_2})^{-1/2} (c_1 w_1/2) n^{1-\tau_1}\} \\
&= Pr\left\{\ell_j'(\boldsymbol{\beta}^*_{s'}) \geq c n^{1-\tau_1-0.5\tau_2}\right\} \\
&= Pr\left\{\ell_j'(\boldsymbol{\beta}^*_{s'}) \geq n c n^{-\tau_1-0.5\tau_2}\right\} \qquad (3.2.8)
\end{aligned}$$

where $c = c_1 w_1/(2\sqrt{2w_2})$ denotes some generic positive constant. Recall (3.1.2), by differentiation and rearrangement of terms, it can be shown as in Andersen and

Gill (1982) that the gradient of $\ell_p(\boldsymbol{\beta})$ is

$$\ell_p'(\boldsymbol{\beta}) \equiv \frac{\partial \ell_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\infty} [\mathbf{x}_i - \bar{\mathbf{x}}_n(\boldsymbol{\beta}, t)]\, d\overline{N}_i(t). \tag{3.2.9}$$

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

Let t_j be the time of the jth jump of the process $\sum_{i=1}^n \int_0^{\infty} R_i(t)d\overline{N}_i(t)$, $j = 1, \ldots, N$ and $t_0 = 0$. Then, t_j are stopping times. For $j = 0, 1, \ldots, N$, define

$$Q_j = \sum_{i=1}^{n}\int_0^{t_j} b_i(u)d\overline{N}_i(u) = \sum_{i=1}^{n}\int_0^{t_j} b_i(u)dM_i(u) \tag{3.2.10}$$

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

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

$$Pr(|Q_N| > nC_0x) \le 2\exp\{-n^2C_0^2x^2/(2N)\} \le 2\exp(-nx^2/2) \tag{3.2.11}$$

By (3.2.10), $Q_N = n\ell_p'(\boldsymbol{\beta})$ if and only if $\sum_{i=1}^n \int_0^{\infty} R_i(t)d\overline{N}_i(t) \le N$. Thus, the left-hand side of (3.15) in Lemma 3.3 of Huang *et al* (2013) is no greater than $Pr(|Q_N| > nC_0x) \le 2\exp(-nx^2/2)$.

So, (3.2.8) can be rewritten as follows.

$$Pr\left\{\ell_j'(\boldsymbol{\beta}_{s'}^*) \ge ncn^{-\tau_1-0.5\tau_2}\right\} \le \exp\{-0.5nn^{-2\tau_1-\tau_2}\} = \exp\{-0.5n^{1-2\tau_1-\tau_2}\} \tag{3.2.12}$$

Also, by the same arguments, we have

$$Pr\{\ell_j'(\boldsymbol{\beta}^*_{s'}) \leq -m^{-1/2}(c_1 w_1/2)n^{1-\tau_1}\} \leq \exp\{-0.5n^{1-2\tau_1-\tau_2}\} \tag{3.2.13}$$

The inequalities (3.2.12) and (3.2.13) imply that,

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

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

$$\begin{aligned}
Pr\left\{\max_{s\in s^m_-}\ell_p(\boldsymbol{\beta}_{s'}) \geq \ell_p(\boldsymbol{\beta}^*_{s'})\right\} &\leq \sum_{s\in S^m_-} Pr\{\ell_p(\boldsymbol{\beta}_{s'}) \geq \ell_p(\boldsymbol{\beta}^*_{s'})\} \\
&\leq 4mp^m\exp\{-0.5n^{1-2\tau_1-\tau_2}\} \\
&= 4\exp\{\log m + m\log p - 0.5n^{1-2\tau_1-\tau_2}\} \\
&\leq 4\exp\{\log w_2 + \tau_2\log n + w_2 n^{\tau_2}\tilde{c}n^{\kappa} - 0.5n^{1-2\tau_1-\tau_2}\} \\
&= 4w_2\exp\{\tau_2\log n + w_2\tilde{c}n^{\tau_2+\kappa} - 0.5n^{1-2\tau_1-\tau_2}\} \\
&= a_1\exp\{\tau_2\log n + a_2 n^{\tau_2+\kappa} - 0.5n^{1-2\tau_1-\tau_2}\} \\
&= o(1) \quad \text{as} \quad n\to\infty
\end{aligned} \tag{3.2.14}$$

for some generic positive constants $a_1 = 4w_2$ and $a_2 = w_2\tilde{c}$. By Condition (C3), $\ell_p(\boldsymbol{\beta}_{s'})$ is concave in $\boldsymbol{\beta}_{s'}$, (3.2.14) holds for any $\boldsymbol{\beta}_{s'}$ such that $\|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}\| \geq w_1 n^{-\tau_1}$.

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

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

The theorem is proved.

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

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, 2000 and 5000. For each combination, we conduct 1000 replicates of Monte Carlo simulation. We compare the performance of feature screening procedures using the following two criteria:

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

2. P_a: the proportion that all active predictors are selected for a given model size m in the 1000 replications.

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

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

Table 3.1. Censoring rates

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

Table 3.2 reports $\mathcal{P}_s$ for the active predictors and $\mathcal{P}_a$ when the covariance matrix of $\mathbf{x}$ is the compound symmetric (i.e., S1). Note that under the design of (S1) with (b1), X_4 is jointly dependent but marginally independent of the survival time for all $\rho \neq 0$. This setting is designed to challenge all screening procedures, in particularly the marginal screening procedures. As shown in Table 3.2, Cox-SIS fails to identify X_4 as an active predictor completely when $\boldsymbol{\beta}$ is set to be the 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 $\mathcal{P}_a$ much larger than zero for every case when $\boldsymbol{\beta}$ is set to be the one in (b1). There is no doubt that SJS outperforms Cox-SIS easily in this setting.

We next discuss the performance of the Cox-SIS and the SJS when the covariance matrix of $\mathbf{x}$ is compound symmetric and $\boldsymbol{\beta}$ is set to be the one in (b2). In

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	0.996	0.995	0.995	0	0	1	1	0.999	0.996	0.996
		b2	0.887	0.896	0.876	0.88	0.598	0.999	1	0.999	1	0.998
1000	0.5	b1	0.96	0.971	0.963	0	0	0.993	0.992	0.993	0.998	0.99
		b2	0.765	0.816	0.793	0.81	0.366	0.993	0.998	0.994	0.992	0.981
1000	0.75	b1	0.863	0.84	0.858	0.007	0.003	0.991	0.984	0.98	0.989	0.964
		b2	0.691	0.694	0.707	0.683	0.202	0.96	0.947	0.96	0.942	0.849
2000	0.25	b1	0.984	0.991	0.991	0	0	0.999	0.995	0.997	0.981	0.975
		b2	0.826	0.817	0.826	0.842	0.437	0.993	0.992	0.993	0.997	0.984
2000	0.5	b1	0.951	0.948	0.937	0.001	0.001	0.961	0.962	0.962	0.983	0.937
		b2	0.73	0.707	0.707	0.734	0.236	0.981	0.976	0.977	0.976	0.936
2000	0.75	b1	0.761	0.783	0.775	0.008	0.005	0.954	0.943	0.942	0.987	0.898
		b2	0.611	0.638	0.619	0.62	0.134	0.887	0.891	0.9	0.898	0.717
5000	0.25	b1	0.977	0.975	0.981	0	0	0.983	0.985	0.987	0.906	0.897
		b2	0.739	0.788	0.763	0.769	0.317	0.972	0.974	0.978	0.975	0.938
5000	0.5	b1	0.892	0.9	0.894	0	0	0.88	0.879	0.885	0.934	0.825
		b2	0.636	0.619	0.643	0.629	0.127	0.919	0.922	0.934	0.923	0.812
5000	0.75	b1	0.701	0.696	0.659	0.008	0.002	0.84	0.832	0.85	0.988	0.724
		b2	0.501	0.501	0.488	0.472	0.045	0.78	0.799	0.784	0.783	0.486

this setting, there is no predictor that is marginally independent of, but jointly dependent with the response. Increasing the sample size n from 100 to 200 generates Table 3.3, which has the similar pattern as Table 3.2 while with better performance due to larger sample size. Tables 3.2 and 3.3 clearly show that how the performance of Cox-SIS and SJS is affected by the sample size n, the dimension of predictors p 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 n increases, the feature dimension p 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

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	0.986	0.987	0.988	0.989	0.951	1	1	1	1	1
1000	0.5	b1	1	1	1	0	0	1	1	1	1	1
		b2	0.964	0.961	0.959	0.954	0.844	1	1	1	1	1
1000	0.75	b1	0.996	0.997	0.994	0.001	0.001	1	1	1	1	1
		b2	0.93	0.929	0.92	0.935	0.731	1	1	1	1	1
2000	0.25	b1	1	1	1	0	0	1	1	1	1	1
		b2	0.977	0.971	0.979	0.964	0.897	1	1	1	1	1
2000	0.5	b1	0.999	1	1	0	0	1	1	1	1	1
		b2	0.95	0.946	0.932	0.942	0.786	1	1	1	1	1
2000	0.75	b1	0.989	0.99	0.994	0.001	0.001	1	1	1	1	1
		b2	0.887	0.873	0.883	0.909	0.597	1	0.998	1	1	0.998
5000	0.25	b1	1	1	1	0	0	1	1	1	1	1
		b2	0.952	0.962	0.949	0.958	0.825	1	1	1	1	1
5000	0.5	b1	0.999	0.998	1	0	0	1	1	1	1	1
		b2	0.904	0.903	0.892	0.885	0.637	1	1	1	1	1
5000	0.75	b1	0.978	0.976	0.985	0.004	0.004	1	1	1	0.999	0.999
		b2	0.823	0.832	0.832	0.812	0.431	0.998	0.999	0.997	0.999	0.993

whether there presents marginally independent but jointly dependent predictors or not.

To investigate that how fast our proposed computing algorithm converges, we also report the summarized information of converge steps in Table 3.4 for every combination scenarios with compound symmetric covariate matrix. Table 3.4 implies that the algorithm will converges within 5 iterations at a average level.

Tables 3.5 and 3.6 depict the simulation results for the AR covariance structure (S2) with sample size $n = 100$ and 200 respectively. It is worth to note that with the AR covariance structure and $\boldsymbol{\beta}$ being set to the one in (b1) or (b2), none of the active predictors $X_1, \cdots, X_4$ is marginally independent of the survival time. Thus, it is expected that the Cox-SIS works well for both cases (b1) and (b2). Tables 3.5 and 3.6 indicate 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), although $\mathcal{P}_a$ becomes much higher when the

Table 3.4. The summarized information of converge steps for SJS

				$\Sigma = (1-\rho)I + \rho\mathbf{1}\mathbf{1}^T$			$\Sigma = (\rho^{\lvert i-j\rvert})$		
n	p	ρ	$\boldsymbol{\beta}$	Median	Mean	Sd	Median	Mean	Sd
100	1000	0.25	b1	4	4.142	0.662	4	3.888	0.725
			b2	4	4.058	0.932	3	3.597	0.782
100	1000	0.5	b1	4	4.477	0.767	4	3.806	0.984
			b2	4	4.45	0.935	3	3.616	0.834
100	1000	0.75	b1	5	4.572	0.779	4	3.8	0.926
			b2	5	4.623	1.006	3	3.586	0.855
100	2000	0.25	b1	4	4.144	0.772	4	3.961	0.717
			b2	4	4.248	0.993	3	3.639	0.860
100	2000	0.5	b1	5	4.603	0.976	4	4.051	1.010
			b2	5	4.599	1.036	3	3.637	0.889
100	2000	0.75	b1	5	4.82	0.844	4	4.092	1.018
			b2	5	4.644	1.080	3	3.655	0.891
100	5000	0.25	b1	4	4.092	0.900	4	4.017	0.685
			b2	4	4.353	1.129	3	3.559	0.883
100	5000	0.5	b1	5	4.687	1.256	4	4.135	1.067
			b2	5	4.682	1.139	3	3.556	0.884
100	5000	0.75	b1	5	5.082	0.815	4	4.045	1.194
			b2	4	4.539	1.206	3	3.625	0.956
200	1000	0.25	b1	4	4.028	0.614	4	3.616	0.683
			b2	3	3.33	0.643	3	3.451	0.659
200	1000	0.5	b1	4	4.133	0.544	3	3.213	0.580
			b2	4	3.769	0.780	3	3.413	0.643
200	1000	0.75	b1	4	3.985	0.526	3	2.932	0.665
			b2	4	4.105	0.717	3	3.317	0.619
200	2000	0.25	b1	4	3.954	0.623	4	3.719	0.719
			b2	3	3.347	0.713	3	3.484	0.706
200	2000	0.5	b1	4	4.226	0.599	3	3.334	0.683
			b2	4	3.92	0.873	3	3.475	0.739
200	2000	0.75	b1	4	4.208	0.550	3	3.252	0.686
			b2	4	4.236	0.824	3	3.433	0.690
200	5000	0.25	b1	4	3.838	0.602	4	3.815	0.714
			b2	3	3.404	0.783	3	3.377	0.744
200	5000	0.5	b1	4	4.203	0.667	3	3.439	0.778
			b2	4	3.959	0.867	3	3.426	0.740
200	5000	0.75	b1	5	4.547	0.595	3	3.499	0.749
			b2	4	4.201	0.913	3	3.421	0.721

Table 3.5. 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	1	1	1	0.302	0.302	1	1	1	0.999	0.999
		b2	0.991	1	1	0.992	0.983	1	1	1	1	1
1000	0.5	b1	1	1	0.965	0.527	0.496	1	1	0.985	0.989	0.983
		b2	0.999	1	1	1	0.999	1	1	1	1	1
1000	0.75	b1	1	0.999	0.61	0.453	0.154	0.997	0.98	0.811	0.967	0.799
		b2	1	1	1	1	1	1	1	1	1	1
2000	0.25	b1	1	1	0.997	0.183	0.182	1	1	1	0.989	0.989
		b2	0.989	1	0.999	0.983	0.971	1	1	1	1	1
2000	0.5	b1	1	1	0.941	0.446	0.394	0.998	0.997	0.936	0.97	0.931
		b2	1	1	1	0.999	0.999	1	1	1	1	1
2000	0.75	b1	1	1	0.525	0.364	0.048	0.985	0.927	0.641	0.907	0.615
		b2	1	1	1	1	1	1	1	1	1	1
5000	0.25	b1	1	1	0.991	0.135	0.131	1	1	1	0.965	0.965
		b2	0.981	0.999	1	0.975	0.955	0.999	1	1	0.999	0.999
5000	0.5	b1	1	1	0.888	0.296	0.214	0.992	0.981	0.821	0.896	0.811
		b2	0.999	1	1	0.999	0.998	1	1	1	1	1
5000	0.75	b1	1	1	0.439	0.23	0.019	0.959	0.82	0.449	0.783	0.415
		b2	1	1	1	1	1	1	1	1	1	1

sample size increases from 100 to 200. In summary, SJS outperform Cox-SIS in all cases considered in Tables 3.5 and 3.6, in particular, when $\boldsymbol{\beta}$ is set to be the one in (b1).

We next compare SJS with the iterative Cox-SIS. Tables 3.2 and 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's 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

Table 3.6. 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	0.681	0.681	1	1	1	1	1
		b2	1	1	1	1	1	1	1	1	1	1
1000	0.5	b1	1	1	1	0.913	0.913	1	1	1	1	1
		b2	1	1	1	1	1	1	1	1	1	1
1000	0.75	b1	1	1	0.96	0.839	0.799	1	1	0.999	1	0.999
		b2	1	1	1	1	1	1	1	1	1	1
2000	0.25	b1	1	1	1	0.592	0.592	1	1	1	1	1
		b2	1	1	1	1	1	1	1	1	1	1
2000	0.5	b1	1	1	0.999	0.869	0.868	1	1	1	1	1
		b2	1	1	1	1	1	1	1	1	1	1
2000	0.75	b1	1	1	0.921	0.757	0.678	1	1	0.999	0.999	0.998
		b2	1	1	1	1	1	1	1	1	1	1
5000	0.25	b1	1	1	1	0.45	0.45	1	1	1	1	1
		b2	1	1	1	1	1	1	1	1	1	1
5000	0.5	b1	1	1	1	0.79	0.79	1	1	1	1	1
		b2	1	1	1	1	1	1	1	1	1	1
5000	0.75	b1	1	1	0.88	0.674	0.554	1	1	0.993	0.997	0.991
		b2	1	1	1	1	1	1	1	1	1	1

(each with the size $m/2$) so that the number of the included predictors equals $m = [n/\log(n)] = 22$ for both Cox-SIS and the SJS.

The simulation results are summarized in Table 3.7. 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. Table 3.7 indicates Cox-ISIS and SJS work well and are comparable. When $\rho = 0.5$ and 0.75, SJS can significantly outperform Cox-ISIS in terms of $\mathcal{P}_s$ and $\mathcal{P}_a$ while Cox-ISIS seems to be more suitable for cases with $\rho = 0.25$. When $p = 2000$, SJS takes less time than Cox-ISIS, and they are comparable in computing time when $p = 5000$.

Table 3.7 with $\tau = 1$ and Table 3.2 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 performance of Cox-SIS, Cox-ISIS and SJS, with the setting of $\beta_j = \tau(-1)^U(a + |Z_j|)$ for $j = 1, 2, 3$ and 4,

Table 3.7. Comparison with Cox-ISIS

			Cox-ISIS						SJS					
			$\mathcal{P}_s$				$\mathcal{P}_a$	Time	$\mathcal{P}_s$				$\mathcal{P}_a$	Time
τ	p	ρ	X_1	X_2	X_3	X_4	ALL	(second)	X_1	X_2	X_3	X_4	ALL	(second)
1	1000	0.25	1	1	1	0.999	0.999	7.59	1	0.999	1	0.996	0.996	1.09
		0.5	0.932	0.946	0.944	1	0.833	7.02	0.987	0.99	0.99	0.998	0.985	1.13
		0.75	0.78	0.785	0.775	1	0.451	7.61	0.983	0.977	0.98	0.993	0.958	1.27
1	2000	0.25	0.998	0.998	0.999	1	0.996	23.34	0.999	0.996	0.995	0.979	0.975	5.75
		0.5	0.898	0.894	0.897	1	0.708	21.47	0.97	0.968	0.975	0.983	0.952	6.05
		0.75	0.697	0.696	0.694	1	0.303	19.03	0.952	0.949	0.953	0.993	0.903	5.72
1	5000	0.25	0.998	0.994	0.999	0.992	0.983	36.47	0.988	0.981	0.984	0.925	0.912	26.00
		0.5	0.819	0.833	0.853	1	0.562	37.81	0.871	0.861	0.862	0.948	0.805	31.89
		0.75	0.579	0.583	0.611	1	0.177	38.81	0.829	0.838	0.828	0.988	0.724	36.73
0.75	1000	0.25	1	1	1	1	1	7.81	1	1	0.998	0.99	0.99	1.19
		0.5	0.943	0.93	0.929	1	0.804	7.21	0.995	0.994	0.995	0.997	0.991	1.27
		0.75	0.799	0.797	0.769	1	0.477	7.72	0.984	0.987	0.98	0.988	0.961	1.33
0.75	2000	0.25	1	0.997	1	0.999	0.996	14.19	0.999	0.998	1	0.98	0.978	3.85
		0.5	0.896	0.899	0.904	1	0.712	14.10	0.97	0.969	0.97	0.987	0.952	4.47
		0.75	0.709	0.687	0.724	1	0.334	22.99	0.936	0.938	0.942	0.99	0.882	7.33
0.75	5000	0.25	0.991	0.996	0.99	0.99	0.972	42.64	0.983	0.985	0.988	0.931	0.914	52.50
		0.5	0.84	0.823	0.844	1	0.563	44.85	0.895	0.89	0.896	0.956	0.848	43.96
		0.75	0.566	0.584	0.555	1	0.167	50.80	0.832	0.819	0.836	0.985	0.7	55.27
0.5	1000	0.25	0.998	1	1	1	0.998	7.49	1	0.999	0.999	0.994	0.992	1.20
		0.5	0.934	0.946	0.936	1	0.827	9.61	0.993	0.994	0.994	0.996	0.988	1.70
		0.75	0.749	0.74	0.756	1	0.408	7.91	0.958	0.958	0.965	0.99	0.907	1.47
0.5	2000	0.25	0.997	0.997	0.999	1	0.994	14.45	1	0.997	0.998	0.981	0.978	3.99
		0.5	0.891	0.888	0.899	1	0.702	26.78	0.957	0.962	0.963	0.987	0.943	8.81
		0.75	0.672	0.678	0.665	1	0.273	13.95	0.883	0.889	0.889	0.99	0.772	4.79
0.5	5000	0.25	0.993	0.995	0.99	0.993	0.975	41.41	0.977	0.983	0.989	0.912	0.897	34.82
		0.5	0.806	0.847	0.805	1	0.527	56.10	0.874	0.867	0.855	0.946	0.803	57.31
		0.75	0.56	0.574	0.544	1	0.161	40.54	0.738	0.761	0.746	0.975	0.564	61.49
0.25	1000	0.25	0.989	0.992	0.99	0.989	0.961	8.01	0.993	0.995	0.99	0.936	0.927	1.20
		0.5	0.881	0.876	0.883	1	0.671	8.72	0.948	0.946	0.951	0.988	0.904	1.69
		0.75	0.671	0.626	0.645	1	0.252	7.96	0.726	0.717	0.735	0.965	0.44	1.41
0.25	2000	0.25	0.97	0.972	0.976	0.973	0.902	14.40	0.971	0.971	0.981	0.853	0.824	3.72
		0.5	0.822	0.806	0.819	1	0.534	14.45	0.866	0.845	0.833	0.966	0.748	5.00
		0.75	0.528	0.536	0.526	1	0.126	14.48	0.552	0.566	0.564	0.952	0.238	4.72
0.25	5000	0.25	0.941	0.936	0.934	0.949	0.805	43.85	0.901	0.914	0.897	0.675	0.592	59.46
		0.5	0.731	0.736	0.709	0.999	0.366	45.25	0.664	0.671	0.645	0.86	0.475	50.66
		0.75	0.466	0.432	0.419	1	0.067	49.79	0.427	0.389	0.372	0.958	0.1	118.30

where $a = 4\log n/\sqrt{n}$, $U \sim$ Bernoulli(0.4) and $Z_j \sim \mathcal{N}(0,1)$ 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 $n = 100$ with the covariance matrix set to be S2. The values of $\mathcal{P}_a$ and computing time are reported in Table 3.8 from which, it can be seen that SJS, comparable to Cox-ISIS, outperforms Cox-SIS in terms of

Table 3.8. Comparison among Cox-SIS, Cox-ISIS and SJS

					Cox-SIS		Cox-ISIS		SJS	
Σ	n	τ	p	ρ	$\mathcal{P}_a$	Time (s)	$\mathcal{P}_a$	Time (s)	$\mathcal{P}_a$	Time (s)
S1	200	1	1000	0.25	0.946	4.53	1	13.26	1	2.18
				0.5	0.862	4.09	1	12.08	1	2.20
				0.75	0.734	4.08	1	12.05	0.999	2.40
S1	200	1	2000	0.25	0.931	8.74	1	25.84	1	8.23
				0.5	0.777	9.65	0.999	27.67	1	9.79
				0.75	0.602	7.12	0.992	21.59	1	7.31
S1	200	1	5000	0.25	0.832	20.27	1	60.35	1	92.34
				0.5	0.671	19.87	0.999	59.47	1	96.00
				0.75	0.446	20.33	0.978	60.08	0.99	107.67
S1	200	0.5	1000	0.25	0.933	4.56	1	13.05	1	2.11
				0.5	0.824	4.48	0.996	12.94	0.997	2.26
				0.75	0.643	4.60	0.871	13.32	0.923	2.53
S1	200	0.5	2000	0.25	0.878	8.61	0.999	24.96	1	7.03
				0.5	0.74	8.55	0.985	24.95	0.995	7.67
				0.75	0.497	8.25	0.684	24.45	0.836	8.14
S1	200	0.5	5000	0.25	0.793	24.21	0.996	69.06	0.997	130.18
				0.5	0.594	20.39	0.958	60.04	0.985	105.37
				0.75	0.356	22.66	0.494	65.93	0.719	139.00
S2	100	1	1000	0.25	0.982	3.41	1	8.11	1	0.96
				0.5	1	3.41	1	8.10	1	0.96
				0.75	1	3.05	1	7.29	1	0.87
S2	100	1	2000	0.25	0.977	6.51	1	15.60	1	3.48
				0.5	1	6.48	1	15.50	1	3.42
				0.75	1	6.72	1	16.05	1	3.58
S2	100	1	5000	0.25	0.946	17.12	1	44.88	1	28.11
				0.5	0.998	16.75	1	40.24	1	25.86
				0.75	1	17.31	1	41.57	1	27.67
S2	100	0.5	1000	0.25	0.974	3.20	1	7.57	0.999	1.00
				0.5	1	3.19	1	7.56	1	0.97
				0.75	1	3.18	1	7.54	0.994	0.96
S2	100	0.5	2000	0.25	0.952	8.55	1	19.98	0.997	4.54
				0.5	1	8.60	1	20.17	1	4.54
				0.75	1	7.27	1	17.21	0.993	3.94
S2	100	0.5	5000	0.25	0.908	19.47	1	46.31	0.977	38.38
				0.5	0.999	17.82	1	42.16	0.999	31.91
				0.75	1	18.35	1	43.70	0.994	32.83

$\mathcal{P}_a$ in all scenarios included this table. While Cox-SIS needs less computing time than Cox-ISIS and Cox-SJS.

3.4 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 3.9. The maximum of partial likelihood function of three corresponding models obtained by SJS, Cox-ISIS and Cox-SIS procedures are -536.9838, -561.8795, and -600.0885, respectively. This implies that both SJS and Cox-ISIS performs much better than Cox-SIS with SJS performing the best. There exits overlapped screened features among the three procedures, and we list them in Table 3.10.

Table 3.9. Four-three gene IDs selected by Cox-SJS, Cox-ISIS and Cox-SIS

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

Table 3.10. Overlapped features by 3 different procedures

SJS ∩ ISIS (10)		SJS ∩ SIS (12)		ISIS∩ SIS (22)				SJS∩ISIS∩ SIS (8)	
1662	4317	1662	3820	1188	1825	3822	5614	1662	7343
1682	6956	1664	3821	1456	2579	3824	5950	1682	7357
1825	7343	1682	3824	1662	3799	3825	7343	1825	
3811	7357	1825	3825	1671	3811	4131	7357	3811	
3824		3810	7343	1681	3812	5055		3824	
3825		3811	7357	1682	3813	5301		3825	

The sparsest model (SJS∩ISIS∩SIS), which contains 8 common genes from three different post-screening models, summaries the common significant gene information. We compare model (SJS∩ISIS), (SJS∩SIS) and (ISIS∩SIS) with model (SJS∩ISIS∩SIS) respectively. The likelihood ratio tests results are listed in Ta-

ble 3.11. Likelihood/BIC/AIC criteria of model (SJS∩ISIS) are the best among the four models listed above. Moreover, p-value 1.57e-06 in Table 3.11 implies that SJS and ISIS indeed capture some extra information other than SIS procedure.

Table 3.11. Likelihood Ratio Tests for models based on (SJS∩ISIS), (SJS∩SIS), (ISIS∩SIS) and (SJS∩ISIS∩SIS)

Model	(SJS∩ISIS)	(SJS∩SIS)	(ISIS∩SIS)	(SJS∩ISIS∩SIS)
Likelihood	-613.8057	-624.1106	-618.1842	-627.1692
BIC	1282.207	1313.736	1356.479	1298.015
AIC	1247.611	1272.221	1280.368	1270.338
p-value	1.57e-06	0.19	0.21	-

We then 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 3.12. The likelihood, the degree of freedom (df), the BIC score and the AIC score of the resulting models are listed in Table 3.13, from which SJS-SCAD results in the best fit model in terms of the AIC and BIC. The partial likelihood ratio test for comparing the model selected by SJS-SCAD and SJS without SCAD is 18.41286 with df=13. This leads to the P-value of this partial likelihood ratio test to be 0.1424632. 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 3.14, which indicates that most selected genes have significant impact on the survival time. We further compare Tables 3.9 and 3.10, and find that Gene 4317 was selected by both SJS and Cox-ISIS, but not by Cox-SIS. From Tables 3.12, 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 3.15 presents likelihoods and AIC/BIC scores for models with and without Gene 4317. The P-values of the likelihood ratio tests in-

dicate that Gene 4317 should be included in the models. This clearly indicates that Cox-SIS fails to identify this significant gene, which is consistent to our conclusion base on Table 3.11.

Table 3.12. IDs of selected genes by SCAD and LASSO

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

Table 3.13. Likelihood, df, AIC and BIC of resulting models.

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

Table 3.14. Estimates and Standard Errors (SE) based on SJS-SCAD

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

Table 3.15. Likelihood, AIC and BIC of models with and without Gene 4317.

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

Ater obtaining estimated covaiate coefficients $\widehat{\boldsymbol{\beta}}$, we further compute the risk scores $\mathbf{x}^T\widehat{\boldsymbol{\beta}}$ for the DLBCL data and divide it to a low-risk group and a high-risk group, where the cutoff value is determined by the median of the estimated scores. Figure 3.1(a), (b) and (c) shows the Kaplan-Meier estimate of survival curves for the two risk groups of patients in the whole data based on procedure SJS-SCAD, ISIS-SCAD and SIS-SCAD respectively. The two curves are well separated in all plots, with the log-rank test yielding a p-value equal to 0, indicating a good fitness of all the models. However, we notice easily that SIS posses small area between two KM curves, which implies SJS and ISIS work better than SIS. And the logrank test statistics for Figure 3.1(a), (b), (c) are 146, 136 and 56.9 with degree freedom 1, which also indicates that SJS serves a better tool in the real data example in terms of fitness.

To assess the predictive performance, the DLBCL data are randomly partitioned into two groups: the training set with 188 patients and the testing set with 47 patients. SJS, Cox-SIS, Cox-ISIS are all applied to the training data to obtain the reduced model with $[188/\log(188)] = 36$ genes. Variable selection tool SCAD is used to analyze the post-screened training data. The corresponding obtained estimates $\widehat{\boldsymbol{\beta}}$ from the training set are then used to predict the testing set. We then compute the risk scores $\mathbf{x}_{test}^T\widehat{\boldsymbol{\beta}}$ for the testing data and divide it into two groups: the low-risk group and the high-risk group. We then plot the KM estimate of survival curves for 10 randomly selected testing data, one would expect the two groups of curves separated well if a procedure performs well.

Figure 3.2(c) shows the Kaplan-Meiercurves for 10 randomly selected testing data where the testing datasets were divided into high- and low-risk groups using the median of the estimated scores by procedure SIS-SCAD. Again the difference in survival curves are not as large as those shown in Figure 3.2(a) and (b). This comparison clearly demonstrates that SIS-SCAD may not work well where there exits gene that is jointly significant while marginal insignificant and SJS or ISIS can capture. This also indicates that SJS serves a better tool than SIS in terms of prediction.

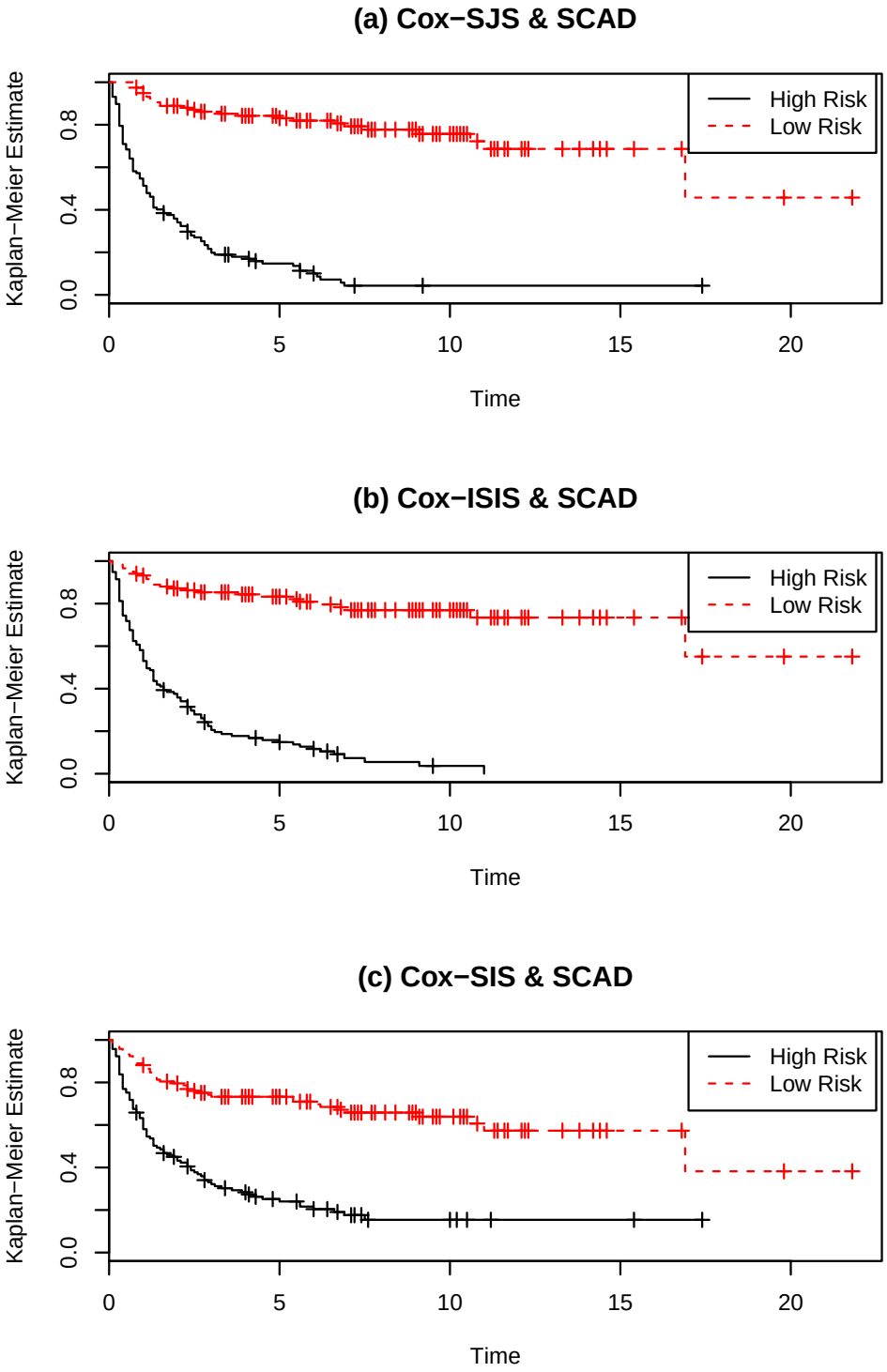

Figure 3.1. Kaplan-Meier curves based on procedures (a) SJS-SCAD (b) ISIS-SCAD and (c) SIS-SCAD for DLBCL data. The whole dataset is divided into two groups of different risks based on the median of the estimated risk scores.

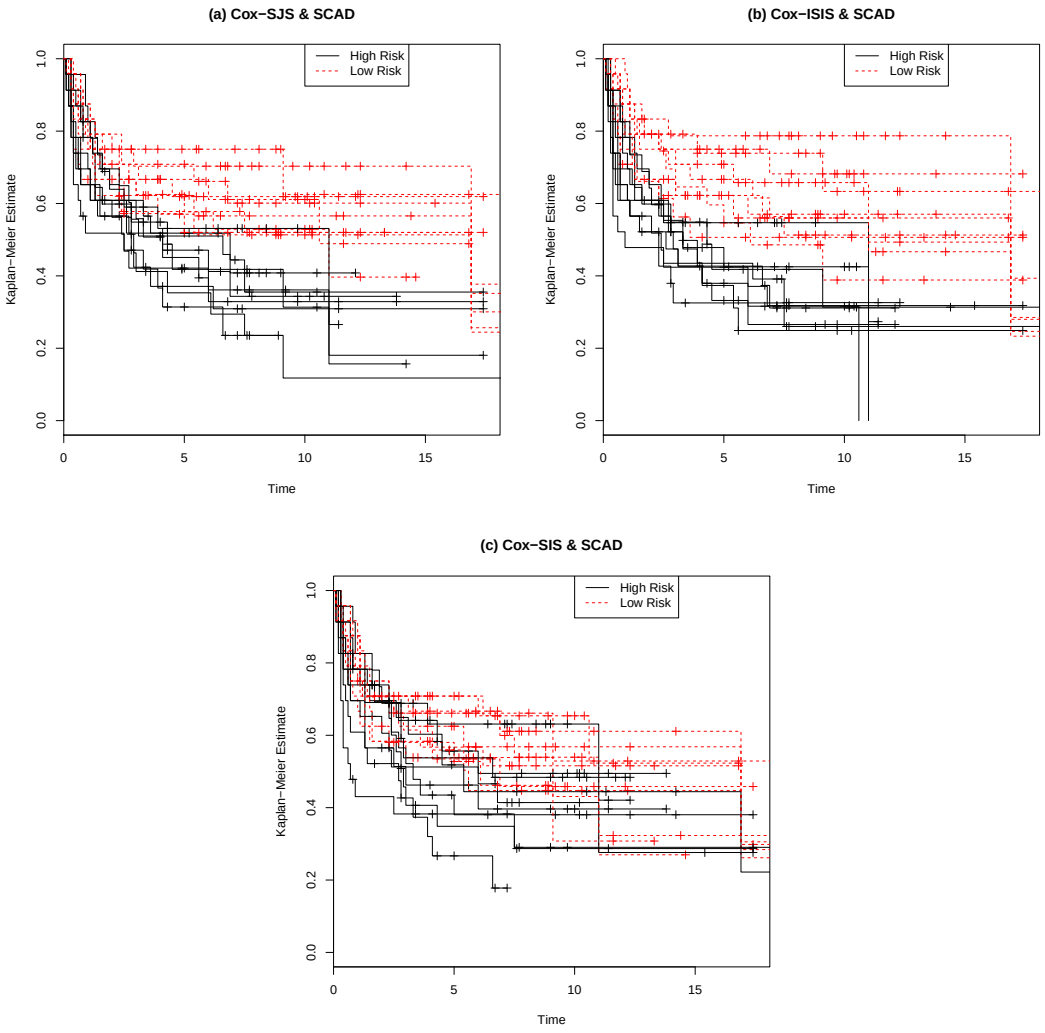

Figure 3.2. Kaplan-Meier curves based on procedures (a) SJS-SCAD (b) ISIS-SCAD and (c) SIS-SCAD for 10 randomly selected testing datasets.

3.5 Conclusions

In this chapter, 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.

Chapter 4

Asymptotic Behavior of Cox's Partial Likelihood

4.1 Introduction

The Cox model (Cox, 1972) has been the most popular model in the survival data analysis during the past decades, and the partial likelihood (Cox, 1975) perhaps is the most commonly-used technique for analysis of right censored data. In practice, many risk factors and covariates are available for the initial analysis, thus an important and challenging task is to identify the significant risk factors and covariates which impact the hazard function depends. Variable selection is a useful technique in the analysis of survival data in the presence of many covariates. Classical variable selection criteria, such as the AIC and BIC, for linear regression models can be naturally extended for the Cox model by replacing the log-likelihood in the AIC (Akaike, 1973) and BIC (Schwarz, 1978) by log-partial likelihood. The LASSO (Tibshirani, 1996) variable selection technique has been extended for the Cox model (Tibshirani, 1996; Zhang and Lu, 2007; Zou, 2008). Folded concave penalized (FCP) partial likelihood variable selection procedures have been developed for the Cox model (Fan and Li, 2002; Bradic, J., Fan, J., and Jiang, J., 2011). To study the theoretical property of these procedures, we need to understand the asymptotic behavior of the partial likelihood.

It seems that little work on asymptotic behavior of the partial likelihood,

though the asymptotic properties of the partial likelihood estimator have extensively studied (Tsiatis, 1981; Andersen and Gill, 1982; Takemi and Toshinari, 1984). Under mild regularity condition, the maximum partial likelihood estimator behaves the same as the ordinary maximum likelihood estimator of independent and identically distributed random samples in terms of asymptotic consistency, asymptotic normality and asymptotic efficiency. See, for example, Murphy and van der Vaart (2000). In this dissertation, we first study the asymptotic behavior of the partial likelihood, and prove that the '*sample average*' of partial likelihood diverges to infinity at a rate of logarithm of the sample size. This clearly indicates that the Cox's partial likelihood does not behave like an ordinary likelihood in that under mild regularity conditions, the sample average of the ordinary likelihood function converges to its expectation (a finite value) in probability as the sample size tends to infinity. This is an interesting result and surprising to us.

With the aid of the asymptotic property of partial likelihood, we study the issue of regularization parameter selection for penalized partial likelihood that was proposed to select significant variables in the Cox model in the literature. Tibshirani (1997) proposed penalized partial likelihood with LASSO penalty for the Cox model. Fan and Li (2002) proposed the partial likelihood with the SCAD penalty for the Cox models, and showed that under certain regularity conditions, the SCAD resulting estimate enjoys the oracle property. Zhang and Lu (2007) and Zou (2008) further proposed adaptive LASSO for the Cox model to improve the SCAD procedure in terms of computational efficiency, while retaining the oracle property. However, the oracle property depends on the choice of the regularization parameter in these proposed procedures. It is well known that the regularization parameter controls the model complexity of the selected models, thus it plays a crucial role in the proposed procedures above mentioned. To our best knowledge, the issue of regularization parameter selection for penalized partial likelihood has not been systematically studied partially because the asymptotic behavior of partial likelihood was not well understood. Wang, Li and Tsai (2007) studied the selection of regularization parameter in the SCAD penalized least squares for linear regression models. They showed that with a positive probability, the *generalized cross-validation* (GCV, Craven and Wahba, 1979) selector yields an over-fitted model, thus the SCAD penalized least squares with the GCV selector does not

enjoy the oracle property. In this dissertation, we shall prove that the GCV selector proposed in Fan and Li (2002) enjoys the model selection consistency for the FCP partial likelihood in the Cox model, which is in contrast to its model selection inconsistency in the least squares setting as demonstrated in Wang, Li and Tsai (2007). This is a surprising result also because the GCV is equivalent to the AIC and the C_p in the context for linear regression models, and it is well known that both AIC and C_p yield an overfitted model with a positive probability, thus are not model selection consistent.

4.2 Asymptotic Behavior of Cox's Partial Likelihood

Let T and $\mathbf{X} = (X_1, \cdots, X_d)^T$ be the survival time and associated d-dimensional vector of covariates, respectively. Consider the following Cox's proportional hazard regression model:

$$h(t \,|\, \mathbf{x}) = h_0(t) \exp(\boldsymbol{x}^T \boldsymbol{\beta}), \tag{4.2.1}$$

where $\boldsymbol{\beta}$ is the regression parameter, and $h(t \,|\, \boldsymbol{x})$ is the conditional hazard function of T given $\boldsymbol{X} = \boldsymbol{x}$ with $h_0(t)$ as an arbitrary baseline hazard function. Suppose that $(\boldsymbol{x}_1, T_1), \ldots, (\boldsymbol{x}_n, T_n)$ is a random sample of $(\boldsymbol{X}, T)$, and the actually observed right censored survival data in practice are as follows: $(\boldsymbol{x}_1, Z_1, \delta_1)$, $\ldots$, $(\boldsymbol{x}_n, Z_n, \delta_n)$, where $Z_i = \min\{T_i, C_i\}, \delta_i = I\{T_i \leq C_i\}$, and C_i is the *right censoring variable* and is independent of T_i given $\boldsymbol{X} = \boldsymbol{x}_i$.

4.2.1 Initial Exploration for Asymptotic Behavior of Partial Likelihood

To study the asymptotic properties in this article, we, without loss of the generality, assume that there are no ties among observed continuous random variables Z_i's. The log-partial likelihood function of the observed data is defined to be

$$\ell_c(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \boldsymbol{x}_i^T \boldsymbol{\beta} - \sum_{i=1}^{n} \delta_i \log\Big(\textstyle\sum_{j=1}^{n} I\{Z_j \geq Z_i\} \exp(\boldsymbol{x}_j^T \boldsymbol{\beta}) \Big). \tag{4.2.2}$$

Table 4.1. Values of U and the corresponding average censoring rates $(1-\rho_1)$ together with $\boldsymbol{\mu}_0 \widehat{=} \mathrm{E}\{(T \le C)X\}$.

U	10.00	5.00	2.75	1.80	1.20	0.80	0.50	0.30
$(1-\rho_1)$	0.1222	0.2055	0.2991	0.3797	0.4613	0.5485	0.6393	0.7311
$\boldsymbol{\mu}_0$	0.0968	0.1429	0.1775	0.1971	0.2007	0.2028	0.1921	0.1652

(Cox, 1975). The goal of this section is to study the asymptotic behavior of $\ell_c(\boldsymbol{\beta})$. Before we pursue further, let us illustrate the different behavior of the log-partial likelihood from the likelihood of independent and identically distributed sample by a hypothetical example.

Example 1. Suppose that we have an independent and identically distributed random sample $\{Y_1, \cdots, Y_n\}$ from a population with probability density/mass function $f(y;\theta)$. Denote by $\ell(\theta) = \sum_{i=1}^n \log\{f(Y_i;\theta)\}$ the log-likelihood function. By the weak law of large number, $n^{-1}\ell(\theta) \to E\log\{f(Y;\theta)\}$ in probability under mild regularity conditions. Furthermore, it has been shown that under mild regularity condition, the maximum partial likelihood estimator (i.e., the maximizer of $\ell_c(\boldsymbol{\beta})$) behaves the same as the ordinary maximum likelihood estimator (i.e., the maximizer of $\ell(\theta)$) in terms of asymptotic consistency, asymptotic normality and asymptotic efficiency. See, for example, Murphy and van der Vaart (2000). However, the partial likelihood itself may behave differently from the ordinary likelihood. In this example, we numerically illustrate the sample average of the partial likelihood function:

$$n^{-1}\ell_c(\boldsymbol{\beta}) \to \infty \quad \text{as} \quad n \to \infty. \tag{4.2.3}$$

We generate a random sample of size n from the following proportional hazard model:

$$h(t|x) = h_0(t)\exp(X\beta),$$

where $h_0(t) \equiv 1$, $\beta = 1$ and $X \sim N(0,1)$. The censoring variable C is generated from an exponential distribution with mean U. Therefore, the average censoring rate varies with different values of U. We list several values of U in Table 4.1 together with their corresponding average censorate rates equals $1 - E\,I(T \le C) \widehat{=} 1-\rho_1$, and take 10 different values of n ranging from $4(=2^2)$ to $1024(=2^{10})$.

Figure 4.1 depicts the scatter plot of $\log(n)$ versus $-n^{-1}\ell_c$ based on a set of typical samples based on the different U listed in Table 4.1. Figure 4.1 clearly shows that $-n^{-1}\ell_c$ increases at $\log(n)$ rate.

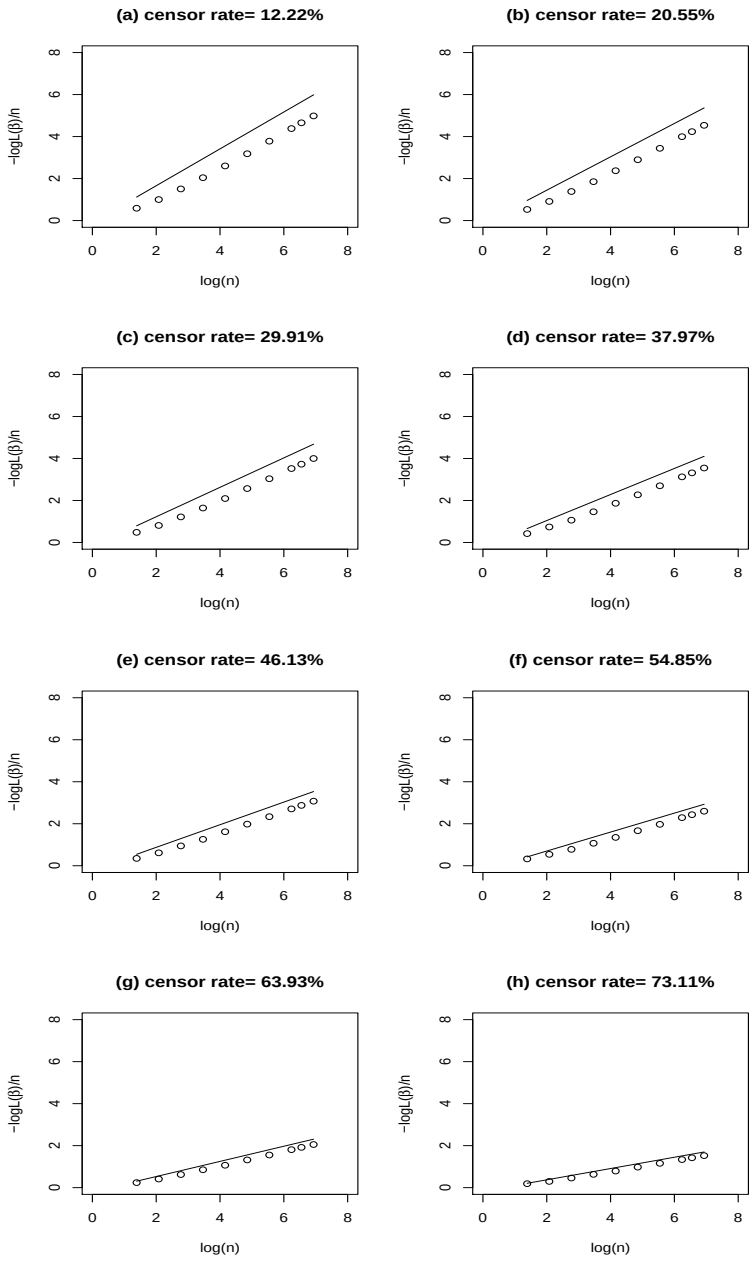

Figure 4.1. Plot of $\log(n)$ versus $-n^{-1}\ell_c$. 'o' is the scatter plot of $\log(n)$ versus $-n^{-1}\ell_c$. The solid line in each plot is $\log(n)\widehat{\rho}_1 - \boldsymbol{\beta}^T\widehat{\boldsymbol{\mu}}_0$ with $\boldsymbol{\beta} = 1$, where $\widehat{\rho}_1$ is an estimate of $EI\{T \leq C\}$ and $\widehat{\boldsymbol{\mu}}_0$ is an estimate of $E\{I\{T \leq C\}X\}$.

We next show that $-n^{-1}\ell_c(\boldsymbol{\beta})$ tends to infinite at the rate of $\log(n)$ by using techniques related to empirical processes. Denote

$$G_n(z,\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} I\{Z_i \le z, \boldsymbol{x}_i \le \mathbf{x}\}, \qquad H_n(z) = n^{-1}\sum_{i=1}^{n} I\{Z_i \le z, \delta_i = 1\}, \tag{4.2.4}$$

and denote $G(z,\boldsymbol{x})$ and $H(z)$ as the limits of the empirical distribution functions $G_n(z,\boldsymbol{x})$ and $H_n(z)$, respectively. Define $\rho_1 = E\{I\{T \le C\}\}$, $\boldsymbol{\mu}_0 = E\{I\{T \le C\}\boldsymbol{X}\}$, and $W(t) = \int\int_{z\ge t} \exp(\boldsymbol{x}^T\boldsymbol{\beta})\, dG(z,\boldsymbol{x})$. We have the following theorem about the asymptotic behavior of the partial likelihood. The proof of this theorem is given in the following section.

Theorem 3. *Suppose that $(\boldsymbol{x}_i, Z_i, \delta_i)$, $i = 1, \cdots, n$, is a random sample from the Cox model (4.2.1) and the censoring is noninformative. That is, C_i is independent of T_i given $\boldsymbol{x}_i$. The following statements are valid.*

(a) If $\boldsymbol{X}$ has a finite bounded support, then

$$-n^{-1}\ell_c(\boldsymbol{\beta}) = \rho_1 log n - \boldsymbol{\mu}_0^T\boldsymbol{\beta} + O_p(1), \quad as \quad n \to \infty. \tag{4.2.5}$$

(b) Assume that $\mu_1 = \int_0^\infty \Big(logW(t)\Big)\, dH(t)$ is well-defined, $E|X_j| < \infty$ for all $j = 1, \cdots, p$, and $0 < E\exp(\boldsymbol{X}^T\boldsymbol{\beta}) < \infty$.

$$-n^{-1}\ell_c(\boldsymbol{\beta}) = \rho_1 log n - \boldsymbol{\mu}_0^T\boldsymbol{\beta} + \mu_1 + o_p(1). \tag{4.2.6}$$

Result in Part (a) of Theorem 3 indicates that $-n^{-1}\ell_c(\boldsymbol{\beta})$ tends to infinity at the rate of $\log(n)$ provided that all covariates are bounded (e.g., all covariates are categorical). Result in Part (b) of Theorem 1 shows that under some mild conditions, $-n^{-1}\ell_c(\boldsymbol{\beta}) - \rho_1\log(n)$ tends to $\mu_1 - \boldsymbol{\mu}_0^T\boldsymbol{\beta}$. When there is no censoring (i.e. $\rho_1 = 1$), it can be shown that $W(t) = f_T(t)/h_0(t)$ and $\mu_1 = \int_0^\infty \log[f_T(t)/h_0(t)]\, dF_T(t)$, where $f_T(t)$ and $F_T(t)$ are the probability density and cumulative distribution function of T in the Cox model (4.2.1), respectively. Thus, assumption about μ_1 holds for many distributions such as the exponential distribution.

4.2.2 Proof of Theorem 3

Without loss of the generality, assume that there are no ties among Z_i's in observed data, and

$$Z_1 < Z_2 < \cdots < Z_n. \tag{4.2.7}$$

This enables us simplifies $n^{-1}\ell_c(\boldsymbol{\beta})$ to the following expression:

$$n^{-1}\ell_c(\boldsymbol{\beta}) = n^{-1}\boldsymbol{\beta}^T \sum_{i=1}^{n} \delta_i \boldsymbol{x}_i - n^{-1} \sum_{i=1}^{n} \delta_i \log\Big(\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) + \cdots + \exp(\boldsymbol{x}_n^T\boldsymbol{\beta})\Big). \tag{4.2.8}$$

It follows by the weak law of large number (WLLN) that $(n^{-1}\sum_{i=1}^{n}\delta_i\boldsymbol{x}_i)^T\boldsymbol{\beta} = \boldsymbol{\mu}_0^T\boldsymbol{\beta} + o_P(1)$. Let

$$R_n = n^{-1} \sum_{i=1}^{n} \delta_i \log\Big(\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) + \cdots + \exp(\boldsymbol{x}_n^T\boldsymbol{\beta})\Big). \tag{4.2.9}$$

Thus,

$$-n^{-1}\ell_c(\boldsymbol{\beta}) = -\boldsymbol{\mu}_0^T\boldsymbol{\beta} + R_n + o_P(1), \tag{4.2.10}$$

Note that from (4.2.7), we have

$$\begin{aligned}
&\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) + \cdots + \exp(\boldsymbol{x}_n^T\boldsymbol{\beta}) = \sum_{j=1}^{n} I\{Z_j \geq Z_i\} e^{\boldsymbol{x}_j^T\boldsymbol{\beta}} \\
&= \int\int I\{z \geq Z_i\} \exp(\mathbf{x}^T\boldsymbol{\beta})\, d\Big\{ \sum_{j=1}^{n} I\{Z_j \leq z, \boldsymbol{x}_j \leq \boldsymbol{x}\}\Big\} \\
&= \int\int_{z \geq Z_i} \exp(\mathbf{x}^T\boldsymbol{\beta})\, d\Big\{ \sum_{j=1}^{n} I\{Z_j \leq z, \boldsymbol{x}_j \leq \boldsymbol{x}\}\Big\} = nW_n(Z_i),
\end{aligned} \tag{4.2.11}$$

where $W_n(t) = \int\int_{z\geq t} \exp(\mathbf{x}^T\boldsymbol{\beta})\, dG_n(z,\boldsymbol{x})$ with $G_n(z,\boldsymbol{x})$ given in (4.2.4). Note that δ_i is a binary random variable, $n^{-1}\sum_{i=1}^{n}\delta_i = \rho_1 + O_P(1/\sqrt{n})$. Denote $A_n = \int_0^\infty \log[W_n(t)]\, dH_n(t)$. It follows that

$$\begin{aligned}
R_n &= n^{-1}\sum_{i=1}^{n} \delta_i \log\Big(nW_n(Z_i)\Big) = n^{-1}\int_0^\infty \log\Big(nW_n(t)\Big)\, d\Big\{\sum_{i=1}^{n}\delta_i I\{Z_i \leq t\}\Big\} \\
&= \int_0^\infty \Big\{\log n + \log\Big(W_n(t)\Big)\Big\}\, dH_n(t) = \log n\Big(H_n(\infty) - H_n(0)\Big) + A_n
\end{aligned}$$

$$= \log n\Big(n^{-1}\sum_{i=1}^{n}\delta_i\Big) + A_n = \rho_1 \log n + \log n\Big(n^{-1}\sum_{i=1}^{n}\delta_i - \rho_1\Big) + A_n$$
$$= \rho_1 \log n + O_P(n^{-1/2}\log n) + A_n = \rho_1 \log n + A_n + o_P(1), \qquad (4.2.12)$$

To prove Part (a), we next deal with A_n. Since

$$(n-i+1)\min_{i\le j\le n}\exp(\boldsymbol{x}_j^T\boldsymbol{\beta}) \le \sum_{j=i}^{n}\exp(\boldsymbol{x}_j^T\boldsymbol{\beta}) \le (n-i+1)\max_{i\le j\le n}\exp(\boldsymbol{x}_j^T\boldsymbol{\beta})$$

and $\boldsymbol{X}$ has a finite bounded support, it follows

$$A_n = n^{-1}\sum_{i=1}^{n}\delta_i \log\Big(W_n(Z_i)\Big) = n^{-1}\sum_{i=1}^{n}\delta_i \log\Big(\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) + \cdots + \exp(\boldsymbol{x}_n^T\boldsymbol{\beta})}{n}\Big)$$
$$= n^{-1}O_P\Big(\sum_{i=1}^{n}\log\big(\frac{n-i+1}{n}\big)\Big) = n^{-1}O_P(\log(n!/n^n)) = O_P(1). \qquad (4.2.13)$$

The last equality is due to the Sterling formula. That is, $n!$ has the same order as $\sqrt{2\pi}(n/e)^n$. This completes the proof of (a).

We next prove Part (b). It suffices to show that

$$A_n \xrightarrow{P} \mu_1, \qquad \text{as } n \to \infty. \qquad (4.2.14)$$

From (4.2.4), we know that $H_n(z)$ is the empirical process of a random sample of Z_i's with $\delta_i = 1$. Thus, $\|H_n - H\| = \sup_z |H_n(z) - H(z)| = O_p(n^{-1/2})$ by DWK inequality (van der Vaart, 1998, P. 268) since $EI\{\delta_i = 1\} = \rho_1 > 0$. Hence, from (4.2.4), (4.2.11), and integration by parts, we have the following:

$$A_n = \int_0^{Z_n}\log[W_n(t)]\,dH_n(t) = B_n + \int_0^{Z_n}\log[W_n(t)]\,d[H_n(t) - H(t)]$$
$$= B_n + [H_n(t) - H(t)]\log[W_n(t)]\Big|_0^{Z_n} - \int_0^{Z_n}[H_n(t) - H(t)]d\{\log[W_n(t)]\}$$
$$= B_n + [H_n(Z_n) - H(Z_n)]\log[W_n(Z_n)] - \int_0^{Z_n}[H_n(t) - H(t)]d\{\log[W_n(t)]\}$$
$$= B_n + [H_n(Z_n) - H(Z_n)]\log\{\exp(\boldsymbol{x}_n^T\boldsymbol{\beta})/n\} - \int_0^{Z_n}[H_n(t) - H(t)]d\{\log[W_n(t)]\}$$

$$= B_n + O_p\Big(\frac{\log n}{\sqrt{n}}\Big) - \int_0^{Z_n} [H_n(t) - H(t)] d\{\log[W_n(t)]\}, \tag{4.2.15}$$

where $B_n = \int_0^{Z_n} \log[W_n(t)] dH(t)$ by using the fact $\mathbf{x}_n^T \boldsymbol{\beta} = O_P(1)$ since $E(|X_j|) < \infty$ by the assumption on $E|X_j| < \infty$ for all $j = 1, \cdots, p$. Note that from (4.2.4) and (4.2.11), we have

$$\begin{aligned}\Big| \int_0^{Z_n} [H_n(t) - H(t)] d\{\log[W_n(t)]\} \Big| &\le \|H_n - H\| \times |\log W_n(Z_n) - \log W_n(0)| \\ &= O_p(n^{-1/2}) \log\Big(\frac{\exp(\boldsymbol{x}_1^T\boldsymbol{\beta}) + \cdots + \exp(\boldsymbol{x}_n^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_n^T\boldsymbol{\beta})}\Big)\end{aligned}$$

By the assumption in Part (b) and the WLLN, it follows that $\frac{1}{n}\sum_{i=1}^n \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) \xrightarrow{P} E\exp\{\boldsymbol{X}^T\boldsymbol{\beta}\}$. This implies that $\log\{\sum_{i=1}^n \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\} - \log(n) = O_P(1)$. Furthermore, $\log\{\exp(\boldsymbol{x}_n^T\boldsymbol{\beta})\} = \boldsymbol{x}_n^T\boldsymbol{\beta} = O_P(1)$. Thus,

$$\log\Big(\frac{\exp(\boldsymbol{x}_1^T\boldsymbol{\beta}) + \cdots + \exp(\boldsymbol{x}_n^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_n^T\boldsymbol{\beta})}\Big) = O_P\{\log(n)\}.$$

As a result, it follows that

$$\Big| \int_0^{Z_n} [H_n(t) - H(t)] d\{\log[W_n(t)]\} \Big| = O_p\Big(\frac{\log n}{\sqrt{n}}\Big). \tag{4.2.16}$$

Therefore, (4.2.14) follows from (4.2.15)-(4.2.16) and the assumption about μ_1 and the Dominate Convergence Theorem, we have

$$B_n \xrightarrow{P} \mu_1, \qquad \text{as } n \to \infty. \tag{4.2.17}$$

4.3 Tuning parameter selector in penalized partial likelihood

Fan and Li (2002) developed a class of variable selection procedures by using the following penalized partial likelihood

$$\sum_{j=1}^{N}[\mathbf{x}_{(j)}^T\boldsymbol{\beta} - \log\{\sum_{i\in R_j}\exp(\mathbf{x}_i^T\boldsymbol{\beta})\}] - n\sum_{j=1}^{d}p_\lambda(|\beta_j|), \tag{4.3.18}$$

where d is the dimension of $\boldsymbol{\beta}$, $p_\lambda(\cdot)$ is a penalty function with a tuning parameter λ (more generally, it is allowed to use λ_j). The penalized partial likelihood estimate of $\boldsymbol{\beta}$ is to maximize (4.3.18) with respect to $\boldsymbol{\beta}$. With a proper choice of p_λ, many of estimated coefficients will be zero and hence their corresponding variables do not appear in the model. This achieves the objectives of variable selection.

Denote by $\boldsymbol{\beta}_0$ the true value of $\boldsymbol{\beta}$, and let $\boldsymbol{\beta}_0 = (\beta_{10}, \cdots, \beta_{d0})^T = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$. Without loss of generality, it is assumed that $\boldsymbol{\beta}_{20} = \mathbf{0}$, and all components of $\boldsymbol{\beta}_{10}$ are not equal to 0. Under some regularity conditions, Fan and Li (2002) showed their SCAD estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T)^T$ possesses the following oracle property. With probability tending to 1, for certain choice of $p_{\lambda_n}(\cdot)$, we have $\widehat{\boldsymbol{\beta}}_2 = 0$ and

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \to N\{\mathbf{0}, I_1^{-1}(\boldsymbol{\beta}_{10}, \mathbf{0})\},$$

where $I_1(\boldsymbol{\beta}_{10}, \mathbf{0})$ is the Fisher information matrix for $\boldsymbol{\beta}_1$ knowing $\boldsymbol{\beta}_2 = \mathbf{0}$.

However, the oracle property depends on the choice of the tuning parameter. Thus, the selection of tuning parameter is fundamental in the penalized likelihood procedure. Wang, Li and Tsai (2007) studied the selection of tuning parameter for penalized least squares for linear regression models and partially linear models. They showed that the GCV tuning parameter used in Fan and Li (2001) cannot yield an oracle estimator. The issue of tuning parameter selection for the penalized partial likelihood has not been studied yet. Based on the asymptotic results of the partial likelihood in the previous section, we will show that the GCV tuning parameter selector proposed in Fan and Li (2002) indeed possesses the model selection consistency, in contrast to the model selection inconsistency of the GCV tuning parameter selector in the penalized least squares setting.

Fan and Li (2002) proposed using local quadratic approximation (LQA) to the penalty function. With the aid of LQA, Newton-Raphson algorithm can be used to maximize the penalized partial likelihood function. The Newton-Raphson algorithm can further enable one to define effective number of parameters for the penalized partial likelihood. Specifically,

$$\nabla \ell_c(\boldsymbol{\beta}) = \frac{\partial \ell_c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}, \quad \nabla^2 \ell_c(\boldsymbol{\beta}) = \frac{\partial^2 \ell_c(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T},$$

and define the effective number of parameters for a given tuning parameter λ to be

$$e(\lambda) = \mathrm{tr}[\{\nabla^2 \ell_c(\widehat{\boldsymbol{\beta}}_\lambda) - n\Sigma_\lambda(\widehat{\boldsymbol{\beta}}_\lambda)\}^{-1} \nabla^2 \ell_c(\widehat{\boldsymbol{\beta}}_\lambda)],$$

where $\Sigma_\lambda(\boldsymbol{\beta}) = \mathrm{diag}\{p'_\lambda(|\beta_1|)/|\beta_1|, \ldots, p'_\lambda(|\beta_d|)/|\beta_d|\}$. Zhang, Li and Tsai (2010) shows that with probability tending to one, e(λ) equals the number of nonzero estimated coefficients. It seems that it is natural to set the number of nonzero estimated coefficients to be degrees of freedom of the selected model when we apply local linear approximation (Zou and Li, 2008) to the penalty function. Thus we set degree of freedom (df_λ) to be the number of the nonzero penalized partial likelihood estimate corresponding to the tuning parameter λ. Similar to Fan and Li (2002), we may define the GCV statistic to be

$$\mathrm{GCV}(\lambda) = \frac{-\ell_c(\widehat{\boldsymbol{\beta}}_\lambda)}{n\{1 - \mathrm{df}_\lambda/n\}^2}, \tag{4.3.19}$$

and $\widehat{\lambda}_{\mathrm{GCV}} = \mathrm{argmin}_\lambda\{\mathrm{GCV}(\lambda)\}$ is selected.

Similar to the classical AIC and BIC, we may define the corresponding AIC and BIC statistics as follows.

$$\mathrm{AIC}(\lambda) = -2\ell_c(\widehat{\boldsymbol{\beta}}_\lambda) + 2\mathrm{df}_\lambda \tag{4.3.20}$$

$$\mathrm{BIC}(\lambda) = -2\ell_c(\widehat{\boldsymbol{\beta}}_\lambda) + \log(n)\mathrm{df}_\lambda. \tag{4.3.21}$$

Then the AIC and BIC tuning parameter selectors are

$$\widehat{\lambda}_{\mathrm{AIC}} = \mathrm{argmin}_\lambda\{\mathrm{AIC}(\lambda)\} \text{ and } \widehat{\lambda}_{\mathrm{BIC}} = \mathrm{argmin}_\lambda\{\mathrm{BIC}(\lambda)\}$$

Note that when t lies in the neighborhood of 0, $(1-t)^{-2} \approx 1-2t$ by Taylor expansion. Thus, when n is large enough,

$$2n\text{GCV}(\lambda) \approx -2\ell_c(\widehat{\boldsymbol{\beta}}_\lambda) + 4(-\ell_c(\widehat{\boldsymbol{\beta}}_\lambda)/n)\text{df}_\lambda.$$

If $-\ell_c(\widehat{\boldsymbol{\beta}}_\lambda)/(n\log(n)) \to EI\{T \leq C\} = \rho_1$ as $n \to \infty$, then

$$2n\text{GCV}(\lambda) \approx -2\ell_c(\widehat{\boldsymbol{\beta}}_\lambda) + 4\rho_1\log(n)\text{df}_\lambda. \tag{4.3.22}$$

Thus, if $\rho_1 \geq 1/4$, the GCV tuning parameter may yield a sparser model than the one selected by the BIC-tuning parameter selector. This is consistent with our observation from the simulation study in Section 4.

4.3.1 Definition and Notation

To study the tuning parameter selectors' asymptotic behavior for variable selection in the scenario of partial likelihood, we first need to define the candidate models considered in model selection.

Let $\bar{\alpha} = \{1, \cdots, d\}$ denote the label of predictors for the full model. Hence, α, the subset of $\bar{\alpha}$, would represent a candidate model including the predictors labeled by α. For each candidate model α, its model size and the corresponding coefficients are defined as df_α and $\boldsymbol{\beta}_\alpha$. Therefore, each tuning parameter λ determined in the penalty function would result in a selected model α_λ with model size df_{α_λ} and the corresponding coefficients $\widehat{\boldsymbol{\beta}}_\lambda$. Moreover, we define the collection of all candidate models as $\mathcal{A}$.

For any given model α, we are able to obtain its non-penalized estimates $\widehat{\boldsymbol{\beta}}^\star_\alpha$ by maximizing the corresponding partial likelihood $\ell_c(\boldsymbol{\beta})$. Similarly, for any selected model α_λ obtained from penalized variable selection with given λ, we are able to obtain the corresponding non-penalized estimates $\widehat{\boldsymbol{\beta}}^\star_{\alpha_\lambda}$.

To study the tuning parameter selectors' asymptotic behaviors with partial likelihood, we define a general tuning parameter selector

$$\text{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}) = \{-2\ell_c(\widehat{\boldsymbol{\beta}}) + \kappa_n\text{df}_{\widehat{\boldsymbol{\beta}}}\}, \tag{4.3.23}$$

where $\widehat{\boldsymbol{\beta}}$ is the parameter estimator and $\mathrm{df}_{\widehat{\boldsymbol{\beta}}}$ is the corresponding degree freedom associated with $\widehat{\boldsymbol{\beta}}$. κ_n is a positive number that denotes different variable selection criterion. When $\kappa_n = 2$, GIC_{κ_n} becomes AIC, while when $\kappa_n = \log(n)$, GIC_{κ_n} becomes BIC.

4.3.2 Theoretical Property

In this subsection, we assume that the set of candidate models contain the unique true model and that the number of parameters in the full model is finite. Assume that the coefficients of the unique true model α_0 in $\mathcal{A}$ are nonzero. Therefore, any candidate model $\alpha \not\supseteq \alpha_0$ is an underfitted model while any model $\alpha \supset \alpha_0$ is an overfitted model. Hence, we can partition the tuning parameter into the three different sets as

$$\begin{aligned}\Omega_- &= \{\lambda : \alpha_\lambda \not\supseteq \alpha_0\}\\ \Omega_0 &= \{\lambda : \alpha_\lambda = \alpha_0\}\\ \Omega_+ &= \{\lambda : \alpha_\lambda \supset \alpha_0\}.\end{aligned}$$

We need the following notation to present the regularity conditions for the partial likelihood and the Cox model. Most notations are adapted from Andersen and Gill (1982), in which counting processes were introduced for the Cox model and the consistency and asymptotic normality of the partial likelihood estimate were established. Following the definition in Chapter 3, denote $\overline{N}_i(t) = I\{T_i \le t, T_i \le C_i\}$ and $R_i(t) = I\{T_i \ge t, C_i \ge t\}$. Assume that there are no two component processes $N_i(t)$ jumping at the same time. For simplicity, we shall work on the finite interval $[0, \tau]$. We need the following conditions to establish the oracle property. Define

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

and

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

where $\mathbf{x}_i^{\otimes 0} = 1$, $\mathbf{x}_i^{\otimes 1} = \mathbf{x}_i$ and $\mathbf{x}_i^{\otimes 2} = \mathbf{x}_i\mathbf{x}_i^T$. Note that $\mathbf{A}^{(0)}(\boldsymbol{\beta}, t)$ is a scalar,

$\mathbf{A}^{(1)}(\boldsymbol{\beta}, t)$ and $E(\boldsymbol{\beta}, t)$ are p-vector, and $\mathbf{A}^{(2)}(\boldsymbol{\beta}, t)$ and $V(\boldsymbol{\beta}, t)$ are $p \times p$ matrices.

(A1) $\int_0^1 h_0(t)dt < \infty$.

(A2) The process $\mathbf{x}_t$ and R(t) are left-continuous with right hand limits, and

$$P\{R(t) = 1\} > 0 \text{ for any } t \in [0, \tau].$$

(A3) (Asymptotic stability). There exists a neighborhood $\mathcal{B}$ of $\boldsymbol{\beta}^*$

$$\sup_{t \in [0,\tau], \boldsymbol{\beta} \in \mathcal{B}} R(t) \exp\{\mathbf{x}^T \boldsymbol{\beta}\} \mathbf{x}^{\otimes 2} < \infty.$$

(A4) (Asymptotic regularity conditions). Define $\mathbf{a}^{(k)}(\boldsymbol{\beta}, t) = E[R(t) \exp\{\mathbf{x}^T \boldsymbol{\beta}\} \mathbf{x}^{\otimes k}]$ for $k = 0, 1, 2.$ and define $e = \mathbf{a}^{(1)}/\mathbf{a}^{(0)}$ and $v = \mathbf{a}^{(2)}/\mathbf{a}^{(0)} - e^{\otimes 2}$, where $\mathbf{a}^{(0)}(\cdot, t)$, $\mathbf{a}^{(1)}(\cdot, t)$ and $\mathbf{a}^{(2)}(\cdot, t)$ are continuous functions of $\boldsymbol{\beta} \in \mathcal{B}$, uniformly in $t \in [0, \tau]$, $\mathbf{a}^{(0)}$, $\mathbf{a}^{(1)}$ and $\mathbf{a}^{(2)}$ are bounded on $\mathcal{B} \times [0, \tau]$; $\mathbf{a}^{(0)}$ is bounded away from zero on $\mathcal{B} \times [0, \tau]$, and the matrix

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

is positive definite.

Conditions (A1)-(A4) guarantee the local asymptotic quadratic (LAQ) property for the partial likelihood function, and hence the asymptotic normality of the maximum partial likelihood estimates. See Andersen and Gill (1982) and Murphy and van der Vaart (2000) for details.

Under the above assumptions, we are able to study the asymptotic property of GIC_{κ_n} by introducing the following conditions.

(E1) Assume that $\lambda_{\max}$ depends on n and satisfies $\lambda_{\max} \to 0$ as $n \to \infty$.

(E2) There exits a constant m such that the penalty $p_\lambda(\xi)$ satisfies $p_\lambda'(\xi) = 0$ for $\xi > m\lambda$.

(E3) If $\lambda_n \to 0$ as $n \to \infty$, then the penalty function satisfies

$$\liminf_{n \to \infty} \liminf_{\xi \to 0} \sqrt{n} p_\lambda'(\xi) \to \infty.$$

(E4) For any candidate model $\alpha \in \mathcal{A}$, there exits $c_\alpha > 0$ such that $-\frac{1}{n}\ell_c(\widehat{\boldsymbol{\beta}}^\star_\alpha) - \log(n)\rho_1 \to c_\alpha$. In addition, for any underfitted model $\alpha \not\supseteq \alpha_0$, $c_\alpha > c_{\alpha_0}$.

Conditions (E1)-(E3) are the penalty conditions while condition (E4) is the technical condition to investigate the asymptotic properties of the tuning parameter selector in the scenario of partial likelihood. Condition (E1) indicates a smaller tuning parameter is required if sample size is large. Condition (E2) indicates that the penalty chosen could result in an asymptotic unbiased penalized estimator. Condition (E3) is used to study the oracle property of the penalized estimator. Condition (E4) assures that the underfitted model yields a larger model deviance than that of the true model.

Following the notation aforementioned, we show the asymptotic performances of GIC_{κ_n} based on the assumptions and conditions described before.

Theorem 4. *Suppose the density function of the Cox model satisfies four regularity condition (A1)-(A4) and that the conditions (E4) holds.*

(A) If there exits a positive constant M such that $\kappa_n < M$, then the tuning parameter $\widehat{\lambda}$ obtained by minimizing $GIC_{\kappa_n}(\lambda)$ satisfies

$$P\{\widehat{\lambda} \in \Omega_-\} \to 0 \quad \textit{and} \quad P\{\widehat{\lambda} \in \Omega_+\} > 0.$$

(B) Under conditions (E1)-(E3), if $\kappa_n \to \infty$ and $\kappa_n/\sqrt{n} \to 0$, then the tuning parameter $\widehat{\lambda}$ obtained by minimizing $GIC_{\kappa_n}(\lambda)$ satisfies

$$P\{\alpha_{\widehat{\lambda}} = \alpha_0\} \to 1.$$

(C) Under conditions (E1)-(E3), if $\rho_1 > 0$, then the tuning parameter $\widehat{\lambda}$ obtained by minimizing the GCV score defined in (4.3.19) satisfies

$$P\{\alpha_{\widehat{\lambda}} = \alpha_0\} \to 1.$$

The proof of theorem 4 is given in the following section.

Theorem 4 provides guidance on the choice of the regularization parameter. Because $\kappa_n = 2$ satisfies the boundedness condition in Theorem 4(A), we name

GIC with the bounded κ_n the AIC-type selector. In contrast, because $\kappa_n = \log n$ fulfills the conditions of Theorem 4(B), we call GIC with $\kappa_n \to \infty$ and $\kappa_n/\sqrt{n} \to 0$ the BIC-type selector. Accordingly, Theorem 4(A) implies that the AIC-type selector tends to overfit without considering which penalty function being used. However, Theorem 4(B) indicates that BIC-type selector enables us to identify the true model consistently. Thus, the penalized partial likelihood with the BIC-type selector possesses the oracle property. Similarly, $\kappa_n = 4\rho_1 \log(n) = \rho \log(n)$ fulfills the conditions of Theorem 4(B). (4.3.22) implies that GCV belongs to the BIC-type selector asymptotically. Therefore, the penalized partial likelihood estimator with the GCV selector also possesses the oracle property.

4.3.3 Proof of Theorem 4

To prove Theorem 4, we show the following two lemmas. And Theorem 4(A) and 4(B) follow Lemma 1 and 2 respectively.

Lemma 1. *Suppose the density function of the Cox model satisfies four regularity condition (A1)-(A4). Assume that there exits a positive constant M such that $\kappa_n < M$. Then under Condition (E4),we have*

$$P\{\inf_{\lambda \in \Omega_-} GIC_{\kappa_n}(\widehat{\boldsymbol{\beta}}_\lambda) > GIC_{\kappa_n}(\widehat{\boldsymbol{\beta}}^\star_{\tilde{\alpha}})\} \to 1 \quad as\ n \to \infty, \tag{4.3.24}$$

$$\liminf_{n\to\infty} P\{\inf_{\lambda \in \Omega_0} GIC_{\kappa_n}(\widehat{\boldsymbol{\beta}}_\lambda) > GIC_{\kappa_n}(\widehat{\boldsymbol{\beta}}^\star_{\tilde{\alpha}})\} \geq \pi. \tag{4.3.25}$$

Proof. Recall that for any given λ, we can obtain a selected model α_λ by penalized variable selection. And based on this selected model α_λ, we are able to obtain its corresponding non-penalized estimates $\widehat{\boldsymbol{\beta}}^\star_{\alpha_\lambda}$ by maximizing the corresponding partial likelihood. Therefore, we have

$$\begin{aligned} \ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha_\lambda}) &\geq \ell_c(\widehat{\boldsymbol{\beta}}_\lambda) \\ -2\ell_c(\widehat{\boldsymbol{\beta}}_\lambda) + \kappa_n \mathrm{df}_\lambda &> -2\ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha_\lambda}) \end{aligned} \tag{4.3.26}$$

Namely,

$$\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_\lambda) > -2\ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha_\lambda}). \tag{4.3.27}$$

Subtract $\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}})$ from both size of (4.3.27), we can obtain that

$$\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda}) - \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}}) > -2\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_\lambda}) - [-2\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}})] - \kappa_n \mathrm{df}_{\bar{\alpha}}$$

For any $\lambda \in \Omega_- = \{\lambda : \alpha \not\supseteq \alpha_0\}$, we can take $\inf_{\lambda\in\Omega_-}$ over $\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda})$. Under Condition (E4) and $\kappa_n < M$, for any $\lambda \in \Omega_-$, we have

$$\begin{aligned}
&P\{\inf_{\lambda\in\Omega_-} \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda}) - \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}}) > 0\} \\
&\geq P\{\inf_{\lambda\in\Omega_-} \frac{-2\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_\lambda})}{n} - \frac{-2\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}})}{n} - \frac{\kappa_n \mathrm{df}_{\bar{\alpha}}}{n} > 0\} \\
&= P\{\min_{\alpha\not\supseteq\alpha_0}[\frac{-2\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha})}{n} - \log(n)\rho_1] - [\frac{-2\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}})}{n} - \log(n)\rho_1] - \frac{\kappa_n \mathrm{df}_{\bar{\alpha}}}{n} > 0\} \quad (4.3.28) \\
&= P\{\min_{\alpha\not\supseteq\alpha_0} c_\alpha - c_{\bar{\alpha}} + o_{\mathbb{P}}(1) > 0\} \to 1, \quad (4.3.29)
\end{aligned}$$

as $n \to \infty$. (4.3.28) is due to the finiteness of $\mathcal{A}$, and (4.3.29) uses both (E4) and the fact that deviance tends to be smaller as covariate dimension increases. (4.3.24) follows from the above equations.
For any $\lambda \in \Omega_0, \alpha_\lambda = \alpha_0$, re-conduct similar steps as above, we have

$$\begin{aligned}
&P\{\inf_{\lambda\in\Omega_0} \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda}) - \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}}) > 0\} \\
&\geq P\{\inf_{\lambda\in\Omega_0} -2\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_\lambda}) - [-2\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}})] - \kappa_n \mathrm{df}_{\bar{\alpha}} > 0\} \\
&= P\{-2[\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_0}) - \ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}})] - \kappa_n \mathrm{df}_{\bar{\alpha}} > 0\} \\
&\geq P\{-2[\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_0}) - \ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}})] > M\mathrm{df}_{\bar{\alpha}}\} \quad (4.3.30) \\
&\to P\{\chi^2_{\mathrm{df}_{\bar{\alpha}}-\mathrm{df}_{\alpha_0}} \geq M\mathrm{df}_{\bar{\alpha}}\} > 0. \quad (4.3.31)
\end{aligned}$$

(4.3.30) is due to $\kappa_n < M$, and (4.3.31) uses the fact that $\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_0}$ and $\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}}$ are asymptotically normal under regular condition (A)-(D) in Fan and Li (2002). Hence, the likelihood ratio test statistics $-2[\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_0}) - \ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\bar{\alpha}})] \xrightarrow{\mathcal{L}} \chi^2_{\mathrm{df}_{\bar{\alpha}}-\mathrm{df}_{\alpha_0}}$. (4.3.25) follows from the above equations. Thus, we complete the proof of Lemma 1.

Lemma 2. *Suppose the density function of the Cox model satisfies four regularity condition (A1)-(A4). Then under Condition (E1)-(E4), and let* $\lambda_n = \kappa_n/\sqrt{n}$. *If*

κ_n satisfies $\kappa_n \to \infty$ and $\lambda_n \to 0$ as $n \to \infty$, we have

$$P\{GIC_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda_n}) = GIC_{\kappa_n}(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_0})\} \to 1, \tag{4.3.32}$$

$$P\left\{\inf_{\lambda \in (\Omega_- \cup \Omega_+)} GIC_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda}) > GIC_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda_n})\right\} \to 1. \tag{4.3.33}$$

Proof. With loss of generality, we assume that the first d_{α_0} component of $\boldsymbol{\beta}_0$ are nonzero for the true model while the rest are zeros. By Conditions (A1)-(A4) together with Condition (E3), Fan and Li (2002) showed that

$$\widehat{\beta}_{\lambda_n j} \xrightarrow{p} 0 \quad \text{for } j = d_{\alpha_0}+1, \cdots, d, \tag{4.3.34}$$

$$\frac{\partial}{\partial \beta_j}\ell_c(\widehat{\boldsymbol{\beta}}_{\lambda_n j}) - p'_{\lambda_n}(|\widehat{\beta}_{\lambda_n j}|)\mathrm{sgn}(\widehat{\beta}_{\lambda_n j}) \xrightarrow{p} 0 \qquad \text{for } j = 1, \cdots, d_{\alpha_0}, \tag{4.3.35}$$

where $\widehat{\beta}_{\lambda_n j}$ is the jth component of $\widehat{\boldsymbol{\beta}}_{\lambda_n}$. Under Condition (E1) and (E2), for $j = 1, \cdots, d_{\alpha_0}$, there exits a m such that

$$p'_{\lambda_n}(|\widehat{\beta}_{\lambda_n j}|) = 0 \text{ for } |\widehat{\beta}_{\lambda_n j}| \geq \min\{|\beta_{\lambda_n j}|\} \geq m\lambda_n.$$

By (4.3.35), with probability tending to 1, we have,

$$\frac{\partial}{\partial \beta_j}\ell_c(\widehat{\boldsymbol{\beta}}_{\lambda_n j}) = 0, \text{ for } j = 1, \cdots, d_{\alpha_0},$$

which is equal to solving the non-penalized partial likelihood function under the true model α_0. Therefore, with probability tending to 1, we have

$$\widehat{\boldsymbol{\beta}}_{\lambda_n} = \widehat{\boldsymbol{\beta}}^{\star}_{\alpha_0},$$
$$\ell_c(\widehat{\boldsymbol{\beta}}_{\lambda_n}) = \ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_0}).$$

Using the fact that $\mathrm{df}_{\alpha_{\lambda_n}} \xrightarrow{p} \mathrm{df}_{\alpha_0}$ when n is large enough, we have,

$$\begin{aligned}
&P\{\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda_n}) = \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_0})\} \\
&= P\{-2\ell_c(\widehat{\boldsymbol{\beta}}_{\lambda_n}) + \kappa_n \mathrm{df}_{\alpha_{\lambda_n}} + 2\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_0}) - \kappa_n \mathrm{df}_{\alpha_0} = 0\} \\
&= P\{-2[\ell_c(\widehat{\boldsymbol{\beta}}_{\lambda_n}) - \ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_0})] + \kappa_n(\mathrm{df}_{\alpha_{\lambda_n}} - \mathrm{df}_{\alpha_0}) = 0\} \\
&\to 1.
\end{aligned}$$

This validates (4.3.32).

Next, we want to show that $\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_\lambda) > \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda_n})$ for any λ that can not result in the true model. First, we consider λ that could result in underfitting models, namely, $\lambda \in \Omega_- = \{\lambda : \alpha_\lambda \not\supseteq \alpha_0\}$. By (4.3.27) and (4.3.32), with probability tending to 1, we have,

$$\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_\lambda) - \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda_n}) > -2\ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha_\lambda}) - [-2\ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha_0})] - \kappa_n \mathrm{df}_{\alpha_0}.$$

For any $\lambda \in \Omega_- = \{\lambda : \alpha \not\supseteq \alpha_0\}$, we can take $\inf_{\lambda\in\Omega_-}$ over $\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_\lambda)$. Under Condition (E4) and $\kappa_n/\sqrt{n} \to 0$, for any $\lambda \in \Omega_-$, we have

$$\begin{aligned}
&P\{\inf_{\lambda\in\Omega_-} \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_\lambda) - \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda_n}) > 0\} \\
&\geq P\{\inf_{\lambda\in\Omega_-} \frac{-2\ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha_\lambda})}{n} - \frac{-2\ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha_0})}{n} - \frac{\kappa_n \mathrm{df}_{\alpha_0}}{n} > 0\} \\
&= P\{\min_{\alpha\not\supseteq\alpha_0}[\frac{-2\ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha})}{n} - \log(n)\rho_1] - [\frac{-2\ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha_0})}{n} - \log(n)\rho_1] - \frac{\kappa_n \mathrm{df}_{\alpha_0}}{n} > 0\} \\
&= P\{\min_{\alpha\not\supseteq\alpha_0} c_\alpha - c_{\alpha_0} + o_{\mathbb{P}}(1) > 0\} \to 1, \qquad (4.3.36)
\end{aligned}$$

as $n \to \infty$. (4.3.36) is due to Condition (E4). Therefore, we validate that

$$P\left\{\inf_{\lambda\in\Omega_-} \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_\lambda) > \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda_n})\right\} \to 1. \qquad (4.3.37)$$

For any $\lambda \in \Omega_+ = \{\lambda : \alpha_\lambda \supset \alpha_0\}$, we have

$$\begin{aligned}
&\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_\lambda) - \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda_n}) \\
&= -2\ell_c(\widehat{\boldsymbol{\beta}}_\lambda) - [-2\ell_c(\widehat{\boldsymbol{\beta}}_{\lambda_n})] + \kappa_n(\mathrm{df}_{\alpha_\lambda} - \mathrm{df}_{\alpha_{\lambda_n}}) \\
&\geq -2\ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha_\lambda}) - [-2\ell_c(\widehat{\boldsymbol{\beta}}^\star_{\alpha_0})] + \kappa_n\tau_n, \qquad (4.3.38)
\end{aligned}$$

where $\tau_n > 0$ due to the fact that $\mathrm{df}_{\alpha_\lambda} - \mathrm{df}_{\alpha_{\lambda_n}} = \tau_n > 0$ when n is large. And (4.3.38) follows (4.3.26). We then take $\inf_{\lambda\in\Omega_+}$ over $\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_\lambda)$. Under Condition (E4) and $\kappa_n/\sqrt{n} \to 0$, for any $\lambda \in \Omega_+$, we have

$$\inf_{\lambda\in\Omega_+} \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_\lambda) - \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda_n})$$

$$\geq \min_{\alpha \not\supseteq \alpha_0} -2[\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha}) - \ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_0})] + \kappa_n \tau_n \tag{4.3.39}$$

$$\approx \kappa_n \tau_n. \tag{4.3.40}$$

(4.3.40) uses the fact that $2[\ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha}) - \ell_c(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha_0})] \to \chi^2_{\mathrm{df}_\alpha - \mathrm{df}_{\alpha_0}}$ for $\alpha \supset \alpha_0$ together with that $\kappa_n \to \infty$. Therefore, (4.3.39) is positive as $n \to \infty$. Hence, we have,

$$P\left\{ \inf_{\lambda \in \Omega_+} \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda}) > \mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda_n}) \right\} \to 1. \tag{4.3.41}$$

Based on (4.3.37) and (4.3.41) together, we proof (4.3.33). Consequently, we complete the proof of Lemma 2.

Lemma 1 implies that for any λ produces the underfitted model, its associated $\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda})$ is consistently larger than $\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}^{\star}_{\alpha})$. Thus, the optimal model selected by minimizing the $\mathrm{GIC}_{\kappa_n}(\boldsymbol{\beta})$ must be either the true model or overfitted models with probability tending to one. In addition, Lemma 1 indicates that there is a nonzero probability that the smallest value of $\mathrm{GIC}_{\kappa_n}(\widehat{\boldsymbol{\beta}}_{\lambda})$ associated with the true model is larger than that of the full model. As a result, there is a positive probability that any λ associated with the true model cannot be selected by $\mathrm{GIC}_{\kappa_n}(\boldsymbol{\beta})$ as the regularization parameter. Theorem 4(A) follows.

Lemma 2 indicates that the model identified by λ_n converges to the true model as the sample size gets large. In addition, it shows that those λ's, which fail to identify the true model, cannot be selected by $\mathrm{GIC}_{\kappa_n}(\boldsymbol{\beta})$ asymptotically. Theorem 4(B) follows.

We next show Theorem 4(C). Applying Taylor expansion to (4.3.19), we have

$$2n\mathrm{GCV}(\lambda) \approx -2\ell_c(\widehat{\boldsymbol{\beta}}_{\lambda}) + 4(-\ell_c(\widehat{\boldsymbol{\beta}}_{\lambda})/n)\mathrm{df}_{\lambda}.$$

Theorem 3 implies $-\ell_c(\widehat{\boldsymbol{\beta}}_{\lambda})/(n\log(n)) \to \rho_1 > 0$ as $n \to \infty$, then

$$\begin{aligned} 2n\mathrm{GCV}(\lambda) &\approx -2\ell_c(\widehat{\boldsymbol{\beta}}_{\lambda}) + 4\rho_1 \log(n)\mathrm{df}_{\lambda} \\ &= -2\ell_c(\widehat{\boldsymbol{\beta}}_{\lambda}) + \kappa_{gcv}\mathrm{df}_{\alpha_\lambda}, \end{aligned}$$

where $\kappa_{gcv} = 4\rho_1 \log(n)$. Hence, the $2n\mathrm{GCV}(\lambda)$ selector could be written as an special case of GIC_{κ_n} selector.

Since $\kappa_{gcv} = 4\rho_1 \log(n)$ satisfies conditions of Theorem 4(B), namely, $\kappa_{gcv} \to \infty$, and $\kappa_{gcv}/\sqrt{n} \to 0$. Therefore, by theorem 4(B), the tuning parameter $\widehat{\lambda}$ obtained by minimizing $2n\text{GCV}(\lambda)$ or equivalently $\text{GCV}(\lambda)$ satisfies

$$P\{\alpha_{\widehat{\lambda}} = \alpha_0\} \to 1.$$

This completes the proof of Theorem 4(C).

4.4 Numerical Results

In this section, we assess the finite sample performance of proposed procedures. In our simulation, we employ the local linear algorithm (LLA, Zou and Li, 2008) to compute the parameter estimates of the SCAD penalized partial likelihood function.

To assess the finite sample performance of the simulation studies, we report the percentage of models correctly fitted, underfitted, and overfitted with 1, 2, 3, 4, 5 or more parameters by SCAD-AIC, SCAD-GCV and SCAD-BIC selectors as well as via the oracle procedure (i.e., the simulated data were fitted with the true model). In addition, we report the average number of zero coefficients that were correctly (C) and incorrectly (IC) identified by the two selectors. To compare model fittings, we further calculate the following model error for the new observation $(\mathbf{x}, Z, \delta)$,

$$ME(\widehat{\boldsymbol{\beta}}) = E_{\mathbf{X}}\{\mu(\mathbf{x}^T\boldsymbol{\beta}) - \mu(\mathbf{x}^T\widehat{\boldsymbol{\beta}})\}^2,$$

where the expectation is taken with respect to the new observed covariate vector $\mathbf{x}$, and $\mu(\mathbf{x}^T\boldsymbol{\beta}) = E(t|\mathbf{x}, \boldsymbol{\beta})$. Then, we report the median of the relative model error (MRME), where the relative model error is defined as $RME = ME/ME_{full}$, and ME_{full} is the model error calculated by fitting the data with the full model.

Example 4.1. We adapt the model structure in Fan and Li (2002) to generate the data from the Cox model with the hazard function

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

where $h_0(y) \equiv 1$, $\boldsymbol{\beta} = (0.8, 0, 0, 1, 0, 0, 0, 0, 0, 0.6, 0, 0)^T$, and $\mathbf{x}$ is distributed accord-

ing to a 12-dimensional normal distribution with the correlation between x_i and x_j being $0.5^{|i-j|}$. Accordingly, $\mu(\mathbf{x}^T\boldsymbol{\beta}) = \exp(-\mathbf{x}^T\boldsymbol{\beta})$. Furthermore, the censoring distribution is exponential with mean $U\exp(-\mathbf{x}^T\boldsymbol{\beta})$, where U is sampled from a uniform distribution over $[1, 3]$. Consequently, the average censoring percentage is 35%. Moreover, we include the case with no censoring as benchmark.

In Fan and Li (2002), it has been shown that

$$ME(\widehat{\boldsymbol{\beta}}) = E_{\mathbf{X}}\{\mu(\mathbf{x}^T\boldsymbol{\beta}) - \mu(\mathbf{x}^T\widehat{\boldsymbol{\beta}})\}^2 = E_{\mathbf{X}}\{\exp(-\mathbf{x}^T\boldsymbol{\beta}) - \exp(-\mathbf{x}^T\widehat{\boldsymbol{\beta}})\}^2$$

for model used above. By using moment generating function of multinormal distribution, we can simplify the formula as

$$ME(\widehat{\boldsymbol{\beta}}) = \exp(2\widehat{\boldsymbol{\beta}}^T\Sigma\widehat{\boldsymbol{\beta}}) + \exp(2\boldsymbol{\beta}^T\Sigma\boldsymbol{\beta}) - 2\exp\{\frac{1}{2}(\widehat{\boldsymbol{\beta}} + \boldsymbol{\beta})^T\Sigma(\widehat{\boldsymbol{\beta}} + \boldsymbol{\beta})\}$$

, where Σ is the covariance matrix of $\mathbf{x}$. And we use this formula to calculate model errors for all our simulations.

For the uncensored case, Table 4.2 shows that the MRME of SCAD-BIC/GCV is smaller than SCAD-AIC. As sample size increases, the MRME of SCAD-BIC/GCV approaches that of the oracle estimator, whereas the MRME of SCAD-AIC remains at the same level. Hecne, SCAD-BIC/GCV is superior to SCAD-AIC in terms of model error measurements in this example.

In model identifications, SCAD-BIC/GCV has a higher chance than SCAD-AIC of correctly setting the nine true zero coefficients to zero. However, SCAD-BIC/GCV are more prone than SCAD-AIC to incorrectly set the three nonzero coefficients to zero when sample size is small, and SCAD-GCV tends to be more aggressive than SCAD-BIC with larger values in "IC" columns. In addition, SCAD-BIC/GCV has a much higher probability of correctly identifying the true model.

For the censored case, Table 4.3 shows findings similar to those presented in Table 4.2. Accordingly, SCAD-BIC/GCV are superior to SCAD-AIC in both identifying the true model and in reducing the model error and complexity. When the data are 35% censored, all methods decline slightly in their efficacy. However, the relative performance of SCAD-BIC/GCV versus SCAD-AIC remains the same as in the uncensored case. In sum, SCAD-BIC/GCV outperforms SCAD-AIC.

Table 4.2. Simulation results for the Cox model (**No Censoring**)

		MRME	Zeros		Under	Exact	Over Fitted (%)				
n	Method	(%)	C	IC	(%)	(%)	1	2	3	4	≥ 5
100	SCAD-AIC	45.75	7.255	0.001	0.1	37.5	15.3	16.1	13.1	8.5	9.4
	SCAD-BIC	20.90	8.576	0.003	0.3	74.0	15.7	5.2	3.8	0.9	0.1
	SCAD-GCV	17.29	8.940	0.059	5.6	89.2	4.9	0.3	0.0	0.0	0.0
	AIC	52.52	7.349	0.001	0.1	20.1	29.4	26.9	15.3	6.0	2.2
	BIC	25.68	8.666	0.004	0.4	72.5	22.6	3.4	1.1	0.0	0.0
	Oracle	15.73	9.000	0.000	0.0	100.0	0.0	0.0	0.0	0.0	0.0
200	SCAD-AIC	58.53	7.591	0.000	0.0	46.6	15.9	13.8	8.5	7.3	7.9
	SCAD-BIC	36.33	8.867	0.000	0.0	90.1	7.3	1.8	0.8	0.0	0.0
	SCAD-GCV	33.96	8.995	0.003	0.3	99.2	0.5	0.0	0.0	0.0	0.0
	AIC	66.37	7.506	0.000	0.0	23.1	32.2	24.8	13.8	4.5	1.6
	BIC	41.89	8.781	0.000	0.0	81.2	16.1	2.3	0.4	0.0	0.0
	Oracle	33.95	9.000	0.000	0.0	100.0	0.0	0.0	0.0	0.0	0.0
400	SCAD-AIC	68.14	7.553	0.000	0.0	45.1	14.7	15.4	9.3	9.1	6.4
	SCAD-BIC	44.10	8.936	0.000	0.0	94.7	4.4	0.7	0.2	0.0	0.0
	SCAD-GCV	42.33	8.999	0.000	0.0	99.9	0.1	0.0	0.0	0.0	0.0
	AIC	74.71	7.530	0.000	0.0	22.6	33.4	25.7	12.2	5.0	1.1
	BIC	47.38	8.875	0.000	0.0	88.6	10.5	0.7	0.2	0.0	0.0
	Oracle	42.30	9.000	0.000	0.0	100.0	0.0	0.0	0.0	0.0	0.0

Example 4.2. (Heart attack data) We downloaded the heart attack data set, which was provided by Dr. Robert Goldberg from the University of Massachusetts and used in the book of Hosmer and Lemeshow (1999), from the web site of John Wiley & Sons Inc. In addition, the data comes from the Worcester Heart Attack Study which describes trends over time in the survival rates following hospital admission for acute myocardial infarction. The total length of follow-up on the admission of 481 hospital patients was recorded for years 1975, 1978, 1981, 1984, 1986, and 1988. Among those patients, 249 died and the rest were censored so that the censoring rate is 48%.

To model the total length of survival time, Hosmer and Lemeshow (1999) suggested fitting the Cox proportional hazards model with the following five explanatory variables: x_1-*age*; x_2-*cpk* (peak cardiac enzymes in international units); x_3-*sex* (male=0 and female=1); x_4-*chf* (left heart failure complications, yes=1 and no=0); x_5-*miord* (MI order, first=0 and recurrent=1). In addition to these variables, we include the six interactions between the two continuous variables (*age* and *cpk*) and

Table 4.3. Simulation results for the Cox model (**35% Censoring**)

		MRME	Zeros		Under	Exact	Over Fitted (%)				
n	Method	(%)	C	IC	(%)	(%)	1	2	3	4	≥ 5
100	SCAD-AIC	42.43	7.235	0.012	1.2	33.4	18.8	16.6	12.0	8.5	9.5
	SCAD-BIC	21.42	8.491	0.060	5.8	63.4	19.7	7.3	2.4	1.1	0.3
	SCAD-GCV	19.04	8.800	0.153	13.6	71.6	12.3	2.1	0.3	0.1	0.0
	AIC	50.03	7.370	0.016	1.6	20.4	30.0	25.9	13.6	6.5	2.0
	BIC	23.45	8.648	0.036	3.6	68.8	23.7	3.5	0.4	0.0	0.0
	Oracle	14.35	9.000	0.000	0.0	100.0	0.0	0.0	0.0	0.0	0.0
200	SCAD-AIC	59.24	7.535	0.000	0.0	42.3	19.5	13.2	10.2	7.7	7.1
	SCAD-BIC	35.53	8.841	0.000	0.0	87.4	9.8	2.3	0.5	0.0	0.0
	SCAD-GCV	32.48	8.963	0.006	0.6	95.9	3.3	0.2	0.0	0.0	0.0
	AIC	64.64	7.513	0.000	0.0	22.8	35.5	21.5	12.1	6.9	1.2
	BIC	37.90	8.830	0.000	0.0	84.8	13.5	1.6	0.1	0.0	0.0
	Oracle	31.45	9.000	0.000	0.0	100.0	0.0	0.0	0.0	0.0	0.0
400	SCAD-AIC	69.31	7.552	0.000	0.0	41.5	19.5	14.7	10.6	7.4	6.3
	SCAD-BIC	45.07	8.920	0.000	0.0	93.2	5.7	1.0	0.1	0.0	0.0
	SCAD-GCV	42.75	8.993	0.000	0.0	99.4	0.5	0.1	0.0	0.0	0.0
	AIC	73.64	7.547	0.000	0.0	23.8	33.7	24.7	11.6	4.5	1.7
	BIC	48.85	8.856	0.000	0.0	86.8	12.0	1.2	0.0	0.0	0.0
	Oracle	43.47	9.000	0.000	0.0	100.0	0.0	0.0	0.0	0.0	0.0

the three indicator variables (*sex, chf*, and *miord*) (i.e., x_6-*age*sex*, x_7-*age*chf*, x_8-*age*miord*, x_9-*cpk*sex*, x_{10}-*cpk*chf*, and x_{11}-*cpk*miord*). In sum, there are a total of 11 variables in the full model. For the sake of comparison, we also fit the data with the penalized partial likelihood approach, and the resulting regularization parameters selected by SCAD-AIC and SCAD-BIC are 0.0533, 0.0878, and 0.1326 respectively. The corresponding tuning parameters selector curves are illustrated in Figure 4.2.

Table 4.4 presents the maximum partial penalized likelihood estimates (MPLE) from the full model as well as the SCAD-AIC/BIC/GCV parameter estimates, together with their standard errors. It shows that the full model contains six spurious variables (x_2, x_3, and x_7 to x_{10}) and SCAD-AIC includes two insignificant variables (x_6 and x_{10}) at a level of 0.05. In contrast, all four variables (x_1, x_4, x_5, and x_{11}) selected by SCAD-BIC are significant. In this specific data, SCAD-GCV tends to be too aggressive by excluding x_5, and x_{11}, which is similar to the results

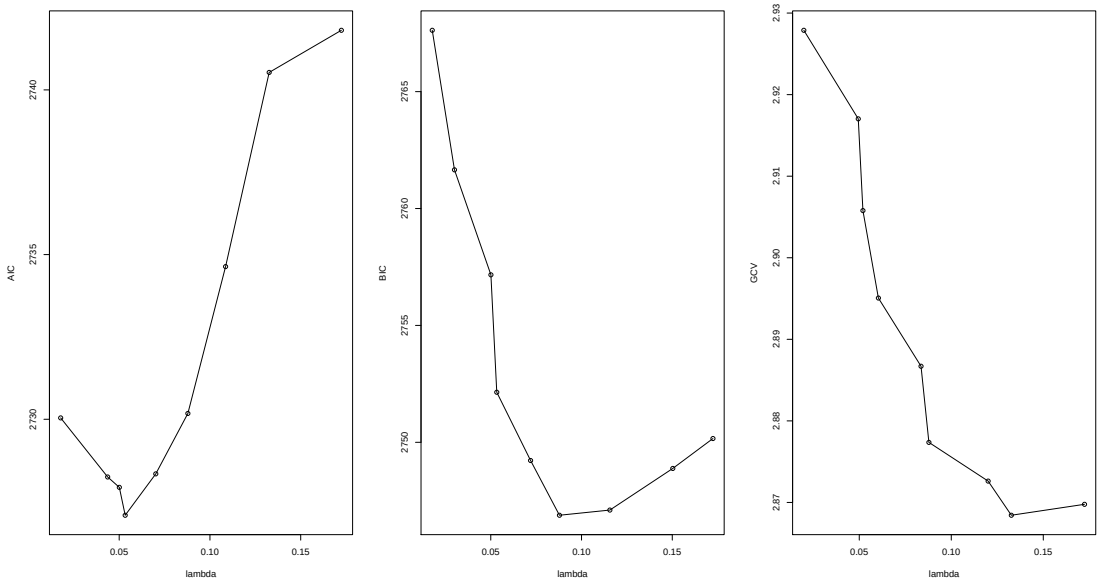

Figure 4.2. The tuning parameter selector curve.
[The tuning parameter selector curve based on SCAD-AIC, SCAD-BIC, and SCAD-GCV respectively from left to the right.]

Table 4.4. Estimates for heart disease survival data with standard deviations in parentheses based on LLA

	MPLE	SCAD-AIC	SCAD-BIC	SCAD-GCV
age (x_1)	0.60(0.13)	0.56(0.09)	0.43(0.07)	0.41(0.05)
cpk (x_2)	0.03(0.14)	0(-)	0(-)	0(-)
sex (x_3)	0.17(0.14)	0(-)	0(-)	0(-)
chf (x_4)	0.80(0.14)	0.80(0.13)	0.80(0.14)	0.82(0.13)
miord (x_5)	0.42(0.14)	0.43(0.13)	0.41(0.13)	0(-)
age*sex(x_6)	-0.29(0.14)	-0.22(0.13)	0(-)	0(-)
age*chf (x_7)	-0.07(0.15)	0(-)	0(-)	0(-)
age*miord (x_8)	0.03(0.15)	0(-)	0(-)	0(-)
cpk*sex (x_9)	-0.16(0.16)	0(-)	0(-)	0(-)
cpk*chf (x_{10})	0.19(0.15)	0.19(0.09)	0(-)	0(-)
cpk*miord (x_{11})	0.29(0.15)	0.25(0.12)	0.21(0.05)	0(-)

of simulations when sample size is not large enough.

Based on the Table 4.4, the p-value of the partial likelihood ratio test for examining the SCAD-AIC, SCAD-BIC, and SCAD-GCV model versus the full model

are 0.6752, 0.1749, and 0.0034 respectively. Consequently, there is no evidence of lack of fit in the SCAD-BIC model. However, SCAD-GCV model may be too aggressive, which is also indicated by our simulation results that GCV tends to be underfitted when sample size is not large enough.

4.5 Discussions

In this chapter, we show that the sample average of Cox's partial likelihood tends to infinity at the rate of $\log(n)$ under some mild conditions. We study how to select the regularization parameter in the penalized partial likelihood with the SCAD penalty, and demonstrate the GCV tuning parameter selector behaves differently from the one for the penalized least squares with random samples from a linear regression models. The results are applicable for penalized partial likelihood with other folded concave penalty such as the MCP (Zhang, 2010).

Chapter 5

Conclusion and Future Research

5.1 Conclusion Remarks

In Chapter 3, a new sure joint screening (SJS) procedure is proposed for ultrahigh dimensional Cox's models based on maximizing the constrained partial likelihood. SJS is different from other existing feature screening methods since it considers the effects of predictors simultaneously rather than the marginal utilities. This mechanism helps SJS to avoid the iterative screening procedure when some of predictors are related to the responses jointly but not independently. For high dimensional problems, it becomes almost impossible to solve the constrained maximization problem (3.1.3) directly. We consider a proxy resulted by the Taylor expansion of the partial likelihood function as the alternative and replace the corresponding Hessian matrix with a diagonal working matrix. Simulation results shows the effectiveness of the algorithm. More importantly, it is proved that we do not lose the sure screening property with the new procedure given constraint partial likelihood approximation.

In Chapter 4, we study the asymptotic behavior of partial likelihood for the Cox model. We find that the partial likelihood does not behave like an ordinary likelihood, whose sample average typically tends to its expected value, a finite number, in probability. Under some mild conditions, we prove that the sample average of partial likelihood tends to infinity at the rate of logarithm of the sample size in probability. We also study tuning parameter selection for penalized partial likelihood in this chapter. Our finding indicates that the penalized partial likeli-

hood with the generalized cross-validation (GCV) tuning parameter proposed in Fan and Li (2002) enjoys the model selection consistency property. This is an surprising result because it is well known that the GCV, AIC and C_p are all equivalent in the context for linear regression models, and are not model selection consistent. Our empirical studies via Monte Carlo simulation and real data example confirms our theoretical finding.

5.2 Future Research

5.2.1 Cox's Model with Varying Coefficient

The Cox proportional hazard model (Cox, 1972) has been widely used to model the relationships between the event time and the time-invariant covariates. In the conventional form of the Cox model, the coefficients are assumed to be constant, thus guaranteeing the hazard is proportional across time. In practice, however, covariates may interact nonlinearly with other exposure variables. For example, the effects of a treatment may vary with age. The variation often cannot be calibrated properly in a parametric form, thus, a varying coefficient Cox model may be considered instead.

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

$$h(t|\mathbf{x}, u) = h_0(t) \exp(\mathbf{x}^T \boldsymbol{\beta}(u)), \tag{5.2.1}$$

where $h_0(t)$ is an unspecified baseline hazard functions and u is a continuous variable associated to covarite effects $\boldsymbol{\beta}$ in Cox's model, and $\boldsymbol{\beta}(u) = (\beta_1(u), \cdots, \beta_p(u))^T$ is consist of unknown coefficient functions. In survival data analysis, the survival time may be censored by a censoring time C. Denote the observed time by $Z = \min\{T, C\}$ and the censoring 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 $\{(u_i, \mathbf{x}_i, Z_i, \delta_i) : i = 1, \cdots, n\}$ is an independently and identically distributed random sample from model (5.2.1), where $t_1^0 < \cdots < t_N^0$ be the ordered observed failure times. Let (j) be 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, that is,

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

The partial likelihood function of the random sample is

$$\ell_p(\boldsymbol{\beta}) = \sum_{j=1}^{N} [\mathbf{x}_{(j)}^T \boldsymbol{\beta}(u_j) - \log\{\sum_{i \in R_j} \exp(\mathbf{x}_i^T \boldsymbol{\beta}(u_i))\}], \tag{5.2.2}$$

where the time-varying coefficients $\boldsymbol{\beta}(u)$ are assumed to be smooth functions of time and need to be estimated nonparametrically. We will considered using smoothing splines, say B-spline, to estimate the coefficient functions $\boldsymbol{\beta}(\cdot)$. Hence, the coefficient functions on a given set of knots can be expressed as a linear combinations of B-splines. Therefore, we have

$$\beta_j(u) \approx \sum_{i=1}^{m_j} \alpha_{j,i} B_{j,i}(u), \ j = 1, \cdots, p, \tag{5.2.3}$$

where $B_{j,i}(u), i = 1, \cdots, m_j$ is a set of B-spline basis function. (5.2.3) may be further written as

$$\beta_j(u) \approx \boldsymbol{\alpha}_j \mathbf{b}_j(u), \ j = 1, \cdots, p, \tag{5.2.4}$$

where $\boldsymbol{\alpha}_j = (\alpha_{j,1}, \cdots, \alpha_{j,m_j})$ and $\mathbf{b}_j = (B_{j,1}(u), \cdots, B_{j,m_j}(u))^T$. Therefore, (5.2.1) could be written as

$$h(t|\mathbf{x}, u) = h_0(t) \exp(\sum_{j=1}^{p} \beta_j(u) X_j) \approx h_0(t) \exp(\sum_{j=1}^{p} \boldsymbol{\alpha}_j \mathbf{b}_j(u) X_j). \tag{5.2.5}$$

Let $\mathbf{y} = (\mathbf{b}_1(u)X_1, \cdots, \mathbf{b}_p(u)X_p)^T$ and $\tilde{\boldsymbol{\alpha}} = (\boldsymbol{\alpha}_1^T, \cdots, \boldsymbol{\alpha}_p^T)^T$, we have

$$h(t|\mathbf{x}, u) \approx h_0(t) \exp(\mathbf{y}^T \tilde{\boldsymbol{\alpha}}). \tag{5.2.6}$$

Therefore, the varying-coefficient Cox model can be casted to a traditional Cox model with group variables by the technique of B-spline approximation. The idea of SJS can be applied to this problem. To conduct the screening procedure, we

need to maximize the pseduo-likelihood function under some constrains. Namely,

$$\max_{\boldsymbol{\beta}(\cdot)} \ell_p\{\boldsymbol{\beta}(\cdot)\} \quad \text{subject to } \#(\{j : \|\beta_j(\cdot)\|_2^2 > 0\}) \leq m$$

for a pre-specified m that is much smaller than p. As

$$\beta_j(u) \approx \boldsymbol{\alpha}_j \mathbf{b}_j(u),\ j = 1, \cdots, p,$$

$\beta_j(u)$ depends on $\boldsymbol{\alpha}_j$ and $\mathbf{b}_j = (B_{j,1}(u), \cdots, B_{j,m_j}(u))^T$. Note that $\|\beta_j(\cdot)\|_2^2 = \boldsymbol{\alpha}_j E\{\mathbf{b}_j(\cdot)\mathbf{b}_j(\cdot)^T\}\boldsymbol{\alpha}_j^T$. Under the assumption that $E\{\mathbf{b}_j(\cdot)\mathbf{b}_j(\cdot)^T\}$ is finite positive definite for all $j = 1, \cdots, p$, the maximization problem above is equivalent to

$$\max_{\tilde{\boldsymbol{\alpha}}} \ell_p(\tilde{\boldsymbol{\alpha}}) \quad \text{subject to } \#(\{j : \|\boldsymbol{\alpha}_j\|_2^2 > 0\}) \leq m \tag{5.2.7}$$

For high dimensional problems, it becomes almost impossible to solve the constrained maximization problem (5.2.7) directly. Following the strategy in Chapter 3, we can consider a proxy of the target function. It follows by the Taylor expansion for the target function $\ell_p(\boldsymbol{\gamma})$ at $\tilde{\boldsymbol{\alpha}}$ lying within a neighbor of $\boldsymbol{\gamma}$ that

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

where $\ell_p'(\tilde{\boldsymbol{\alpha}}) = \partial \ell_p(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}|_{\boldsymbol{\gamma}=\tilde{\boldsymbol{\alpha}}}$ and $\ell_p''(\tilde{\boldsymbol{\alpha}}) = \partial^2 \ell_p(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}^T|_{\boldsymbol{\gamma}=\tilde{\boldsymbol{\alpha}}}$. The computational complexity of calculating the inverse of $\ell_p''(\tilde{\boldsymbol{\alpha}})$ is $O\{(\sum_m^p m_j)^3\}$. For the setting of large p and small n, $\ell_p''(\tilde{\boldsymbol{\alpha}})$ 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 $\ell_p''(\boldsymbol{\gamma})$

$$h(\boldsymbol{\gamma}|\tilde{\boldsymbol{\alpha}}) = \ell_p(\tilde{\boldsymbol{\alpha}}) + (\boldsymbol{\gamma} - \tilde{\boldsymbol{\alpha}})^T \ell_p'(\tilde{\boldsymbol{\alpha}}) - \frac{u}{2}(\boldsymbol{\gamma} - \tilde{\boldsymbol{\alpha}})^T \mathbf{W}(\tilde{\boldsymbol{\alpha}})(\boldsymbol{\gamma} - \tilde{\boldsymbol{\alpha}}), \tag{5.2.8}$$

where u is a scaling constant to be specified and $\mathbf{W}(\tilde{\boldsymbol{\alpha}}) = \text{diag}(W_1(\tilde{\boldsymbol{\alpha}}), \cdots, W_p(\tilde{\boldsymbol{\alpha}}))$ is a diagonal blockmatrix. $W_j(\tilde{\boldsymbol{\alpha}})$ is an $m_j \times m_j$ matrix and defined as $W_j(\tilde{\boldsymbol{\alpha}}) = \ell_p(\tilde{\boldsymbol{\alpha}})/\partial \boldsymbol{\alpha}_j \partial \boldsymbol{\alpha}_j{}^T$. This implies that we approximate $\ell_p''(\tilde{\boldsymbol{\alpha}})$ by $-u\mathbf{W}(\tilde{\boldsymbol{\alpha}})$, which has the same spirits of SJS.

Therefore, a similar 2-step efficient computational algorithm may be employed to obtain the solution of $\tilde{\boldsymbol{\alpha}}$. It can be seen that $h(\tilde{\boldsymbol{\alpha}}|\tilde{\boldsymbol{\alpha}}) = \ell_p(\tilde{\boldsymbol{\alpha}})$, and under some conditions, $h(\boldsymbol{\gamma}|\tilde{\boldsymbol{\alpha}}) \leq \ell_p(\tilde{\boldsymbol{\alpha}})$ for all $\boldsymbol{\gamma}$. This ensures the ascent property, which needs to be studied and proved in future. Since $\mathbf{W}(\tilde{\boldsymbol{\alpha}})$ is a diagonal block matrix, $h(\boldsymbol{\gamma}|\tilde{\boldsymbol{\alpha}})$ is an additive function of $\boldsymbol{\gamma}_j$ for any given $\tilde{\boldsymbol{\alpha}}$. The additivity enables us to have a closed form solution for the following maximization problem

$$\max_{\boldsymbol{\gamma}} h(\boldsymbol{\gamma}|\tilde{\boldsymbol{\alpha}}) \qquad \text{subject to } \#(\{j : \|\boldsymbol{\alpha}_j\|_2^2 > 0\}) \leq m \tag{5.2.9}$$

for given $\tilde{\boldsymbol{\alpha}}$ and m.

Define $\tilde{\boldsymbol{\gamma}}_j = \boldsymbol{\alpha}_j + u^{-1}W^{-1}\partial\ell_p(\tilde{\boldsymbol{\alpha}})/\partial\boldsymbol{\alpha}_j$ for $j = 1, \cdots, p$. Then $\tilde{\boldsymbol{\gamma}} = (\tilde{\boldsymbol{\gamma}}_1^T, \cdots, \tilde{\boldsymbol{\gamma}}_p^T)^T = \tilde{\boldsymbol{\alpha}} + u^{-1}W^{-1}(\tilde{\boldsymbol{\alpha}})\ell_p'(\tilde{\boldsymbol{\alpha}})$ is the maximizer of $h(\boldsymbol{\gamma}|\tilde{\boldsymbol{\alpha}})$. Denote $g_j = \tilde{\boldsymbol{\gamma}}_j^T W_j(\tilde{\boldsymbol{\alpha}})\tilde{\boldsymbol{\gamma}}_j$ for $j = 1, \cdots, p$, and sort g_j so that $g_{(1)} \geq g_{(2)} \geq \cdots \geq g_{(p)}$. The solution of maximization problem (5.2.9) is the hard-thresholding rule defined as

$$\widehat{\boldsymbol{\gamma}}_j = \tilde{\boldsymbol{\gamma}}_j I\{g_j > g_{(m+1)}, \}$$

which enables us to effectively screen features by using the following algorithm.

Step 1. Set the initial value $\tilde{\boldsymbol{\alpha}}_j^{(0)} = \mathbf{0}$, $j = 1, \cdots, p$.

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

Step 2a. Calculate

$$\tilde{\boldsymbol{\gamma}}_j^{(t)} = \boldsymbol{\alpha}_j^{(t)} + u_t^{-1}W_j^{-1}\partial\ell_p(\tilde{\boldsymbol{\alpha}}^{(t)})/\partial\boldsymbol{\alpha}_j,$$

and

$$g_j^{(t)} = \{\tilde{\boldsymbol{\gamma}}_j^{(t)}\}^T W_j(\tilde{\boldsymbol{\alpha}}^{(t)})\tilde{\boldsymbol{\gamma}}_j.$$

Let $g_{(1)}^{(t)} \geq g_{(2)}^{(t)} \geq \cdots \geq g_{(p)}^{(t)}$, the order statistics of $g_j^{(t)}$s. Set $S_t = \{j : \|g_j^{(t)} \geq g_{(m+1)}^{(t)}\}$, the nonzero index set.

Step 2b. Update $\tilde{\boldsymbol{\alpha}}$ by $\tilde{\boldsymbol{\alpha}}^{(t+1)} = (\boldsymbol{\alpha}_1^{(t+1)}, \cdots, \boldsymbol{\alpha}_p^{(t+1)})^T$ as follows. If $j \notin S_t$, set $\boldsymbol{\alpha}_j^{(t+1)} = \mathbf{0}$, otherwise, set $\{\boldsymbol{\alpha}_j^{(t+1)} : j \in S_t\}$ be the maximum likelihood estimate of the submodel S_t.

5.2.2 Extension to other Survival Models with Parametric Covariate Effect

It is assumed for the Cox proportional hazards model that the survival time of subjects are independent. However, this assumptions might be violated when the collected data are correlated. One popular approach to model correlated survival times is to use a frailty model. A frailty corresponds to a random block effect that acts on the hazard rates of all subjects in a group.

Cox's frailty model (Vaupel et al., 1979) is defined as follows. The hazard rate for the jth subject in ith group is

$$h_{ij}(t|\mathbf{x}_{ij}) = h_0(t)u_i\exp(\mathbf{x}_{ij}^T\boldsymbol{\beta}),\ i = 1, \cdots, m,\ j = 1, \cdots, J_m, \tag{5.2.10}$$

where u_i is associated with frailties and follows a specific distribution, such as gamma frailty.

The full likelihood of the observed data $\{(u_i, \mathbf{x}_{ij}, T_{ij}, \delta_{ij}) :\ i = 1, \cdots, m,\ j = 1, \cdots, J_m\}$ could be written as $L(u_i, \mathbf{x}_{ij}, Z_{ij}, \delta_{ij})$. Integrating the full likelihood function with respect to $u_1, \cdots, u_m$, the likelihood could expressed as $L\{\alpha, \boldsymbol{\beta}, H(\cdot)|\mathbf{X}, \mathbf{T}, \boldsymbol{\delta}\}$, where α is the parameter in the distribution of frailty. $H(\cdot)$ is the baseline cumulative hazard function which could be estimated by

$$\widehat{H}(z) = \sum_{i}^{N} h_i\mathbf{I}(t_i^0 \leq z), \tag{5.2.11}$$

where $t_i^0, \cdots, t_N^0$ are the pooled failure time as defined in Chapter 3. Therefore, the logarithm of the profile likelihood could be expressed as $\ell\{\alpha, \boldsymbol{\beta}, H(\cdot)\}$. This is distinguished from the partial likelihood function in that $\ell\{\alpha, \boldsymbol{\beta}, H(\cdot)\}$ contains nuisance parameter α which is not subject to screening.

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. Therefore, the ***Proportional odds model*** was introduced as

$$\frac{S(t|\mathbf{x})}{1 - S(t|\mathbf{x})} = \frac{S_0(t|\mathbf{x})}{1 - S_0(t|\mathbf{x})}\exp(\mathbf{x}^T\boldsymbol{\beta}), \tag{5.2.12}$$

where $S_0(t|\mathbf{x})$ is the baseline function of an unspecific form. Similarly, the logarithm of likelihood function of proportional odds model could be written as $\ell\{\boldsymbol{\beta}, H(\cdot)\}$.

Accelerated failure time (AFT) model (Collett, 2003) is another common used survival model in practice. It describes the logarithm of survival time using a conventional linear model. The AFT model can be expressed as

$$\log(T) = \mathbf{x}^T\boldsymbol{\beta} + \sigma\epsilon, \tag{5.2.13}$$

where ϵ is a random error term with a distribution to be specified, and σ is a scalar. This model specifies the distribution of log-survival as a simple shift of a baseline distribution represented by the error term. Similarly, the logarithm for its corresponding likelihood function could be expressed as $\ell\{\boldsymbol{\alpha}, \boldsymbol{\beta}, H(\cdot)\}$, where $\boldsymbol{\alpha} = (\sigma)$ describes the signal strengh of error measurement.

In summary, the logarithm of profile likehood for all these three models above could be expressed in the same form $\ell\{\boldsymbol{\alpha}, \boldsymbol{\beta}, H(\cdot)\}$. We can adopt the same ideas as SJS and conduct feature screening with the corresponding efficient algorithm as follows. To conduct the screening procedure, we need to maximize the profile likelihood function under some constrains. Substituting (5.2.11) into $\ell\{\boldsymbol{\alpha}, \boldsymbol{\beta}, H(\cdot)\}$, then differentiating it with respect to H and the roots of the corresponding score function is $\widehat{H}(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Hence, we need to figure out the constraint maximum of $\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}, \widehat{H}) = \ell(\boldsymbol{\alpha}, \boldsymbol{\beta})$ with $\widehat{H}(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Namely,

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \ell(\boldsymbol{\alpha}, \boldsymbol{\beta}) \quad \text{subject to } \#(\{j : \beta_j > 0\}) \leq m \tag{5.2.14}$$

for a pre-specified m that is smaller than p.

Denote $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$. For high dimensional problems, it becomes almost impossible to solve the constrained maximization problem (5.2.14) directly. Following the computational algorithm idea in Chapter 3, we can consider a proxy of the target function. Denote $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{\boldsymbol{\alpha}}^T, \gamma_1, \cdots, \gamma_p)^T$. It follows by the Taylor expansion for the target function $\ell(\boldsymbol{\gamma})$ at $\boldsymbol{\theta}$ lying within a neighbor of $\boldsymbol{\gamma}$ that

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

where $\ell'(\boldsymbol{\theta}) = \partial\ell(\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}|_{\boldsymbol{\gamma}=\boldsymbol{\theta}}$ and $\ell''(\boldsymbol{\theta}) = \partial^2\ell(\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T|_{\boldsymbol{\gamma}=\boldsymbol{\theta}}$. When If $\ell''(\boldsymbol{\theta})$

is invertible, The computational complexity of calculating the inverse of $\ell''(\boldsymbol{\theta})$ is $O\{(p+d_{\boldsymbol{\alpha}})^3\}$, where $d_{\boldsymbol{\alpha}}$ is the dimension of covariate $\boldsymbol{\alpha}$. For the setting of large p and small n, $\ell''(\boldsymbol{\theta})$ is not invertible. Low computational cost is always desirable for feature screening. To deal with singularity of the Hessian matrix and save computational costs, we propose using the following approximation for $\ell''(\boldsymbol{\gamma})$

$$h(\boldsymbol{\gamma}|\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + (\boldsymbol{\gamma}-\boldsymbol{\theta})^T\ell'(\boldsymbol{\theta}) - \frac{1}{2}(\boldsymbol{\gamma}-\boldsymbol{\theta})^T\mathbf{W}(\boldsymbol{\theta})(\boldsymbol{\gamma}-\boldsymbol{\theta}), \tag{5.2.15}$$

where $\mathbf{W}(\boldsymbol{\theta}) = \text{diag}(W_{\boldsymbol{\alpha}}(\boldsymbol{\theta}), uW_1(\boldsymbol{\theta}), \cdots, uW_p(\boldsymbol{\theta}))$ is a diagonal block matrix and u is a scaling constant to be specified. $W_{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = -\partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T$ and $W_j(\boldsymbol{\theta}) = -\partial^2\ell(\boldsymbol{\theta})/\partial^2\beta_j$. This implies that we approximate $\ell''(\boldsymbol{\theta})$ by $-\mathbf{W}(\boldsymbol{\theta})$, which share the same spirits of SJS.

A two-step efficient computational algorithm could be advocated to obtain the solution of $\boldsymbol{\theta}$. It can be proved that $h(\boldsymbol{\theta}|\boldsymbol{\theta}) = \ell(\boldsymbol{\theta})$, and under some conditions, $h(\boldsymbol{\gamma}|\boldsymbol{\theta}) \leq \ell(\boldsymbol{\theta})$ for all $\boldsymbol{\gamma}$. This ensures the ascent property. Since $\mathbf{W}(\boldsymbol{\theta})$ is a diagonal block matrix, $h(\boldsymbol{\gamma}|\boldsymbol{\theta})$ is an additive function of $\boldsymbol{\gamma}_j$ for any given $\boldsymbol{\theta}$. The additivity enable us to have a closed form solution for the following maximization problem

$$\max_{\boldsymbol{\gamma}} h(\boldsymbol{\gamma}|\boldsymbol{\theta}) \qquad \text{subject to } \#(\{j:\beta_j>0\}) \leq m \tag{5.2.16}$$

for given $\boldsymbol{\theta}$ and m. Define $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{\alpha}} = \boldsymbol{\alpha} + W_{\boldsymbol{\alpha}}^{-1}\partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\alpha}$, $\tilde{\gamma}_j = \beta_j + u^{-1}W_j^{-1}\partial\ell(\boldsymbol{\theta})/\partial\beta_j$ for $j = 1, \cdots, p$, and $\widehat{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\gamma}}_{\boldsymbol{\alpha}}^T, \tilde{\gamma}_1, \cdots, \tilde{\gamma}_p)^T$ is the maximizer of $h(\boldsymbol{\gamma}|\boldsymbol{\theta})$. Denote $g_j = W_j\tilde{\gamma}_j^2$ for $j = 1, \cdots, p$, and sort g_j so that $g_{(1)} \geq g_{(2)} \geq \cdots \geq g_{(p)}$. The solution of maximization problem (5.2.16) is the hard-thresholding rule defined below

$$\widehat{\boldsymbol{\gamma}}_{\boldsymbol{\alpha}} = \tilde{\boldsymbol{\gamma}}_{\boldsymbol{\alpha}};\ \widehat{\gamma}_j = \tilde{\gamma}_j I\{g_j > g_{(m+1)}\}$$

This enable us to effectively screen features by using the following algorithm.

Step 1. Set the initial value $\boldsymbol{\alpha} = \mathbf{0}$, $\beta_j = 0$, $j = 1, \cdots, p$.

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

Step 2a. Calculate

$$\tilde{\boldsymbol{\gamma}}_{\boldsymbol{\alpha}}^{(t)} = \boldsymbol{\alpha}^{(t)} + W_{\boldsymbol{\alpha}}^{-1}\partial\ell(\boldsymbol{\theta}^{(t)})/\partial\boldsymbol{\alpha},$$

$$\tilde{\gamma}_j^{(t)} = \beta_j^{(t)} + u^{-1}W_j^{-1}\partial\ell(\boldsymbol{\theta}^{(t)})/\partial\beta_j,$$

and

$$g_j^{(t)} = W_j(\tilde{\boldsymbol{\alpha}}^{(t)})\{\tilde{\gamma}_j^{(t)}\}^2.$$

Let $g_{(1)}^{(t)} \geq g_{(2)}^{(t)} \geq \cdots \geq g_{(p)}^{(t)}$, the order statistics of $g_j^{(t)}$s. Set $S_t = \{j : g_j^{(t)} \geq g_{(m+1)}^{(t)}\}$, the nonzero index set.

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

The asymptotic properties, such as the ascent property of the proposed algorithm and sure screening property, deserves being further studied in details based on different scenarios aforementioned while with similar proof techniques in Chapter 3.

Bibliography

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

[2] Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Trans. on Automatic Control*, **19**, 716–723.

[3] Allen, D. M. (1974), "The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction," *Technometrics,* **16**, 125–127.

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

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

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

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

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

[9] Breheny, P. and Huang, J. (2011), "Coordinate Descent Algorithms for Nonconvex Penalized Regression, with Applications to Biological Feature Selection," *The Annals of Applied Statistics*, **5**, 232–253.

[10] Breiman, L. (1995), "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, **37**, 373–384.

[11] Cai, J., Fan, J., Li, R., and Zhou, H. (2005), "Variable Selection for Multivariate Failure Time Data," *Biometrika*, **92**, 303–316.

[12] Candes, E. and Tao, T. (2007), "The Dantzig Selector: Statistical Estimation When p is Much Larger Than n" (with discussion), *The Annals of Statistics*, **35**, 2313–2404.

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

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

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

[16] 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.

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

[18] Faraggi, D. and Simon, R. (1998), "Bayesian Variable Selection Method for Censored Survival Data," *Biometrics*, **54**, 1475–1485.

[19] 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.

[20] 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.

[21] 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.

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

[23] Fan, J. and Li, R. (2006), "Statistical Challenges with High-dimensionality: Feature Selection in Knowledge Discovery," *Proceedings of International Congress of Mathematicians*, **Vol. III**, 595–622.

[24] 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.

[25] 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.

[26] Fan, J. and Peng, H. (2004), "Nonconcave Penalized Likelihood With a Diverging Number of Parameters," *The Annals of Statistics*, **32**, 928–961.

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

[28] Fan, J. and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models with NP-dimensionality," *The Annals of Statistics*, **38**, 3567–3604.

[29] Fan, J., Xue, L. and Zou, H. (2014), "Strong Oracle Optimality of Folded Concave Penalized Estimation," *The Annals of Statistics*, **42**, 819–849.

[30] Foster, D. and George, E. (1994), "The Risk Inflation Criterioin for Multiple Regression," *The Annals of Statistics*, **22**, 1947–1975.

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

[32] Friedman, J., Hastie, T. and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, **33(1)**, 1–22.

[33] Fu, W. J. (1998), "Penalized Regression: The Bridge Versus the LASSO," *Journal of Computational and Graphical Statistics*, **7**, 397–416.

[34] Hoeffding, W. (1963), "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association*, **5**8, 13–30.

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

[36] Hosmer, D. W. and Lemeshow, S. (1999), "Applied Survival Analysis: Regression Modeling of Time to Event Data," *John Wiley & Sons Inc., New York, NY*.

[37] 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.

[38] Hunter, D. R. and Li, R. (2005), "Variable Selection Using MM Algorithms," *The Annals of Statistics*, **27**, 1491–1518.

[39] Kim, Y., Choi, H. and Oh, H. S. (2008), "Smoothly Clipped Absolute Deviation on High Dimensions," *Journal of the American Statistical Association*, **103**, 1665C1673.

[40] Konishi, S. and Kitagawa, G. (1996), "Generalized Information Criteria in Model Selection," *Biometrika*, **83**, 875–890.

[41] Li, H. and Luan, Y. (2005), "Boosting Proportional Hazards Models Using Smoothing Spline, With Application to High-dimensional Microarray Data," *Bioinformatics* **21**, 2403–2409.

[42] Li, K. C. (1987), "Asymptotic Optimality for C_p, C_l, Cross-validation and gGneralized cross validation: Discrete Index Set," *The Annals of Statistics*, **15**, 958–975.

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

[44] Li, R., Zhong, W. and Zhu, L. (2012), "Feature Screening via Distance Correlation Learning," *Journal of American Statistical Association*, **107**, 1129–1139.

[45] Lindley, D. V. (1968), "The Choice of Variables in Multiple Regression," *Journal of the Royal Statistical Society, Series B* **30**, 31–66.

[46] Liu, J., Li, R. and Wu, R. (2014), "Feature Selection for Varying Coefficient Models with Ultrahigh Dimensional Covariates," *Journal of American Statistical Association*.

[47] Loh, P.-L. and Wainwright, M. J.(2013) "Regularized M-estimators with Nonconvexity: Statistical and Algorithmic Theory for Local Optima", Preprint. 2013 Available at arXiv:1305.2436.

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

[49] McQuarrie, A. D. R. and Tsai, C.-L. (1998), "Regression and Time Series Model Selection (1st ed.)," *Singapore: World Scientific Publishing Co.*.

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

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

[52] Nesterov, Y. (2007), "Gradient Methods for Minimizing Composite Functions,", *preprint.*

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

[54] 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.

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

[56] Shao, J. (1997), "An Asymptotic Theory for Linear Model Selection," *Statistica Sinica,* **7**, 221–264.

[57] Shibata, R. (1981), "An Optimal Selection of Regression Variables," *Biometrika*, **68**, 45–54.

[58] Shibata, R. (1984), "Approximation Efficiency of a Selection Procedure for the Number of Regression Variables," *Biometrika*, **71**, 43–49.

[59] Takemi, Y. and Toshinari K. (1984), " Maximum Full and Partial Likelihood Estimators in the Proportional Hazard Model," *Annals of the Institute of Statistical Mathematics*, **36**, 363–373.

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

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

[62] Tibshirani, R. (2009), "Univariate shrinkage in the Cox model for high dimensional data," *Statistical Applications in Genetics and Molecular Biology*, **8**, 1–18.

[63] Tsiatis, A. A. (1981), "A Large Sample Study of Cox's Regression Model," *The Annals of Statistics*, **9**, 93–108.

[64] 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.

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

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

[67] Wang, L., Kim, Y. and Li, R. (2013), "Calibrating nonconvex penalized regression in ultra-high dimension," *The Annals of Statistics*, **41**, 2505–2536.

[68] Wang, Z., Liu, H. and Zhang, T. (2014), "Optimal Computational and Statistical Rates of Convergence for Sparse Nonconvex Learning Problems," *The Annals of Statistics*, **42**, 2164–2201.

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

[70] Yang, Y. (2005), "Can the Strengths of AIC and BIC Be Shared? A Conflict Between Model Identification and Regression Estimation," *Biometrika,* **92**, 937–950.

[71] Yuille, A. and Rangarajan, A. (2003) "The Concave-Convex Procedure (CCCP)," *Neural Computation,* **15**, 915–936.

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

[73] Zhang, H. and Lu, W. (2007), "Adaptive Lasso for Coxs Proportional Hazards Model," *Biometrika,* **1**, 1–13.

[74] Zhang, C.-H. and Zhang, T. (2012), "A General Theory of Concave Regularization for Highdimensional Sparse Estimation Problems," *Statistical Science*, **27**, 576–593.

[75] Zhang, Y., Li, R. and Tsai, C.-L. (2010), "Regularization Parameter Selections via Generalized Information Criterion," *Journal of American Statistical Association*, **105**, 312–323.

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

[77] 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.

[78] Zhu, L, Li, L., Li, R. and Zhu, L.-X. (2011), "Model-free Feature Screening for Ultrahigh Dimensional Data," *Journal of American Statistical Association*, **106**, 1464–1475.

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

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

[81] Zou, H. and Li, R. (2008), "One-step Sparse Estimates in Nonconcave Penalized Likelihood Models"(with discussion), *The Annals of Statistics*, **36**, 1509–1566.

Ye Yu

Curriculum Vitae

326 Thomas Building
Penn State University
University Park, PA 16802-2111
(814) 753 2859
yealfish@gmail.com
Ye(Alex) YU

Seek Truth From Facts

Education

2010–present **Ph.D on Statistics**, *Penn State University (PSU)* , USA, *3.97/4.00*.
M.S degree in December, 2014.
Ph.D degree expected in December, 2015.

2006–2010 **B.E on Mechanical Engineering**, *Tianjin University (TJU)*, P.R China, *3.93/4.00*.
Major degree.

B.S on Finance, *Nankai University (NKU)*, P.R China, *3.7/4.00*.
Dual degree.

Publication and Manuscripts

1. Yang*, G., **Yu*, Y.**, Li, R., and Buu, A. (2014). Feature Screening in Ultrahigh Dimensional Cox's Model. Technical Report No. 14-129. The Methodology Center, The Pennsylvania State University, University Park. Revised for *Statistica Sinica*.

   Authors with $\star$ both are first author.

2. Li, R., Ren J., and **Yu, Y**. (2015). Asymptotic Behavior of Cox's Partial Likelihood and its Application to Variable Selection. To be submitted to *Biometrika*.

   Authors equally contribute to this paper. The order of authors are listed in alphabetic order.

Computer Skills

R, SAS, MATLAB, C++

Experience

2015 summer **Summer Intern**, *Model Development(CMoR)*, Wells Fargo, NC(Supervisor: Jie Chen).
2014 summer **Summer Intern**, *Biostatistics(PDBB)*, Genentech Inc., CA(Supervisor: James Lymp).
2010-present **Research Assistant**, *Methodology Center*, Penn State University(Supervisor: Runze Li).
2012-2013 **Graduate Statistical Consultant**, *Statistical Consulting Center*, Penn State University.
2012.10 **Group Leader**, *Data Mining Projects*, Penn State University.

Honors/Awards

2010-2011 **University Graduate Fellowship**, *Graduate School of Penn State University*.
2006-2009 **National Scholarship**, *Department of Education, P.R China*.