2014-01-01

Variable selection for Cox Proportional Hazards Models via Subtle Uprooting

Chalani S. Wijayasinghe
University of Texas at El Paso, lswijayasinghe@miners.utep.edu

Variable Selection for Cox Proportional Hazards Models via Subtle Uprooting

CHALANI WIJAYASINGHE

Department of Mathematical Sciences

APPROVED:

Xiaogang Su, Chair, Ph.D.

Joan G. Staniswalis, Ph.D.

Jose H. Ablanedo Rosas, Ph.D.

Feng Yang, Ph.D.

Charles Ambler, Ph.D.
Dean of the Graduate School

to my

Family

with LOADS of LOVE

Variable Selection for Cox Proportional Hazards Models via Subtle Uprooting

by

CHALANI WIJAYASINGHE

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Mathematical Sciences

THE UNIVERSITY OF TEXAS AT EL PASO

August 2014

Acknowledgements

I would like to express my deep-felt gratitude to my advisor, Dr. Xiaogang Su, for his advice, encouragement, enduring patience and constant support. He was never ceasing in his belief in me (though I was often doubting in my own abilities), always providing clear explanations when I was struggling, constantly driving me with energy.

My eternal gratitude goes to our graduate coordinator, Dr.Staniswalis for mentoring, encouraging and specially showing me the the correct path whenever I felt I am lost.

I would like to express my gratitude to the members of my committee, Dr.Staniswalis, Dr. Ablanedo and Dr. Yang. Their suggestions, comments and additional guidance were invaluable to the completion of this work.

A special thanks also goes to all the Professors and staff at the University of Texas at El Paso Mathematical Sciences Department for their dedication , support and guidance which has helped me to complete my degree and shaped me into who I am today. I should mention and offer my sincere gratitude to all the staff members for being so friendly and supportive to me.

I am heartily grateful to my dear husband for doing that much sacrifices for the goodness of my carrier with loving support.

Last but not the least, I want to thank my family and friends who appreciated me for my work and motivated me towards the success.

Abstract

Cox proportional hazards model (Cox PH model) is heavily used in survival analysis to assess the importance of various covariates on the survival times of individuals or objects through the hazard function. This study suggests a new variables selection method for Cox PH models, under the title 'Subtle uprooting', that does variable selection and model estimation for Cox proportional hazards (PH) models simultaneously.

There are subset selection methods and shrinkage selection methods suggested in the context of Cox PH model. However the subset selection methods become infeasible in higher dimensions and the available shrinkage methods need tuning of parameters making the approach expensive and time consuming. Most attractive feature of the suggested method against available methods is that it does not require tuning of parameters anymore.

Subtle uprooting uses hyperbolic tangent function as the penalty function based on its appropriate properties such as being a unit dent function, convenience of deriving derivatives and close relationship with the logit function. The procedure include three steps. First, it approximates the cardilaity using surrogate penalty function. Then uprooting and ϵ thresholding are used to enhance the shrinkage.

In the simulation, subtle uprooting, best subset selection, SCAD, LASSO and adaptive LASSO methods are used for simulated data sets and comparisons are made between the methods based on the model error (ME), overfiting, underfitting and correct selection percentages. Furthermore performance of the methods are studied for different sample sizes, censoring rates and input signals (strong and weak). It is found that subtle uprooting outperforms SCAD, LASSO and adaptive LASSO methods under strong signals and higher sample sizes.

Subtle uprooting, best subset selection, SCAD, LASSO and adaptive LASSO methods are applied to the PBC data set. It is found that subtle uprooting estimates are significantly closer to the best subset selection results.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this research, we put forward a new method, coined as 'subtle uprooting', that does variable selection and model estimation for Cox proportional hazards (PH) models simultaneously by approximating information criteria with smooth surrogate functions. The usages and advantages of the proposed method are demonstrated by comparing with other available methods with extensive simulation studies.

1.1 Background

Censored time-to-event data commonly arise in various application fields, including biomedical research, engineering, education, and economics. Cox proportional hazards model is the most popular approach for modeling event time data and quantifying the relationship between the time to event and a set of predictors or covariates. Several reasons account for its popularity: First of all, proportional hazards formulate the covariate effect as multiplicative in terms of the hazard rate, which is found as a good approximation to many phenomena in real application. Secondly, Cox models are flexible in many ways. They are semiparametric since the model contains the baseline hazard rate which is nonparametric and the covariate effects in parametric form. Various extensions of Cox models are available for incorporating time-dependent covariates, time-varying coefficients, stratification, dependent data, etc. Most importantly, the inference of Cox models can be conveniently and efficiently made via partial likelihood. Implementations of Cox models are widely available for practitioners.

1.1.1 Variable Selection in Cox PH model

Variable selection refers to the process of selecting a subset of important and predictive predictors out of all available predictors. These methods are heavily used to extract important information from the data set. This has become a major focus of research in recent years, motivated by application areas which deal with high dimensional data. When there are large number of variables available, researcher has to decide which variables should be included in the model. This is because model misspecification may lead to inaccurate and/or imprecise inference.

Consider the cases of underfitting and overfitting for example. If a smaller number of variables are selected or some important predictors are erroneously omitted form the model, then the model will be overly simplistic. This gives a model that can be easily interpreted. However the resulting model will yield biased estimates. On the other hand, selecting too many variables would result in an unnecessarily complex model that is hard to interpret. This overfitted model will also yield highly variable estimates. In terms of predictive performance, both overfitting and underfitting are suboptimal, referring to the bias-variance trade off.

Variable selection essentially helps to improve the interpretability (simple models), signifies important effects (thus reducing noise), increase the model predictive ability, and speed up modeling time. This has been an important issue to address in the area of statistics. With the advancement of data collection tools and methods, investigators are able to extract more features or covariates from a subject. Therefore, it is important to conduct variable selection.

There are two basic types of variable selection methods, subset selection and shrinkage. Subset selection keeps only a subset of variables while eliminating others from the model by examining a number of model choices. Best subset selection and step wise selection are examples of subset selection methods. There are number of criteria that one may used to compare the models to choose the best one. Since the variables are either retained or discarded, subset selection is a discrete process. Shrinkage selection, also called regularization,

facilitates a continuous variable selection process. Most of the shrinkage methods select variables and estimate parameters simultaneously by solving a constrained optimization problem. Here the coefficients are shrunk towards zero by imposing some penalty function. The least absolute shrinkage and selection operator (LASSO), adaptive LASSO, Smoothly clipped absolute deviance (SCAD) are example of shrinkage methods by producing zero coefficient estimates.

1.2 Significance of the study

Cox proportional hazards models are very important and heavily used and widely applied in survival analysis. While both subset selection and shrinkage selection methods have been proposed and studied in the context of Cox proportional hazards models, each of these methods has its own drawbacks. Variable selection and model estimation are separately done in subset selection. Best subset selection methods examine a large number of models and hence are time and recourse consuming. Although best subset selection often performs well in terms of variable selection, they may not be applicable in dealing with higher dimensional data.

Regularization methods have the attractive feature of conducting variable selection and parameter estimation simultaneously. However, they entail the selection of tuning parameters, for which the researcher have to resort to procedures such as cross validation and generalized cross validation. As a result, the computational burden of these methods could remain high. Also, it is often found that regularization methods may not perform as well as best subset selection in terms of variable selection.

Therefore, it is desirable to explore a new method which borrows strength from both subset selection and regularization. In this thesis, we propose a new variable selection method, titled 'subtle uprooting', for the Cox PH models. The significance of this approach is that it conducts the variable selection and parameter estimation in one optimization step by approximating the best subset selection. This method avoids tuning of the penalty

parameters in traditional shrinkage methods. The procedure results in a nonconvex yet smooth programming problem, which can be solved by a modified BFGS algorithm among other solutions.

1.3 Outline of the thesis

The remainder of the thesis is organized in the following manner. Chapter 2 provides a literature review on available variable selection methods for Cox PH models. Their strengths and weakness are discussed. In Chapter 3, our proposed variable selection method is presented and explained in detail. Chapter 4 presents extensive simulation studies that are designed to evaluate the proposed method and compare it to other competitive methods. An real example based on the analysis of PBC data is also presents. Chapter 5 concludes the thesis with remarks and discussions of further extensions and future work.

Chapter 2

Literature Review

Our aim in this chapter is to present a literature review on available variable selection methods for Cox PH models (Cox, 1972). Let (T_i', C_i') denote the failure and censoring times of the i^{th} individual for $i = 1, ..., n$. Let $T_i = \min(T_i', C_i')$ be the observed failure time and $\delta_i = \mathbb{1}\{T_i' \leq C_i'\}$ indicate the failure status. Let $\mathbf{x}_i \in \mathbb{R}^p$ denote the p-dimensional covariate vector associated with subject i. For identifiability reasons, we assume T_i' and C_i' are independent for given $\mathbf{x}_i$. Thus the observed data consist of $\{(T_i, \delta_i, \mathbf{x}_i) : i = 1, ..., n\}$. Throughout this thesis, our discussion is restricted to the scenarios where covariates are time independent and proportional hazard assumption follows.

A variety of regression models are proposed for quantifying the relationship between failure time with the associated covariates. Cox PH models (Cox, 1972) is the most popular one. Built on the hazard function of the failure time, a Cox model postulates a multiplicative structure that combines a parametric form for the covariate effects with an unspecified underlying hazard function. Hence Cox PH models are labeled as semi-parametric. Specifically, the hazard function $\lambda_i(t|\mathbf{x})$ of subject i conditional on $\mathbf{x}_i$ is specified as

$$\lambda_i(t|\mathbf{x}) = \lambda_0(t) \exp(\boldsymbol{\beta}^T \mathbf{x}_i) \tag{2.1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown regression parameter vector and $\lambda_0(t)$ is the baseline hazard at time t when $\mathbf{x} = \mathbf{0}$.

Estimation of model (2.1) is based on the maximum partial likelihood (Cox, 1975). Let $T_1^0 < T_2^0 < ... < T_K^0$ denote K distinct observed failure times that are sorted in ascending order. Set $T_0^0 \equiv 0$ and $T_{K+1}^0 \equiv \infty$. Assuming no or few ties in the observed times, the

logarithm of the partial likelihood function for $\boldsymbol{\beta}$ is given by,

$$\begin{aligned} L(\boldsymbol{\beta}) = & \log\left\{\prod_{k=1}^{K}\frac{\exp(\mathbf{x}_k^T\boldsymbol{\beta})}{\sum_{i\in R_k}\exp(\mathbf{x}_i^T\boldsymbol{\beta})}\right\} = \sum_{k=1}^{K}\left\{\mathbf{x}_k^T\boldsymbol{\beta} - \log\sum_{i\in R_k}\exp(\mathbf{x}_i^T\boldsymbol{\beta})\right\} \\ & = \sum_{i=1}^{n}\delta_i\left\{\mathbf{x}_i^T\boldsymbol{\beta} - \log\sum_{i'\in R_i}\exp(\mathbf{x}_{i'}^T\boldsymbol{\beta})\right\} \end{aligned} \tag{2.2}$$

where $R_k = \{i : T_i \geq T_k^0\}$ denotes the set of all individuals at risk at time T_k^0, for $k = 1, 2, ..., K$.

Concerning variable selection, the true $\boldsymbol{\beta}$ is often sparse in the sense that some of its components are zeros. We shall review several variable selection methods for fitting Cox PH models. These include the commonly used best subset selection and several newly proposed regularization methods. For the latter, least absolute shrinkage and selection operator (LASSO), adaptive LASSO, and smoothly clipped absolute deviation (SCAD) methods are covered in particular. It is worth noting that many of these variable selection methods is supplemented with parameter estimation, as well as standard error computation for nonzero regression coefficient estimates. We shall outline the main ideas of each method and discuss its strength and weakness.

2.1 Best subset selection

Conventional variable selection techniques which were initially proposed for linear regression models have been adopted for Cox PH models. One such standard method is the best subset selection. In this approach, one finds the best model according to some model selection criterion, by examining all possible subsets of the p covariates. Commonly used criteria are Akaike information criterion (AIC) (Akaike, 1974) or Bayesian information criterion (BIC) (Schwarz, 1978), which can be generally formulated as

$$\min_{\boldsymbol{\beta}} \quad -2L(\boldsymbol{\beta}) + \lambda_0\|\boldsymbol{\beta}\|_0 \tag{2.3}$$

where penalty parameter λ_0 is fixed as 2 in AIC and $\log(n)$ in BIC and $\|\boldsymbol{\beta}\|_0 = \sum_{j=1}^{p}\mathbb{1}(\beta_j \neq 0)$ is known as the cardinality or the number of nonzero coefficients. In efforts of extending

BIC to censored data, Volinsky and Raftery (Volinsky and Raftery, 2000) recommended to use $\lambda_0 = \log(K)$ instead of $\log(n)$, where n is the total number of observations and K is the number of uncensored ones.

Owing to the discrete nature, solving (2.3) involves two major steps: first fit every model choice with known sparsity for $\boldsymbol{\beta}$ and then calculate information criteria and compare across all model choices. So the resultant information criteria will be,

$$-2L(\hat{\boldsymbol{\beta}}^{PL}) + \lambda_0 \|\hat{\boldsymbol{\beta}}^{PL}\|_0 \tag{2.4}$$

where $\hat{\boldsymbol{\beta}}^{PL}$ denotes the maximum partial likelihood estimates of $\boldsymbol{\beta}$ and $L(\hat{\boldsymbol{\beta}}^{PL})$ denotes the maximized log partial likelihood. Faster combinatorial optimization algorithms such as the branch and bound method (Furnival and Wilson, 1974) are available to reduce the amount of computation by finding the best subset without examining all possible subsets.

By retaining a subset of the predictors and discarding the rest, best subset selection produces an interpretable model. In practice, best subset selection often performs well in terms of variable selection when available. However, as a result of being a discrete optimization problem, best subset selection becomes infeasible in high dimension (i.e., large p). Furthermore since the procedure is discrete, selected model can be unstable, having high variance and reduced predictive accuracy. This gives rise to various regularization approaches.

2.2 Shrinkage methods

The regularization or shrinkage methods essentially takes the penalized partial likelihood approach. The general form of shrinkage methods can be given by

$$\min_{\boldsymbol{\beta}} \quad -L(\boldsymbol{\beta}) + \sum_{j}^{p} p_\lambda(\beta_j), \tag{2.5}$$

where $p_\lambda(.)$ denotes a penalty function with regularization parameter $\lambda > 0$.

Several penalty functions have been suggested in the literature. Ridge regression (Hoerl and Kennard, 1970) uses L_2 regularization with the penalty function $p_{\lambda_r}(\beta_j) = \lambda_r \beta_j^2$. In

this case, the resulting estimates are shrunken towards zero as λ_r increases. Ridge regression helps with ill-posed problems that stem from multicolinearity and other numerical difficulties. The objective function remains smooth, which is mathematically appealing. Stable numerical algorithms are available for solving L_2 regularization; in some scenarios, explicit solutions in closed form are available. However, it does not come up with zero estimates and hence the resulting model is not easily interpretable.

Least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1997) is a variable selection method which uses L_1 penalty. Initially LASSO was proposed in the context of linear regression (Tibshirani, 1996) and later extended to Cox PH models. Here an estimate of $\boldsymbol{\beta}$ is obtained through solving the following convex optimization problem

$$\min_{\boldsymbol{\beta}} \ -L(\boldsymbol{\beta}) + \lambda_l \sum |\beta_j|, \tag{2.6}$$

or, equivalently,

$$\max_{\boldsymbol{\beta}} L(\boldsymbol{\beta}), \text{ subject} \quad \text{to } \sum |\beta_j| \leq s, \tag{2.7}$$

where $s > 0$ is a constraint parameter which can be tuned. Compared to L_2 ridge penalty, the penalty function $\sum_j^p \beta_j^2$ is now replaced by $\sum_j^p |\beta_j|$ in LASSO. If $s \geq \sum |\hat{\boldsymbol{\beta}}^{PL}|$ where $\hat{\boldsymbol{\beta}}^{PL}$ is the maximum partial likelihood estimates, LASSO estimates are the same as maximum partial likelihood estimates, otherwise solutions are shrunken towards zero. As s decreases, variables will be systematically removed from the model. The tuning parameter λ_l in equation (2.6) controls the amount of regularization.

The LASSO method can produce some exactly zero-valued coefficient estimates. Consider the special case with an orthogonal design where $\mathbf{x}^T\mathbf{x} = I$. It can be shown that LASSO estimators will be,

$$\hat{\beta}_j^{LASSO} = \text{ sign } (\hat{\beta}_j^{PL})\{|\hat{\beta}_j^{PL}| - \gamma\}_+, \tag{2.8}$$

where γ is determined by the condition $\sum |\beta_j^{LASSO}| = s$. All the coefficients are shrunk by unit of γ towards zero. If the coefficient value is less than γ in absolute value, then it will be set to zero, which enforces variable selection.

More insights into the difference between L_2 and L_1 regularization can be gained via a quadratic approximation of L, the logarithm of partial likelihood. Expanding $L(\boldsymbol{\beta})$ at $\hat{\boldsymbol{\beta}}^{PL}$ gives,

$$\text{Ł}(\boldsymbol{\beta}) \approx L(\hat{\boldsymbol{\beta}}^{PL}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{PL})\nabla\ L(\hat{\boldsymbol{\beta}}^{PL}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{PL})^T \left\{\nabla^2 L(\hat{\boldsymbol{\beta}}^{PL})/2\right\} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{PL}) \quad (2.9)$$

Noting $\nabla\ L(\hat{\boldsymbol{\beta}}^{PL}) = 0$ and ignoring the irrelevant term $L(\hat{\boldsymbol{\beta}}^{PL})$, the optimization problem for Ridge and LASSO methods can then be written as,

$$\min_{\boldsymbol{\beta}}\ (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{PL})^T \left\{-\nabla^2 L(\hat{\boldsymbol{\beta}}^{PL})/2\right\} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{PL}) \text{ subject to } \|\boldsymbol{\beta}\|_2 \leq r, \quad (2.10)$$

$$\min_{\boldsymbol{\beta}}\ (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{PL})^T \left\{-\nabla^2 L(\hat{\boldsymbol{\beta}}^{PL})/2\right\} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{PL}) \text{ subject to } \|\boldsymbol{\beta}\| \leq s, \quad (2.11)$$

where $\|\boldsymbol{\beta}\|_\gamma = \sum_{i=1}^{p} |\boldsymbol{\beta}_i|^\gamma$ and r, and s are the constraint parameters for Ridge, and LASSO respectively. In the two dimensional case, the contour of the objective function are ellipsoids centered at the maximum partial likelihood estimate $\hat{\boldsymbol{\beta}}^{PL}$. The constraint region for Ridge regression is a disk, $\beta_1^2 + \beta_2^2 = r$, and in LASSO it becomes a quadrilateral, $|\beta_1| + |\beta_2| = s$. In either method, the solution is the first point where the ellipsoid contour lines intersect with the constrained region as the value of the quadratic form decreases. Unlike in the circle, the quadrilateral has corner points on the axes and hence it is more likely to have the solution at a corner. According to Figure 2.1 (b), LASSO estimates β_1 as zero. Therefore, LASSO conducts parameter estimation and variable selection simultaneously.

To solve (2.6) and (2.7), Tibshirani (1997) proposed an algorithm that integrates methods for solving LASSO into the iteratively reweighted least squares (IRWLS) procedure. In order to estimate the constraint parameter s, generalized cross validation (GCV) was used. The value, which minimizes GCV statistic was taken as the estimated value for s. Many algorithms are available for finding LASSO solutions. Among others, the Least angle regression (LAR) (Efron et al., 2004) and coordinate descent (CD) (Kim et al., 2008) have gradually become dominant. The LAR method solves the optimization problem with an homotopy algorithm which is similar to forward stagewise method. At each step, the estimates of parameters are updated based on the correlations with the residual. CD is another

algorithm for solving LASSO, especially when p is large. CD solves a multi-parameter optimization problem by optimizing it with respect to one parameter at a time. Recall that explicit solution of LASSO is available with the orthogonal design. It can be easily seen that the threshold formed solution in (2.8) holds true for one parameter optimization as well. Several variants of CD algorithms are available. One appealing way of speeding up CD is combining it with gradient descent. This approach optimizes (2.6) directly from a starting value of $\boldsymbol{\beta}$ by cycling through the coordinates of $\boldsymbol{\beta}$ and sequentially updating them on the basis of the gradient of the penalized likelihood. Gradient based optimization was discussed in the context of Cox PH models in (Sohn et al., 2009).

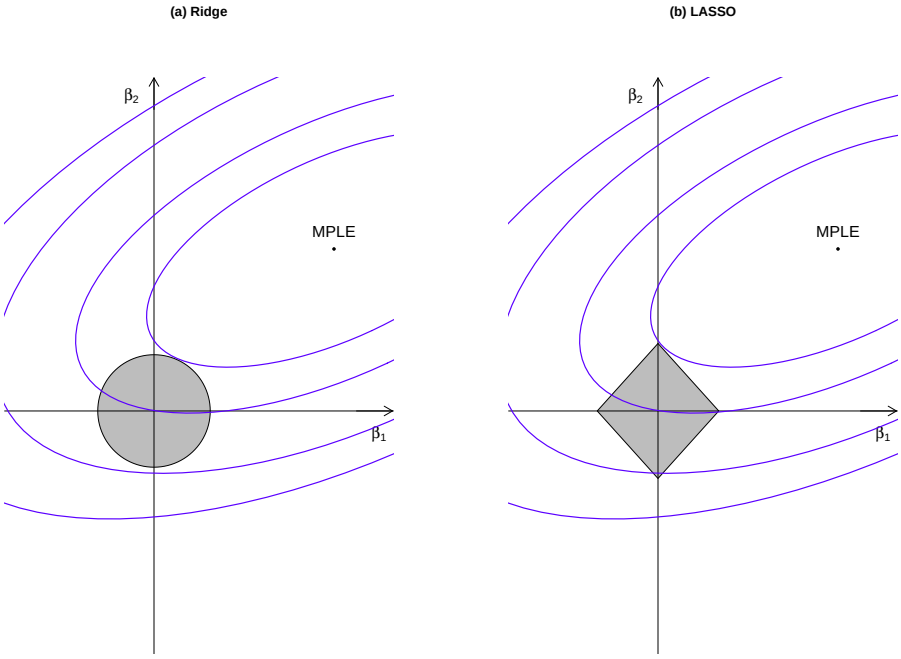

Figure 2.1: Illustration of Ridge and LASSO Estimators in the two dimensional case

Since the first proposal of LASSO, it has generated wide popularity and tremendous interest in statistical literature and other related fields. Despite its success, its performance in variable selection is often unsatisfactory. Theoretical studies shows that LASSO

is unlikely to achieve consistent variable selection unless a strong irrepresentable condition holds (Meinshausen and Bhlmann, 2006). Furthermore LASSO shrinkage produces biased estimates for the large coefficients, and thus it could be also suboptimal in terms of estimation.

To amend these problems, several proposals have been made. Adaptive LASSO (ALASSO) (Zhang and Lu, 2007) is another L_1 regularization method that improves LASSO so that it meets desired theoretical properties. Unlike the LASSO which apply the same penalty to all the coefficients, the ALASSO applies different penalties that are data-adaptive. Specifically, it solves

$$\min_{\boldsymbol{\beta}} \ -L(\boldsymbol{\beta}) + \lambda_{al} \sum |\beta_j| \tau_j, \tag{2.12}$$

where positive weights $\tau = (\tau_1, \tau_2, \tau_3, \cdots, \tau_p)^T$ are adaptively chosen based on data. The common choice is $\tau_j = 1/|\hat{\beta}_j^{PL}|$, as studied in Zhang and Lu (2007). Since $\hat{\boldsymbol{\beta}}^{PL}$ are consistent estimators of $\boldsymbol{\beta}$, their values well reflect the relative importance of the covariates. With these weights, the optimization problem in adaptive LASSO can be written as,

$$\min_{\boldsymbol{\beta}} \ -L(\boldsymbol{\beta}) + \lambda_{al} \sum |\beta_j| / |\hat{\beta}_j^{PL}|. \tag{2.13}$$

ALSSO in (2.13) remains a convex optimization problem and can be conveniently solved by its association with LASSO via re-parameterization.

Fan and Li (2001) (Fan and Li, 2002) provided more insights into regularization methods that conduct variable selection and parameter estimation simultaneously. They indicated that a desirable penalty function is expected to help achieving three properties, as listed below:

- Unbiasness in estimating nonzero parameters — the resulting estimators should be nearly unbiased for the parameter to the avoid unnecessary modeling bias;
- Sparsity in terms of enforcing zero estimates — the resulting estimator should be a threshold rule such that it sets small estimated coefficients to zero to reduce model complexity;

- Continuity in terms of model spectrum — resulting estimator is continuous in data to avoid instability in model prediction.

Based on these considerations, they proposed a smoothly clipped absolute deviation (SCAD) method, which improves LASSO by reducing bias in estimating nonzero coefficients and obtaining better variable selection. The SCAD penalty function is a quadratic spline and symmetric around the origin. It is often given in its first derivative form

$$p'_{\lambda_s}(|\boldsymbol{\beta}|) = \lambda_s \left\{ \mathbb{1}(|\boldsymbol{\beta}| \leq \lambda_s) + \frac{(\alpha\lambda_s - |\boldsymbol{\beta}|)_+}{(\alpha - 1)\lambda_s} \mathbb{1}(|\boldsymbol{\beta}| > \lambda_s) \right\} \tag{2.14}$$

which involves two parameters $\alpha > 2$ and $\lambda_s > 0$. Cross validation and generalized cross validation can be used to estimate α and λ_s. However tuning two parameters can be computationally expensive. It is recommended to use $\alpha = 3.7$, based on a Bayesian statistical point of view. The SCAD penalty is non-convex, rending (2.5) a non-convex non-smooth optimization problem. To solve SCAD, Fan and Li (2001) proposed an modified Newton-Raphson based on a local quadratic approximation of the penalty function. Along the similar lines, Zou and Li (2008) proposed the local linear approximation (LLA) and recommended using one-step LLA as the final estimates. In addition, (Wu and Liu, 2009) proposed a difference convex algorithm (DCA) for solving SCAD by decomposing SCAD penalty function as the difference of two convex functions. The use of SCAD in the Cox PH model setting has been studied by Fan and Li (Fan and Li, 2002).

Although LASSO, adaptive LASSO and SCAD methods are capable of conducting variable selection and parameter estimation simultaneously, the formulation of regularization incurs a loss of knowledge in the fixed value of λ. As a result, λ is left as the so-called tuning parameter. Solving regularization also entails two steps: first compute the regularization path $\hat{\boldsymbol{\beta}}(\lambda)$ for every λ and then select the best λ by referring to some selection criterion. In the latter step, the use of information criteria is often advised. Note that the regularization path is one-dimensional, parameterized by λ. Thus, the whole procedure of regularization amounts to search for the best $\boldsymbol{\beta}$ (with minimum information criterion) from the one-dimensional regularization path $\hat{\boldsymbol{\beta}}(\lambda)$ only, as opposed to the p-dimensional

search space considered in best subset selection. As a result, the performance of these regularization may not be expected to outperform best subset selection, when referring to the same information criterion. Besides, the current practice of regularization can be computationally demanding, owing to selection of one or two tuning parameters.

These above-mentioned deficiencies motivate us to consider a new method that borrows strength from both best subset selection and regularization.

Chapter 3

Subtle Uprooting

This chapter entails detailed explanation of the steps involved in the proposed variable selection method. We discussed strengths and weaknesses of the subset selection and regularization methods in the literature review chapter. The regularization approach has attracted intensive research efforts in recent years. A desirable penalty function in regularization is expected to help in achieving unbiasness when estimating nonzero coefficients, sparsity in terms of enforcing zero estimates, and continuity in terms of model spectrum (Fan and Li, 2001). However these regularization methods entail tuning penalty parameters. On the other hand, an information criteria such as AIC or BIC is used in best subset selection and the penalty parameter λ_0 in equation (3.1) is held fixed. However the best subset selection becomes infeasible due to its discrete nature. We propose a new variable selection method, coined "subtle uprooting", which is intended to borrow strengths of both regularization and subset selection.

In its final form, subtle uprooting solves the following optimization problem.

$$\min_{\boldsymbol{\beta}} \quad -2L(\boldsymbol{\beta}') + \lambda_0 \sum_{j=1}^{p} w(\beta_j) \tag{3.1}$$

where

$$\beta_{\epsilon j} = \operatorname{sgn}(\beta_j) \cdot (|\beta_j| - \epsilon)_+ = \begin{cases} \beta_j + \epsilon & \text{if } \beta_j < -\epsilon; \\ 0 & \text{if } |\beta_j| \leq \epsilon; \\ \beta_j - \epsilon & \text{if } \beta_j > \epsilon. \end{cases}$$

$$w_j = w(\beta_{\epsilon j}) = \tanh(a \cdot \beta_{\epsilon j}^2)$$

$$\beta_j' = \beta_j \cdot w_j$$

The procedure involves three parameters, $\epsilon > 0, \lambda_0,$ and $a > 0$. However all the parameters are fixed, hence further tuning is not required. It is recommended to use $\lambda_0 = \log K$, $a = 50,$ and $\epsilon = 10^{-4}$, where K is the number of events. In this formulation, subtle uprooting is intended as an approximation to the BIC criterion, suggested by Volinsky and Raftery (2000).

There are three steps involved in the procedure. First a surrogate function is used to approximate the cardinality. Then, uprooting and ϵ-thresholding are carried out to enhance shrinkage and achieve sparsity while retaining differentiability of the objective function.

3.1 Penalty function for Subtle uprooting

In the first step, the penalty function in equation (2.5) is replaced by a smooth penalty function $w(\cdot)$. A suitable surrogate penalty function must be an unit dent function (Su, 2013) as defined below:

Definition 3.1.1. *A unit dent function is a continuous function that satisfies the following properties,* $w: \mathbb{R} \to [0, 1]$

i. $w(.)$ *is an even function such that* $w(\beta) = w(-\beta)$

ii. $w(0) = 0$ *and* $lim_{|\beta|\to\infty} w(\beta) = 1$

iii. $w(\beta)$ *is increasing on* $\mathbb{R}_+$ *and decreasing on*$\mathbb{R}_-$

The $[0, 1]$ range restriction essentially makes $w(\cdot)$ non-convex. If the surrogate function $w(\beta)$ is differentiable, then its first derivative satisfies $\dot{w}(\beta) \geq 0$ on $\mathbb{R}_+$ and $\dot{w}(\beta) \leq 0$ on $\mathbb{R}_-$. It is preferred to have a smooth unit dent function so that it will be a natural extension of maximum likelihood. Furthermore use of a smooth unit dent function is important since it allows to use well developed theories in optimization and and statistical estimation. Based on these considerations, subtle uprooting recommends the hyperbolic tangent function

$$w(\beta) = \tanh(a\beta^2) = \frac{\exp(2a\beta^2) - 1}{\exp(2a\beta^2) + 1} \tag{3.2}$$

for routine use. Hyperbolic tangent function is a good choice because it is a smooth unit dent function and its derivatives are easy to derive. Furthermore, tanh(·) function has a connection with logistic function which is widely used in statistics.

Figure 3.1 plots the surrogate functions for SCAD method and for $\tanh(a\beta^2)$. It can be observed that both are unit dent functions. According to figure 3.1 (b), $\tanh(a\beta^2)$ is smooth around the neighborhood of zero.

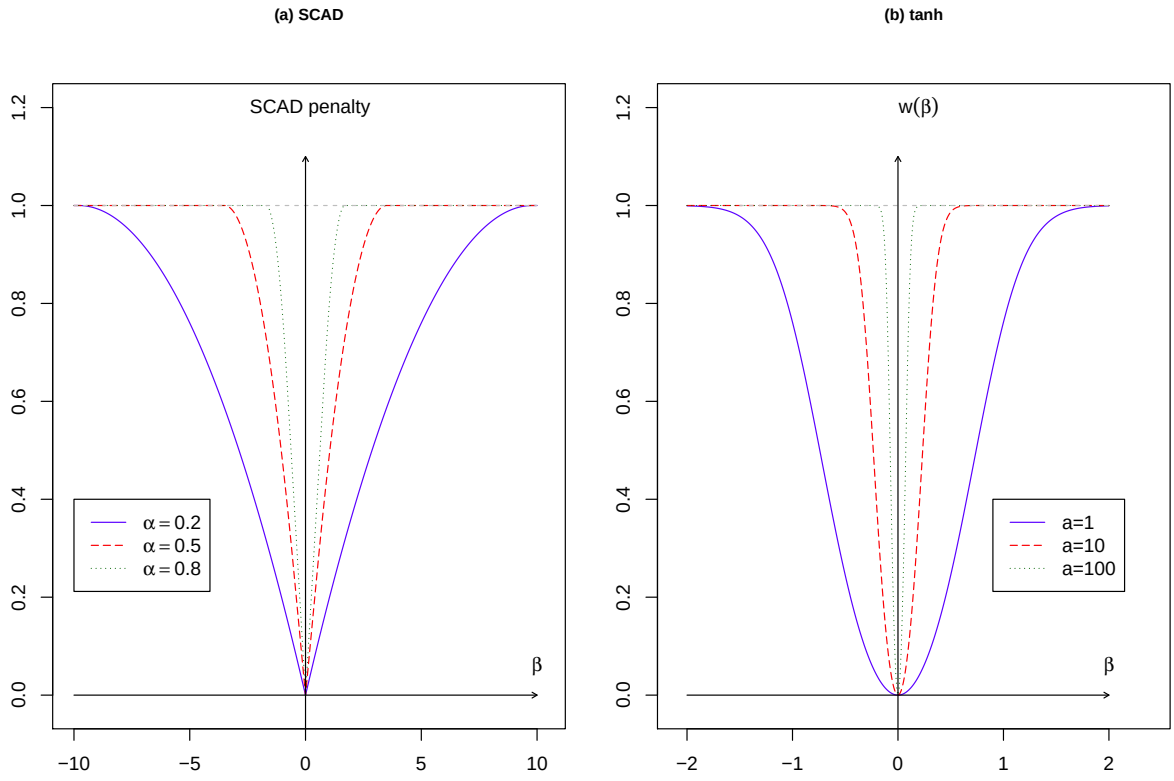

Figure 3.1: plot of surrogate penalty functions for cardinality (a) SCAD, (b) $\tanh(a\beta^2)$

The proposition lists a few properties of the hyperbolic tangent penalty $w(\beta)$.

Proposition 3.1.1. *Let* $\pi(t)$ *be the expit function such that* $\pi(t) = expit(t) = [1 + \exp(-t)]^{-1}$. *Denote* $\pi = \pi(a\beta^2)$ *and* $w = w(\beta) = \tanh(a\beta^2)$. *Then we have,*

(i). $w(\beta) = 2\pi - 1$

(ii). As for its comparison with the weight elimination penalty, $w(\beta) \geq \frac{a\beta^2}{1+a\beta^2}$

(iii). The two derivatives of $w(\beta)$ *are given by*

$$\dot{w}(\beta) = 2a\beta sech^2(a\beta^2) = 2a\beta(1 - w^2) = 8a\beta\pi(1 - \pi)$$

and

$$\ddot{w}(\beta) = 8a\pi(1 - \pi)[1 + (1 - 2\pi)\beta]$$

(iv). By Taylor expansion, $w(\beta) = a\beta^2 + \mathcal{O}(\beta^6)$ *around* 0

Proof. We only show (i). See Su (2013) for other parts. Let $\pi(\cdot)$ be the expit function. Then

$$w(\beta) = \frac{\exp(2a\beta^2) - 1}{\exp(2a\beta^2) + 1} = \frac{\left(\frac{\pi}{1-\pi}\right) - 1}{\left(\frac{\pi}{1-\pi}\right) + 1} = 2\pi - 1.$$

□

From (iii), $w(\beta) \approx \beta^2$ when β is around 0. This implies that this hyperbolic tangent penalty function behaves similar to ridge penalty (the ℓ_2 penalty) which does not enforce the sparsity.

The optimization problem in Equation (3.1) can be equivalently considered as,

$$\min_{\boldsymbol{\beta}} \ (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \left\{-\nabla^2 L(\hat{\boldsymbol{\beta}})/2\right\} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \ \text{ subject to } \sum_{j=1}^{p} w(\beta_j) \leq t_0 \tag{3.3}$$

for some $t_0 \geq 0$. This optimization problem can be visualize for two dimensional case as in figure 3.2. Figure 3.2 (a) shows that $w(\beta_j)$ behave similar to ridge (L_2 penalty) for smaller values of β_j. From the figure it is clear that, by sharpening the edges of diamonds, sparsity can be enforced. Hence the second step of subtle uprooting is used to addresses this issue.

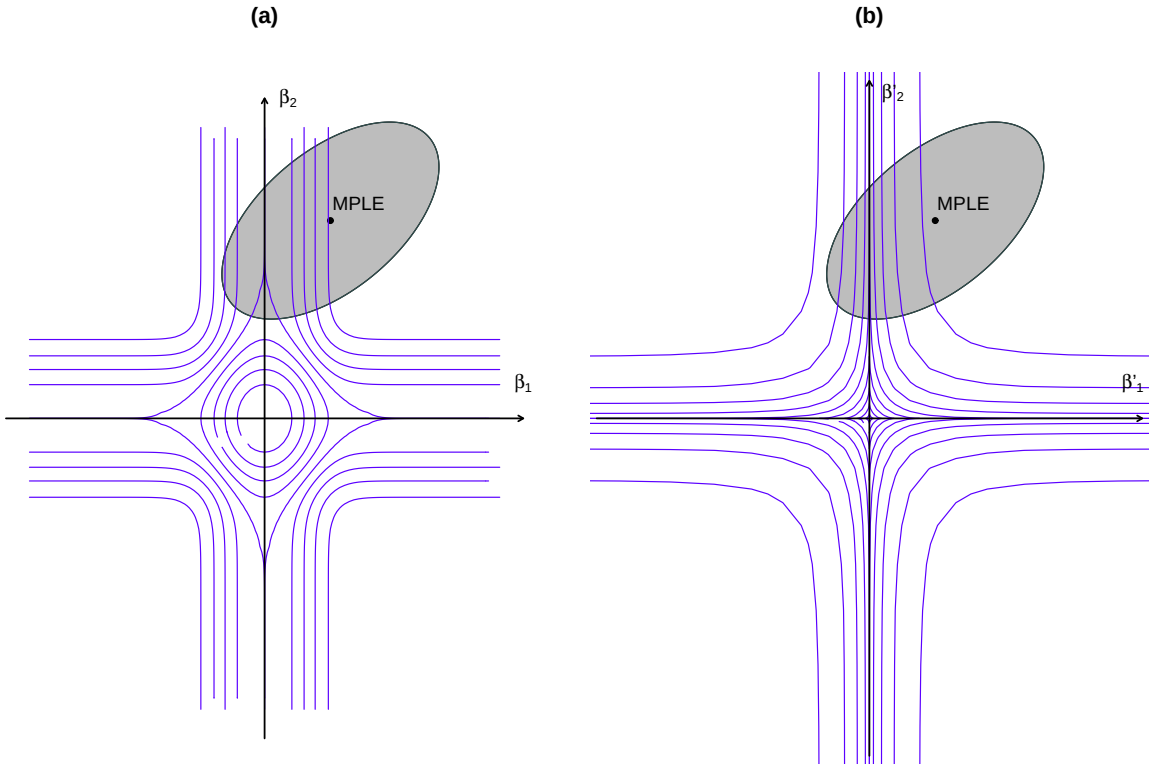

Figure 3.2: Illustration of optimization in two-dimensional case (a) without uprooting (b) with uprooting

3.2 Uprooting

Following Su (2013), we next consider an uprooting step that helps enforce the sparsity. It enhances the shrinkage around the neighborhood of 0 by applying a non convex penalty.

$$\beta = \beta \cdot \mathbb{1}\{\beta \neq 0\} \approx \beta w(\beta) = \beta' \tag{3.4}$$

By reparametrization, optimization problem 3.1 can be obtained. Then the coefficient of x_j becomes ${\beta_j}'$.

$$L(\boldsymbol{\beta}') = \log \prod_{k=1}^{K} \frac{\exp(\mathbf{x}_k^T \boldsymbol{\beta}')}{\sum_{i \in R_k} \exp(\mathbf{x}_i^T \boldsymbol{\beta}')} \tag{3.5}$$

As in Figure 3.3 (a), the relationship between β and β' is one to one except for a small neighborhood of 0. $w : \mathbb{R} \to [0, 1]$ and $\beta' = \beta w(\beta)$ gives $|\beta'| \leq |\beta|$. $w(\beta)$ is a monotonically increasing in $\mathbb{R}_+$. Therefore $w(\beta) \geq w(\beta')$. This follows that a larger penalty is applied for the values in the neighborhood of 0.

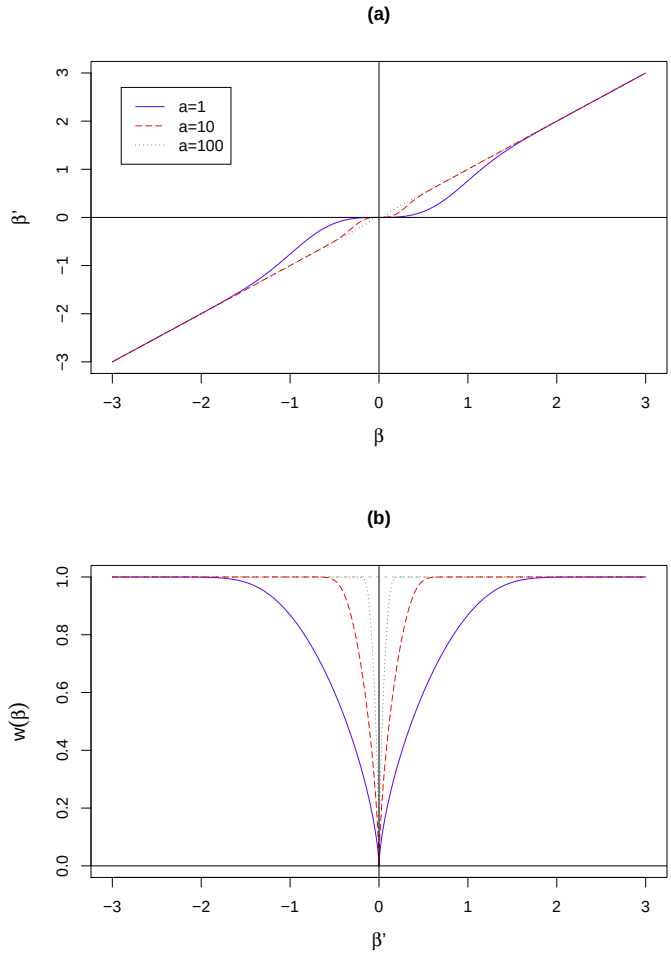

Figure 3.3: Illustration of uprooting step (a) β vs β' (b) $w(\beta)$ vs β'

The following proposition is taken directly from Su (2013).

Proposition 3.2.1. *Let $\mathbb{D}$ denote the space of all unit dent functions. Given a smooth function $w(\beta) \in \mathbb{D}$, denote $w = w(\beta)$, $\dot{w} = \dot{w}(\beta)$, $\ddot{w} = \ddot{w}(\beta)$. Let $\beta' = \beta w(\beta)$. As a function of β', $w(\beta) \in \mathbb{D}$ is smooth everywhere except at $\beta' = 0$. In particular,*

i. $$\frac{dw(\beta)}{d\beta'} = \frac{\dot{w}}{w + \beta\dot{w}}$$

ii. $$\frac{d^2w(\beta)}{d\beta'^2} = \frac{w\ddot{w} - 2\dot{w}^2}{(w + \beta\dot{w})^3}$$

The first two derivatives do not exist at $\beta' \neq 0$

Based on proposition 3.1 and 3.2 it can be shown that the uprooting drastically intensifies the change rate of the penalty function around 0. This property allows establishing the asymptotic consistency in variable selection.

3.3 ϵ-Threshold

Uprooting shrinks the estimates to very small values which can be virtually taken as 0. The estimates are unlikely to be exactly zero without rounding. The ϵ-threshold method is an optional step that helps with this rounding problem. It assigns a non-zero probability to get exactly zero estimates. In the procedure, a threshold is first applied on the estimate to obtain β_ϵ, and then apply $w(\cdot)$ for β_ϵ. Specifically,

$$\beta_{\epsilon j} = \text{sgn}(\beta_j) \cdot (|\beta_j| - \epsilon)_+ = \begin{cases} \beta_j + \epsilon & \text{if } \beta_j < -\epsilon; \\ 0 & \text{if } |\beta_j| \leq \epsilon; \\ \beta_j - \epsilon & \text{if } \beta_j > \epsilon. \end{cases}$$

Now the objective function becomes

$$-2L(\mathbf{W}\boldsymbol{\beta}') + \lambda_0 \text{tr}(\mathbf{W}) \tag{3.6}$$

in matrix form, where $\mathbf{W} = \text{diag}(w(\beta_{\epsilon j}))$ is a $p \times p$ diagonal matrix and $\text{tr}(\cdot)$ denotes the trace of a matrix.

Figure 3.4 shows the behavior of the penalty as a function of β with addition of thresholding. The gray rectangle in figure 3.4 highlights the effect of thresholding. It sets the penalty function between $(-\epsilon, \epsilon)$ to 0, enforcing sparsity.

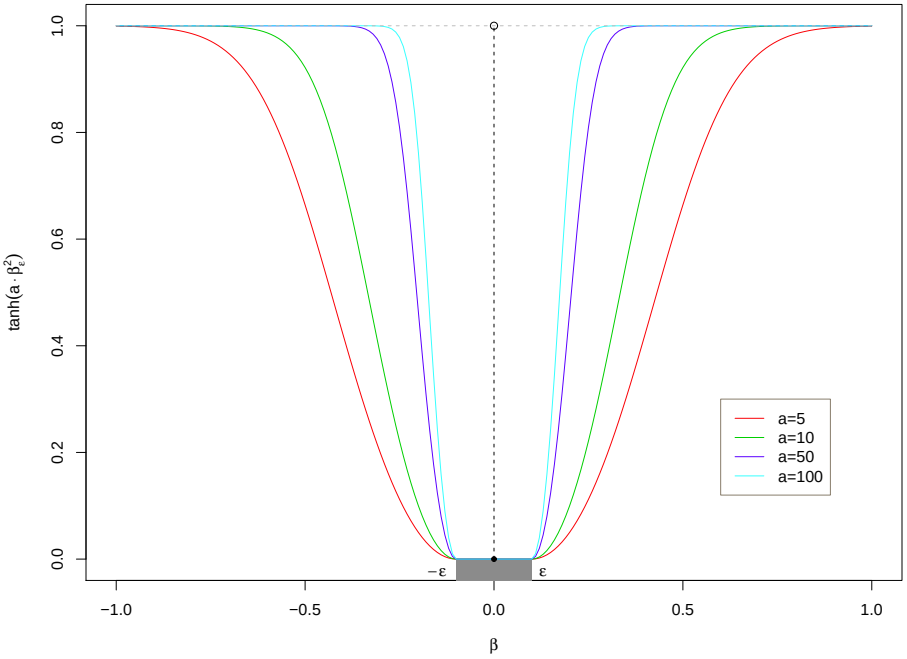

Figure 3.4: Penalty function for Subtle uprooting with the ϵ

Chapter 4

Results

This chapter presents extensive simulation studies that are designed to evaluate the proposed method and compare it to other competitive methods. Different simulation settings are used to investigate the performance of each method in different circumstances.

4.1 Simulation Setting

In the simulation, data sets are generated with known parameter values. Then variable selection and estimation for the Cox PH models is conducted using subtle uprooting, best subset selection, SCAD, LASSO and adaptive LASSO methods. Actual and estimated values are compared to assess the performance of each method.

For subtle uprooting method, value of a is taken as 50 as recommended in chapter 3. BFGS quasi-Newton method (Broyden, 1970) is a well known method used for solving nonlinear optimization problems. Later BFGS algorithm was extended to handle nonconvex optimization problems (Li and Fukushima, 2001). This algorithm is slightly modified and used to solve the nonconvex optimization in subtle uprooting method. BIC information criteria is used in the best subset selection. Here all possible model choices are considered and the model that produces the smallest BIC value is selected. The `SIS` package in R is used in order to fit the SCAD method and BIC is used for selecting the tuning parameter. For LASSO and adaptive LASSO methods, the `glmnet` package in R is used, where the 10-fold cross validation is used to tune the constraint parameters. First it finds a lambda sequence by running nfold+1 times, and then compute the fit with each of the folds omitted. The error is accumulated, and the average error and standard deviation over the folds is

computed. The λ value that yields minimum mean cross validated error is used for LASSO and adaptive LASSO.

Measures for assessing the performance of each method include model error (ME), model size and underfitting, overfitting, correct selection proportions, all being calculated and recorded for each method under 300 runs. Here ME is calculated as $(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T S(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$ where S is the sample variance covariance matrix. The model size is the number of nonzero coefficient estimates in a model. The averaged model size is reported. Underfitting occurs if at least one coefficient is estimated as zero when its true value is nonzero. A model is considered as an overfit if the model is not an underfit and at least one zero coefficient is estimated as nonzero. If the model selects the variables correctly (as the true model) then it is considered as a correct selection. Hence overfitting, underfitting, and correct selection proportions add up to 1.

To conveniently compare with other methods, we employed the similar simulation setting as in (Zhang and Lu, 2007). Covariates $z_j, j = 1, 2, ..., 9$ are generated such that each z_j is marginally normal and the correlation between z_i and $x_{j'}$ is $\rho^{|j-j'|}$ where ρ is taken as 0.5 for first four settings. The actual parameter values are taken as $\boldsymbol{\beta} = (-0.7, -0.7, 0, 0, 0, -0.7, 0, 0, 0)^T$ and $(-0.4, -0.3, 0, 0, 0, -0.2, 0, 0, 0)^T$, which correspond to scenarios with strong and weak signals, respectively. Two sets of survival times $(C_0\ , T_0)$ are generated based on exponential distribution with rate $\exp(\boldsymbol{\beta}^T\mathbf{x})$. A parameter, censoring control, is used to control the censoring rate. The status is taken as 1 (as an event), if the $T_0 \leqslant C_0 *$ censoring control and 0 otherwise. Through the empirical studies it is found that censoring control =1.5, 3 values yield empirical censoring rates of 40% and 25% respectively. Three sample sizes, $n = 100,\ 200,$ and 300 are used in the simulation study. To sum up, the following four simulation settings are used.

Table 4.1: Simulation settings

setting	$\boldsymbol{\beta}$	emphirical censoring rate
setting 1	$(-0.7, -0.7, 0, 0, 0, -0.7, 0, 0, 0)^T$	40%
setting 2	$(-0.4, -0.3, 0, 0, 0, -0.2, 0, 0, 0)^T$	40%
setting 3	$(-0.7, -0.7, 0, 0, 0, -0.7, 0, 0, 0)^T$	25%
setting 4	$(-0.4, -0.3, 0, 0, 0, -0.2, 0, 0, 0)^T$	25%

4.2 Simulation Results

Table 4.2 summarizes mean values of the model error, size, underfitting, overfitting and correct selection proportions in the presence of a strong signal and higher censoring rate (40%), over 300 runs.

Table 4.2: Setting 1: Model Statistics

nrun=300, strong signal, censoring rate ≈ 40%						
n	Method	ME	SIZE	Under	Over	correct selection
100	SUBTLE	0.1763	3.6133	0.0500	0.3900	0.5600
	Best Subset	0.1592*	3.3733	0.0600	0.2867	0.6533*
	SCAD	0.1687	3.3867	0.1167	0.2867	0.5967
	LASSO	0.1877	5.6133	0.0033	0.9467	0.0500
	ALASSO	0.1858	5.5467	0.0033	0.9467	0.0500
200	SUBTLE	0.0565	3.2600	0.0033	0.2200	0.7767
	Best Subset	0.0534*	3.1700	0.0033	0.1467	0.8500*
	SCAD	0.0642	3.7833	0.0000	0.4033	0.5967
	LASSO	0.0863	5.6533	0.0000	0.9633	0.0367
	ALASSO	0.0901	5.5733	0.0000	0.9467	0.0533
300	SUBTLE	0.0337	3.3100	0.0000	0.2633	0.7367
	Best Subset	0.0332*	3.1667	0.0000	0.1400	0.8600*
	SCAD	0.0408	3.8400	0.0000	0.4100	0.5900
	LASSO	0.0540	5.6133	0.0000	0.9367	0.0633
	ALASSO	0.0527	5.6133	0.0000	0.9367	0.0633

* minimum value for ME and maximum correct proportion within each sample size

According to Table 4.2, model error for all the methods decreases as the sample size increases. Subtle method, best subset selection and SCAD method perform better than LASSO and adaptive LASSO methods both in variable selection and estimation aspects. Subtle method outperforms SCAD method with larger sample sizes. Furthermore, it can be observed that the subtle uprooting method resembles best subset selection with larger sample sizes.

Table 4.3: Setting 1: Median of the estimates

nrun=300, strong signal, censoring rate ≈ 40%										
n	Method	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$
100	SUBTLE	-0.7413	-0.7361	0	0	0	-0.7313	0	0	0
	Best Subset	-0.7344	-0.7279	0	0	0	-0.7278	0	0	0
	SCAD	-0.6951	-0.7036	0	0	0	-0.6643	0	0	0
	LASSO	-0.5976	-0.5927	0	0	0	-0.5237	0	0	0
	ALASSO	-0.5938	-0.5825	0	0	0	-0.5206	0	0	0
200	SUBTLE	-0.7148	-0.7078	0	0	0	-0.7060	0	0	0
	Best Subset	-0.7101	-0.7088	0	0	0	-0.7016	0	0	0
	SCAD	-0.7086	-0.7056	0	0	0	-0.7038	0	0	0
	LASSO	-0.6245	-0.6108	0	0	0	-0.5928	0	0	0
	ALASSO	-0.6258	-0.6144	0	0	0	-0.5920	0	0	0
300	SUBTLE	-0.7145	-0.7160	0	0	0	-0.7105	0	0	0
	Best Subset	-0.7107	-0.7161	0	0	0	-0.7105	0	0	0
	SCAD	-0.7152	-0.7141	0	0	0	-0.7079	0	0	0
	LASSO	-0.6435	-0.6415	0	0	0	-0.6101	0	0	0
	ALASSO	-0.6442	-0.6456	0	0	0	-0.6151	0	0	0

Table 4.3 shows the median values for parameter estimates. Note that the true parameter values for this setting is $\boldsymbol{\beta} = (-0.7, -0.7, 0, 0, 0, -0.7, 0, 0, 0)^T$. Thus only $\beta_1,\ \beta_2,$ and β_6 are nonzero. All five methods give zero median value for deselecting those covariates with zero true coefficients. Subtle uprooting, best subset selection and SCAD perform better than LASSO and adaptive LASSO in estimating the nonzero coefficients by producing median values closer to true parameter values.

Table 4.4 shows the median standard errors for estimates of $\beta_1,\ \beta_2$ and β_6 together with the actual standard errors. In this case, only the correct selections are used for standard error calculation; otherwise, the standard errors can be underestimated or overestimated. Overall, standard error decreases when the sample size increases. There is no much dis-

crepancy between the estimated and actual standard errors for all five methods.

Table 4.4: Setting 1; SE of the estimates: Actual Standard Error (ASE) and Estimated Standard Error (ESE)

nrun=300, strong signal, censoring rate ≈ 40%							
n	Method	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_6$	
		ASE	ESE	ASE	ESE	ASE	ESE
100	SUBTLE	0.1781	0.1663	0.1782	0.1734	0.1589	0.1591
	Best Subset	0.1782	0.1583	0.1784	0.1807	0.1591	0.1569
	SCAD	0.1778	0.2120	0.1774	0.2054	0.1586	0.1750
	LASSO	0.1701	0.1503	0.1701	0.1501	0.1489	0.1432
	ALASSO	0.1694	0.1807	0.1739	0.1062	0.1487	0.1524
200	SUBTLE	0.1227	0.1279	0.1225	0.1301	0.1093	0.1095
	Best Subset	0.1227	0.1257	0.1227	0.1273	0.1094	0.1088
	SCAD	0.1223	0.1355	0.1217	0.1370	0.1087	0.1084
	LASSO	0.1216	0.1107	0.1231	0.0984	0.1083	0.1095
	ALASSO	0.1204	0.0968	0.1218	0.1141	0.1066	0.0966
300	SUBTLE	0.0993	0.1079	0.0989	0.0979	0.0882	0.0930
	Best Subset	0.0993	0.1055	0.0988	0.0960	0.0883	0.0904
	SCAD	0.0992	0.1148	0.0988	0.1053	0.0884	0.0950
	LASSO	0.0971	0.1064	0.0985	0.0842	0.0860	0.0758
	ALASSO	0.0978	0.1155	0.0983	0.0701	0.0863	0.0847

Table 4.5 summarizes the mean values of the model statistics for weak signal in the presence of higher censoring rate (40%). Overall, all the methods fail to perform well with the smaller sample size $n = 100$. However subtle uprooting and best subset selection have an improved performance with larger sample sizes by producing higher correct selections and lower model errors. Subtle method outperforms all five methods in the concern of correct selection for $n = 200$ and $n = 300$. When compared with the setting 1 results, it

can be observed that all the methods perform better with strong signals. Furthermore it is interesting to notice that ALASSO provides the lowest model error for all sample sizes in this setting.

Table 4.5: Setting 2 – Model Statistics

nrun=300, weak signal, censoring rate ≈ 40%						
n	Method	ME	SIZE	Under	Over	correction.selection
100	SUBTLE	0.1867	2.3700	0.8733	0.0500	0.0767
	Best Subset	0.1840	2.0000	0.9300	0.0200	0.0500
	SCAD	0.1324	2.9067	0.7300	0.1567	0.1133*
	LASSO	0.1239	4.0333	0.5400	0.4033	0.0567
	ALASSO	0.1183*	4.2100	0.5267	0.4167	0.0567
200	SUBTLE	0.0856	2.5867	0.6867	0.0800	0.2333*
	Best Subset	0.0899	2.3833	0.7333	0.0500	0.2167
	SCAD	0.0773	3.6967	0.4267	0.4133	0.1600
	LASSO	0.0643	4.8533	0.2667	0.6533	0.0800
	ALASSO	0.0628*	4.8867	0.2400	0.6733	0.0867
300	SUBTLE	0.0444	2.8833	0.4200	0.1067	0.4733*
	Best Subset	0.0485	2.6933	0.5000	0.0667	0.4333
	SCAD	0.0464	3.9700	0.2600	0.4967	0.2433
	LASSO	0.0430	5.1800	0.1233	0.7933	0.0833
	ALASSO	0.0425*	5.0967	0.1100	0.8000	0.0900

* minimum value for ME and maximum correct proportion within each sample size

Table 4.6 presents the the median values of the parameter estimates for all five methods. Here the actual parameters are $\boldsymbol{\beta} = (-0.4, -0.3, 0, 0, 0, -0.2, 0, 0, 0)^T$. With the small sample sizes, all the methods fail to provide median values that are closer to true parameters. However in the $n = 300$ case, subtle uprooting, best subset selection, and SCAD provide

better estimates for parameters. This shows that, a larger sample size is needed in order to detect smaller signals.

Table 4.6: Setting 2 – Median values of the estimates

nrun=300, weak signal, censoring rate ≈ 40%										
n	Method	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$
100	SUBTLE	-0.4463	-0.3069	0	0	0	0	0	0	0
	Best Subset	-0.4586	0	0	0	0	0	0	0	0
	SCAD	-0.4130	-0.2397	0	0	0	0	0	0	0
	LASSO	-0.2919	-0.1984	0	0	0	-0.0292	0	0	0
	ALASSO	-0.3062	-0.2110	0	0	0	-0.0528	0	0	0
200	SUBTLE	-0.4320	-0.3137	0	0	0	0	0	0	0
	Best Subset	-0.4351	-0.3122	0	0	0	0	0	0	0
	SCAD	-0.4250	-0.2842	0	0	0	-0.1400	0	0	0
	LASSO	-0.3467	-0.2371	0	0	0	-0.1059	0	0	0
	ALASSO	-0.3528	-0.2340	0	0	0	-0.1105	0	0	0
300	SUBTLE	-0.4131	-0.2888	0	0	0	-0.1819	0	0	0
	Best Subset	-0.4141	-0.2915	0	0	0	-0.2016	0	0	0
	SCAD	-0.4142	-0.2844	0	0	0	-0.1867	0	0	0
	LASSO	-0.3550	-0.2400	0	0	0	-0.1406	0	0	0
	ALASSO	-0.3581	-0.2411	0	0	0	-0.1354	0	0	0

Table 4.7 provides the actual standard errors for $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_6$. Although there are some discrepancies between the actual and estimated standard errors in the small sample size case, they become closer to each other in the $n = 300$ case.

Table 4.7: Setting 2: SE of the estimates ASE:Actual Standard Error, ESE: Estimated Standard Error

nrun=300, weak signal, censoring rate ≈ 40%							
n	Method	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_6$	
		ASE	ESE	ASE	ESE	ASE	ESE
100	SUBTLE	0.1656	0.1232	0.1625	0.0899	0.1436	0.0859
	Best Subset	0.1589	0.1003	0.1594	0.0886	0.1435	0.0907
	SCAD	0.1634	0.1728	0.1605	0.1219	0.1418	0.1136
	LASSO	0.1579	0.1770	0.1581	0.1102	0.1364	0.0833
	ALASSO	0.1607	0.1994	0.1625	0.1052	0.1372	0.0755
200	SUBTLE	0.1136	0.1135	0.1129	0.0902	0.0966	0.0796
	Best Subset	0.1142	0.1141	0.1132	0.0781	0.0974	0.0613
	SCAD	0.1129	0.1477	0.1131	0.1178	0.0959	0.1008
	LASSO	0.1125	0.1303	0.1106	0.1015	0.0942	0.0782
	ALASSO	0.1117	0.1175	0.1107	0.0903	0.0954	0.0822
300	SUBTLE	0.0916	0.0932	0.0903	0.0835	0.0774	0.0709
	Best Subset	0.0915	0.0857	0.0903	0.0665	0.0774	0.0510
	SCAD	0.0912	0.1076	0.0900	0.1028	0.0776	0.0759
	LASSO	0.0907	0.0985	0.0927	0.0804	0.0762	0.0667
	ALASSO	0.0910	0.1069	0.0913	0.0891	0.0768	0.0722

Table 4.8 summarizes the mean values for model statistics for subtle uprooting, best subset selection, SCAD, LASSO and adaptive LASSO methods where the strong signal is used with censoring rate ≈ 25%. Best subset selection provides the highest correct selection in all sample sizes and smallest model error in $n = 100$ and $n = 200$ cases. Subtle method provides the smallest model error in $n = 300$ case. In this setting, the subtle uprooting method also outperforms SCAD, LASSO and adaptive LASSO methods by exhibiting higher correct selection and lower model error.

Table 4.8: Setting 3:Model Statistics

nrun=300, strong signal, censoring rate ≈ 25%						
n	Method	ME	SIZE	Under	Over	correct selection
100	SUBTLE	0.1257	3.4533	0.0100	0.3333	0.6567
	Best Subset	0.1222*	3.3067	0.0300	0.2433	0.7267*
	SCAD	0.1300	3.3533	0.0567	0.2867	0.6567
	LASSO	0.1500	5.5467	0.0000	0.9133	0.0867
	ALASSO	0.1531	5.5033	0.0033	0.9100	0.0867
200	SUBTLE	0.0398	3.2567	0.0000	0.2167	0.7833
	Best Subset	0.0383*	3.1533	0.0000	0.1367	0.8633*
	SCAD	0.0503	3.8200	0.0000	0.3900	0.6100
	LASSO	0.0666	5.5267	0.0000	0.9233	0.0767
	ALASSO	0.0655	5.5767	0.0000	0.9267	0.0733
300	SUBTLE	0.0287*	3.3600	0.0000	0.3133	0.6867
	Best Subset	0.0292	3.2100	0.0000	0.1700	0.8300*
	SCAD	0.0347	3.8767	0.0000	0.4167	0.5833
	LASSO	0.0455	5.6733	0.0000	0.9400	0.0600
	ALASSO	0.0459	5.6100	0.0000	0.9200	0.0800

* minimum value for ME and maximum correct proportion within each sample size

Table 4.9 provides the median values for parameter estimates. Here the actual parameters are $\boldsymbol{\beta} = (-0.4, -0.3, 0, 0, 0, -0.2, 0, 0, 0)^T$. For the sample sizes $n = 200$ and 300, subtle method, best subset selection and SCAD produce very close median values to the actual parameter values. However LASSO and adaptive LASSO median estimates are deviated from the true values.

Table 4.9: Setting 3:Meadian values of estimates

nrun=300, strong signal, censoring rate ≈ 25%										
n	Method	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$
100	SUBTLE	-0.7332	-0.7040	0	0	0	-0.7301	0	0	0
	Best Subset	-0.7205	-0.7005	0	0	0	-0.7220	0	0	0
	SCAD	-0.7013	-0.6673	0	0	0	-0.6594	0	0	0
	LASSO	-0.5970	-0.5720	0	0	0	-0.5494	0	0	0
	ALASSO	-0.6022	-0.5715	0	0	0	-0.5465	0	0	0
200	SUBTLE	-0.7064	-0.7100	0	0	0	-0.6993	0	0	0
	Best Subset	-0.7078	-0.7085	0	0	0	-0.6992	0	0	0
	SCAD	-0.6989	-0.7100	0	0	0	-0.6947	0	0	0
	LASSO	-0.6295	-0.6236	0	0	0	-0.5940	0	0	0
	ALASSO	-0.6291	-0.6311	0	0	0	-0.5932	0	0	0
300	SUBTLE	-0.7112	-0.7034	0	0	0	-0.7121	0	0	0
	Best Subset	-0.7110	-0.7058	0	0	0	-0.7134	0	0	0
	SCAD	-0.7060	-0.7026	0	0	0	-0.7111	0	0	0
	LASSO	-0.6513	-0.6400	0	0	0	-0.6308	0	0	0
	ALASSO	-0.6471	-0.6404	0	0	0	-0.6260	0	0	0

Table 4.10 shows the actual standard errors for $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_6$. Similar to other settings, not much discrepancies between the actual and estimated standard errors are observed. Standard errors for the estimates in all five methods decreases as the sample size increases.

Table 4.10: Setting 3: SE of the estimates ASE:Actual Standard Error, ESE: Estimated Standard Error

nrun=300, strong signal, censoring rate ≈ 25%							
n	Method	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_6$	
		ASE	ESE	ASE	ESE	ASE	ESE
100	SUBTLE	0.1591	0.1645	0.1590	0.1669	0.1424	0.1616
	Best Subset	0.1597	0.1578	0.1589	0.1658	0.1425	0.1607
	SCAD	0.1584	0.1890	0.1585	0.1908	0.1411	0.1712
	LASSO	0.1515	0.1651	0.1519	0.1410	0.1380	0.1600
	ALASSO	0.1506	0.1390	0.1495	0.1525	0.1378	0.1514
200	SUBTLE	0.1093	0.1127	0.1094	0.1047	0.0975	0.0931
	Best Subset	0.1094	0.1131	0.1096	0.1060	0.0976	0.0924
	SCAD	0.1086	0.1324	0.1091	0.1273	0.0974	0.1055
	LASSO	0.1068	0.1197	0.1071	0.1158	0.0949	0.0994
	ALASSO	0.1075	0.1241	0.1074	0.1159	0.0952	0.1034
300	SUBTLE	0.0888	0.0909	0.0887	0.0873	0.0793	0.0825
	Best Subset	0.0886	0.0872	0.0887	0.0888	0.0793	0.0836
	SCAD	0.0886	0.1005	0.0884	0.0988	0.0792	0.0847
	LASSO	0.0868	0.1073	0.0855	0.0650	0.0769	0.0926
	ALASSO	0.0873	0.0964	0.0862	0.0688	0.0771	0.0887

Table 4.11 shows the model statistics for five methods under weak signal and lower censoring rate. Although SCAD method provides the highest correct selection in the $n = 100$ case, subtle method outperforms all five methods in terms of correct selection for both $n = 200, 300$ cases. In the $n = 300$ case, subtle method outperforms all other by providing highest correct selection and smallest model error. This result exhibits that subtle method performs comparatively better than other methods with higher sample sizes even in the presence of weak signal.

Table 4.11: Setting 4:Model Statistics

nrun=300, weak signal, censoring rate ≈ 25%						
n	Method	ME	SIZE	Under	Over	correct selection
100	SUBTLE	0.1535	2.4967	0.8033	0.0833	0.1133
	Best Subset	0.1537	2.2333	0.8600	0.0400	0.1000
	SCAD	0.1133	3.2067	0.6300	0.2400	0.1300*
	LASSO	0.0999	4.5667	0.4067	0.5500	0.0433
	ALASSO	0.0988*	4.4067	0.4300	0.5033	0.0667
200	SUBTLE	0.0598	2.7633	0.5400	0.0833	0.3767*
	Best Subset	0.0647	2.5933	0.6100	0.0533	0.3367
	SCAD	0.0566	3.9800	0.3267	0.4667	0.2067
	LASSO	0.0502*	5.0533	0.1900	0.7300	0.0800
	ALASSO	0.0512	4.9167	0.1900	0.7133	0.0967
300	SUBTLE	0.0287*	3.0967	0.2000	0.1667	0.6333*
	Best Subset	0.0324	2.8767	0.3067	0.0767	0.6167
	SCAD	0.0321	3.9967	0.1233	0.4900	0.3867
	LASSO	0.0301	5.1167	0.0700	0.8300	0.1000
	ALASSO	0.0300	5.1733	0.0667	0.8367	0.0967

* minimum value for ME and maximum correct proportion within each sample size

The median values of the estimates for the weak signal in the presence of lower censoring rate (25%) are presented in table 4.12. With the large sample sizes, all methods provide much closer values to true values. Although LASSO and adaptive LASSO methods provide non zero median estimates in the $n = 100$ case, those values greatly deviate from the true values.

Table 4.12: Setting 4:Median Values of estimates

nrun=300, weak signal, censoring rate ≈ 25%										
n	Method	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$
100	SUBTLE	-0.4541	-0.3202	0	0	0	0	0	0	0
	Best Subset	-0.4582	-0.3157	0	0	0	0	0	0	0
	SCAD	-0.4147	-0.2625	0	0	0	0	0	0	0
	LASSO	-0.3204	-0.2272	0	0	0	-0.0723	0	0	0
	ALASSO	-0.3180	-0.2218	0	0	0	-0.0658	0	0	0
200	SUBTLE	-0.4078	-0.2917	0	0	0	-0.1777	0	0	0
	Best Subset	-0.4065	-0.2922	0	0	0	-0.2037	0	0	0
	SCAD	-0.4041	-0.2732	0	0	0	-0.1661	0	0	0
	LASSO	-0.3450	-0.2319	0	0	0	-0.1266	0	0	0
	ALASSO	-0.3364	-0.2283	0	0	0	-0.1236	0	0	0
300	SUBTLE	-0.4056	-0.3014	0	0	0	-0.2105	0	0	0
	Best Subset	-0.4059	-0.3019	0	0	0	-0.2159	0	0	0
	SCAD	-0.4096	-0.3025	0	0	0	-0.1922	0	0	0
	LASSO	-0.3589	-0.2559	0	0	0	-0.1527	0	0	0
	ALASSO	-0.3562	-0.2594	0	0	0	-0.1521	0	0	0

According to table 4.13, there are some discrepancies between actual standard errors and the estimated standard errors, specially when the sample size is small. However in larger sample size ($n = 300$), the actual and the estimated standard errors are much close to each other.

Table 4.13: Setting 4: SE of the estimates ASE:Actual Standard Error, ESE: Estimated Standard Error

nrun=300, weak signal, censoring rate ≈ 25%							
n	Method	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_6$	
		ASE	ESE	ASE	ESE	ASE	ESE
100	SUBTLE	0.1492	0.1015	0.1469	0.0750	0.1259	0.0584
	Best Subset	0.1491	0.1012	0.1436	0.0710	0.1248	0.0601
	SCAD	0.1506	0.1394	0.1476	0.1343	0.1268	0.0891
	LASSO	0.148583	0.119676	0.1494	0.1180	0.1250	0.0911
	ALASSO	0.1508	0.1057	0.1484	0.1214	0.1251	0.0810
200	SUBTLE	0.0999	0.0953	0.0980	0.0771	0.0859	0.0705
	Best Subset	0.0999	0.0902	0.0982	0.0661	0.0859	0.0525
	SCAD	0.1001	0.1111	0.0982	0.1066	0.0862	0.0879
	LASSO	0.1003	0.0985	0.0987	0.0922	0.0848	0.0502
	ALASSO	0.1001	0.0966	0.0987	0.0906	0.0853	0.0616
300	SUBTLE	0.08215	0.0806	0.0808	0.0837	0.0692	0.0728
	Best Subset	0.0819	0.0733	0.0805	0.0700	0.0694	0.0518
	SCAD	0.0818	0.0881	0.0803	0.0941	0.0696	0.0750
	LASSO	0.0816	0.0785	0.0805	0.0757	0.0681	0.0632
	ALASSO	0.0811	0.0789	0.0808	0.0790	0.0679	0.0672

4.2.1 Performance in the presence of correlation

In real world data, it is common to have correlated variables. Usually, there are limitations applied to many conventional statistical methods when dealing with correlated data. Therefore it is important to access the performance of these variables selection methods when the covariates are correlated. In order to assess this effect, two extra simulation settings are considered.

Table 4.14: Simulation settings

setting	$\boldsymbol{\beta}$	emphirical censoring rate	ρ
setting 5	$(-0.7, -0.7, 0, 0, 0, 0, 0, -0.7)^T$	25%	0.2
setting 6	$(-0.7, -0.7, 0, 0, 0, 0, 0, -0.7)^T$	25%	0.8

In both settings, the same signal and censoring rate are used, but covariates in Setting 5 are less correlated than those in Setting 6. In particular, the correlation between x_i and x_j is $0.2^{|j-i|}$ in setting 5 whilst it is $0.8^{|j-i|}$ in setting 6. However instead of using the signal $(-0.7, -0.7, 0, 0, 0, -0.7, 0, 0)^T$ which is used in previous settings, $(-0.7, -0.7, 0, 0, 0, 0, 0, -0.7)^T$ is used. Here the methods should select the variables $x_1,\ x_2$ and x_9 and estimate their associated coefficients. In this case the variables $x_1,\ x_2$ have the maximum possible correlation while $x_1,\ x_9$ have the minimum correlation under each setting. In particular the correlation between $x_1,\ x_9$ in setting 5 is approximately 0 and in setting 6 it is approximately 0.1677. Covariates $x_1,\ x_2$ have a correlation of 0.2 in setting 5 and 0.8 in setting 6.

Table 4.15: Setting 5:Model Statistics

nrun=300, strong signal, censoring control 3, rho=0.2						
n	Method	ME	SIZE	Under	Over	correct selection
100	SUBTLE	0.1119	3.4033	0.0033	0.3200	0.6767
	Best Subset	0.1025*	3.2667	0.0033	0.2267	0.7700*
	SCAD	0.1059	3.4767	0.0100	0.3667	0.6233
	LASSO	0.1468	5.4633	0	0.9267	0.0733
	ALASSO	0.1484	5.4633	0	0.9233	0.0767
200	SUBTLE	0.0391	3.3033	0	0.2667	0.7333
	Best Subset	0.0383*	3.1567	0	0.1433	0.8567*
	SCAD	0.0477	3.7533	0	0.3467	0.6533
	LASSO	0.0713	5.7267	0	0.9267	0.0733
	ALASSO	0.0701	5.7333	0	0.95	0.05
300	SUBTLE	0.0249	3.3433	0	0.3100	0.6900
	Best Subset	0.0245*	3.1300	0	0.1267	0.8733*
	SCAD	0.0288	3.6700	0	0.3533	0.6467
	LASSO	0.0463	5.9300	0	0.9733	0.0267
	ALASSO	0.0459	6.0033	0	0.9533	0.0467

* minimum value for ME and maximum correct proportion within each sample size

Table 4.15 summarizes the mean values of model statistics over 300 runs. In this case covariates are less correlated. The maximum correlation between two covariates is 0.2. For all sample sizes, best subset selection shows the minimum model error and maximum correct selection proportion. However, the second best performance is obtained via subtle uprooting method. In this case LASSO and adaptive LASSO methods also fail to perform well.

Table 4.16: Setting 5:Median Values of estimates

nrun=300, strong signal, censoring control 3, rho=0.2										
n	Method	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$
100	SUBTLE	-0.7399	-0.7176	0	0	0	0	0	0	-0.7258
	Best Subset	-0.7391	-0.7163	0	0	0	0	0	0	-0.7266
	SCAD	-0.7159	-0.6896	0	0	0	0	0	0	-0.6942
	LASSO	-0.5912	-0.5731	0	0	0	0	0	0	-0.5568
	ALASSO	-0.5776	-0.5667	0	0	0	0	0	0	-0.5601
200	SUBTLE	-0.6938	-0.7055	0	0	0	0	0	0	-0.7166
	Best Subset	-0.6941	-0.7025	0	0	0	0	0	0	-0.7149
	SCAD	-0.6906	-0.7009	0	0	0	0	0	0	-0.6960
	LASSO	-0.6001	-0.6153	0	0	0	0	0	0	-0.6100
	ALASSO	-0.6065	-0.6133	0	0	0	0	0	0	-0.6086
300	SUBTLE	-0.7082	-0.7079	0	0	0	0	0	0	-0.7035
	Best Subset	-0.7086	-0.7078	0	0	0	0	0	0	-0.7026
	SCAD	-0.7058	-0.7031	0	0	0	0	0	0	-0.6987
	LASSO	-0.6326	-0.6355	0	0	0	0	0	0	-0.6218
	ALASSO	-0.6365	-0.6377	0	0	0	0	0	0	-0.6251

Table 4.16 shows the median estimates. While all the methods provide median estimates that are close to the true values $(-0.7, -0.7, 0, 0, 0, 0, 0, -0.7)^T$, subtle uprooting, best subset and SCAD methods provide much closer median estimates than LASSO and adaptive LASSO.

Table 4.17: Setting 5: SE of the estimates ASE:Actual Standard Error, ESE: Estimated Standard Error

nrun=300, strong signal, censoring control 3, rho=0.2							
n	Method	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_9$	
		ASE	ESE	ASE	ESE	ASE	ESE
100	SUBTLE	0.1461	0.1577	0.1450	0.1439	0.1431	0.1350
	Best Subset	0.1460	0.1533	0.1448	0.1425	0.1434	0.1360
	SCAD	0.1454	0.1767	0.1434	0.1552	0.1420	0.1577
	LASSO	0.1418	0.1326	0.1397	0.1129	0.1360	0.1077
	ALASSO	0.1429	0.1355	0.1407	0.0939	0.1342	0.0972
200	SUBTLE	0.0993	0.0906	0.0997	0.1042	0.0980	0.1002
	Best Subset	0.0993	0.0917	0.0995	0.1011	0.0978	0.0980
	SCAD	0.0990	0.1050	0.0995	0.1165	0.0978	0.1040
	LASSO	0.0977	0.0935	0.0962	0.0892	0.0954	0.1145
	ALASSO	0.0992	0.0685	0.0972	0.0999	0.0960	0.1174
300	SUBTLE	0.0803	0.0737	0.0802	0.0862	0.0788	0.0770
	Best Subset	0.0803	0.0742	0.0802	0.0878	0.0788	0.0768
	SCAD	0.0801	0.0816	0.0803	0.0941	0.0787	0.0827
	LASSO	0.0777	0.0838	0.0781	0.0710	0.0772	0.0801
	ALASSO	0.0783	0.0882	0.0781	0.0689	0.0776	0.0830

According to Table 4.17 we can observe that there is no significant discrepancy between the actual and the estimated standard errors. Further more, standard errors decreases with the increasing sample size.

Table 4.18 summarizes the model statistics for setting 6. Overall performance is poor in $n = 100$ case. However subtle and best subset selection methods greatly improve the correct selection proportion with the increasing sample size. However, performance of SCAD method is decreased considerably after introducing the correlation to data. In this case also best subset selection shows the best performance and subtle uprooting provides

second best correct selection proportions.

Table 4.18: Setting 6:Model Statistics

nrun=300, strong signal, censoring control 3, rho=0.8						
n	Method	ME	SIZE	Under	Over	correct selection
100	SUBTLE	0.1751	3.6067	0.2300	0.3533	0.4167
	Best Subset	0.1698	3.0967	0.3400	0.1467	0.5133*
	SCAD	0.3251	3.1700	0.5933	0.1533	0.2533
	LASSO	0.1325*	5.1967	0.0133	0.9067	0.0800
	ALASSO	0.1343	5.1700	0.0133	0.8967	0.0900
200	SUBTLE	0.0570	3.4400	0.0233	0.2433	0.7333
	Best Subset	0.0558*	3.2533	0.0500	0.1533	0.7967*
	SCAD	0.0660	4.1200	0.0100	0.5233	0.4667
	LASSO	0.0631	5.3900	0	0.9200	0.0800
	ALASSO	0.0642	5.4433	0	0.9233	0.0767
300	SUBTLE	0.0304	3.3200	0.0033	0.2100	0.7867
	Best Subset	0.0285*	3.2033	0.0067	0.1500	0.8433*
	SCAD	0.0365	3.9233	0	0.4433	0.5567
	LASSO	0.0408	5.4333	0	0.8967	0.1033
	ALASSO	0.0411	5.4267	0	0.9033	0.0967

* minimum value for ME and maximum correct proportion within each sample size

Observed median estimates are recorded in Table 4.19. In the $n = 100$ case, SCAD method provides the median estimate for $\hat{\beta}_9$ as zero. Surprisingly it gives the zero median estimate for the least correlated variable. SCAD method provides much closer median estimates in $n = 200,\ 300$ cases. Median estimates of best subset and subtle method do not deviate greatly from the actual values.

Table 4.19: Setting 6:Median Values of estimates

nrun=300, strong signal, censoring control 3, rho=0.8										
n	Method	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$
100	SUBTLE	-0.7246	-0.7435	0	0	0	0	0	0	-0.7269
	Best Subset	-0.7334	-0.7249	0	0	0	0	0	0	-0.7184
	SCAD	-0.6617	-0.6641	0	0	0	0	0	0	0
	LASSO	-0.6327	-0.6014	0	0	0	0	0	0	-0.5683
	ALASSO	-0.6263	-0.5923	0	0	0	0	0	0	-0.5645
200	SUBTLE	-0.7028	-0.7268	0	0	0	0	0	0	-0.7158
	Best Subset	-0.7028	-0.7257	0	0	0	0	0	0	-0.7171
	SCAD	-0.7036	-0.7040	0	0	0	0	0	0	-0.7084
	LASSO	-0.6483	-0.6235	0	0	0	0	0	0	-0.5955
	ALASSO	-0.6568	-0.6262	0	0	0	0	0	0	-0.5995
300	SUBTLE	-0.7191	-0.6970	0	0	0	0	0	0	-0.7056
	Best Subset	-0.7204	-0.6970	0	0	0	0	0	0	-0.7085
	SCAD	-0.7189	-0.6979	0	0	0	0	0	0	-0.7025
	LASSO	-0.6761	-0.6292	0	0	0	0	0	0	-0.6265
	ALASSO	-0.6807	-0.6247	0	0	0	0	0	0	-0.6282

Table 4.20 provides the actual and estimated standard error. Standard errors for all methods are higher than the standard errors obtained under setting 6.

Table 4.20: Setting 6: SE of the estimates ASE:Actual Standard Error, ESE: Estimated Standard Error

nrun=300, strong signal, censoring control 3, rho=0.8							
n	Method	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_9$	
		ASE	ESE	ASE	ESE	ASE	ESE
100	SUBTLE	0.2183	0.1819	0.2191	0.2086	0.1456	0.1317
	Best Subset	0.2188	0.1738	0.2178	0.1895	0.1450	0.1323
	SCAD	0.2185	0.2210	0.2171	0.2113	0.1500	0.1431
	LASSO	0.2115	0.1827	0.2199	0.2120	0.1403	0.1375
	ALASSO	0.2141	0.1926	0.2209	0.2218	0.1417	0.1525
200	SUBTLE	0.1482	0.1454	0.1498	0.1432	0.0991	0.0955
	Best Subset	0.1484	0.1474	0.1502	0.1390	0.0993	0.0978
	SCAD	0.1477	0.1603	0.1494	0.1686	0.0992	0.1097
	LASSO	0.1479	0.1714	0.1492	0.1589	0.0973	0.0898
	ALASSO	0.1501	0.1518	0.1497	0.1536	0.0975	0.0978
300	SUBTLE	0.1215	0.1176	0.1218	0.1195	0.0802	0.0766
	Best Subset	0.1215	0.1212	0.1220	0.1204	0.0803	0.0770
	SCAD	0.1214	0.1437	0.1222	0.1434	0.0804	0.0821
	LASSO	0.1213	0.1343	0.1212	0.1423	0.0800	0.0743
	ALASSO	0.1226	0.1261	0.1219	0.1300	0.0796	0.0691

4.3 Summary

Simulation study results were presented in the previous section and this section entails a summary of the simulation results. Setting 1 to setting 4 are used to access the performance with the different strengths of the signal, censoring rates and the sample sizes. Mean values for model error, size, overfitting, underfitting and correct selection proportions are calculated for each setting. Out of these five measures model error can be used to measure

how good the estimation, whilst correct selection proportion says how good the method is as a variable selection tool.

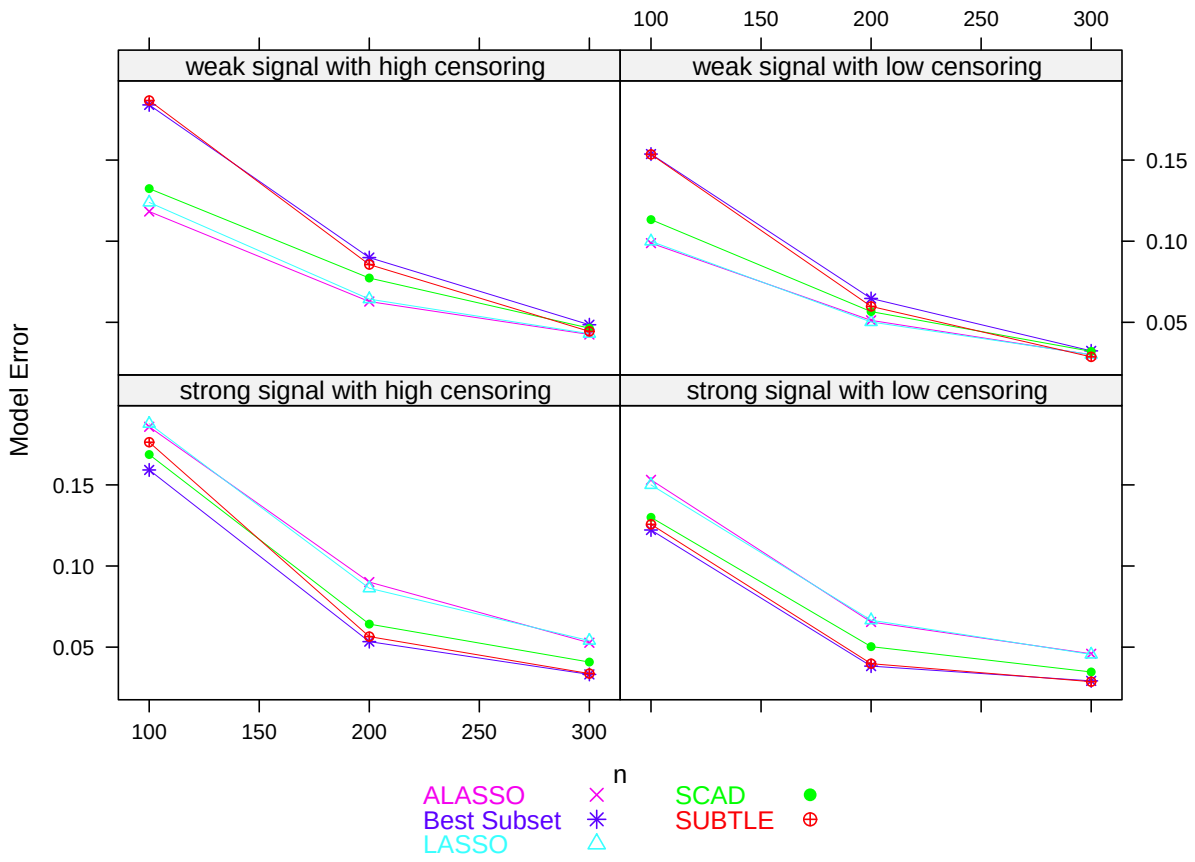

Figure 4.1: Model Error in different settings

Figure 4.1 summarizes the mean estimates of the model errors in different simulation settings. According to the results, mean value of model error for all five methods decreases with the sample size. It can be noticed that subtle uprooting method performs close to the best subset selection under all settings. Both subtle uprooting and best subset methods provide better estimates with a strong signal. However they higher model errors with weak signal in small sample sizes, but they perform well with the higher sample sizes even with a weak signal. This figure exhibits that subtle method performs the estimation well with

stronger signal or with higher sample sizes.

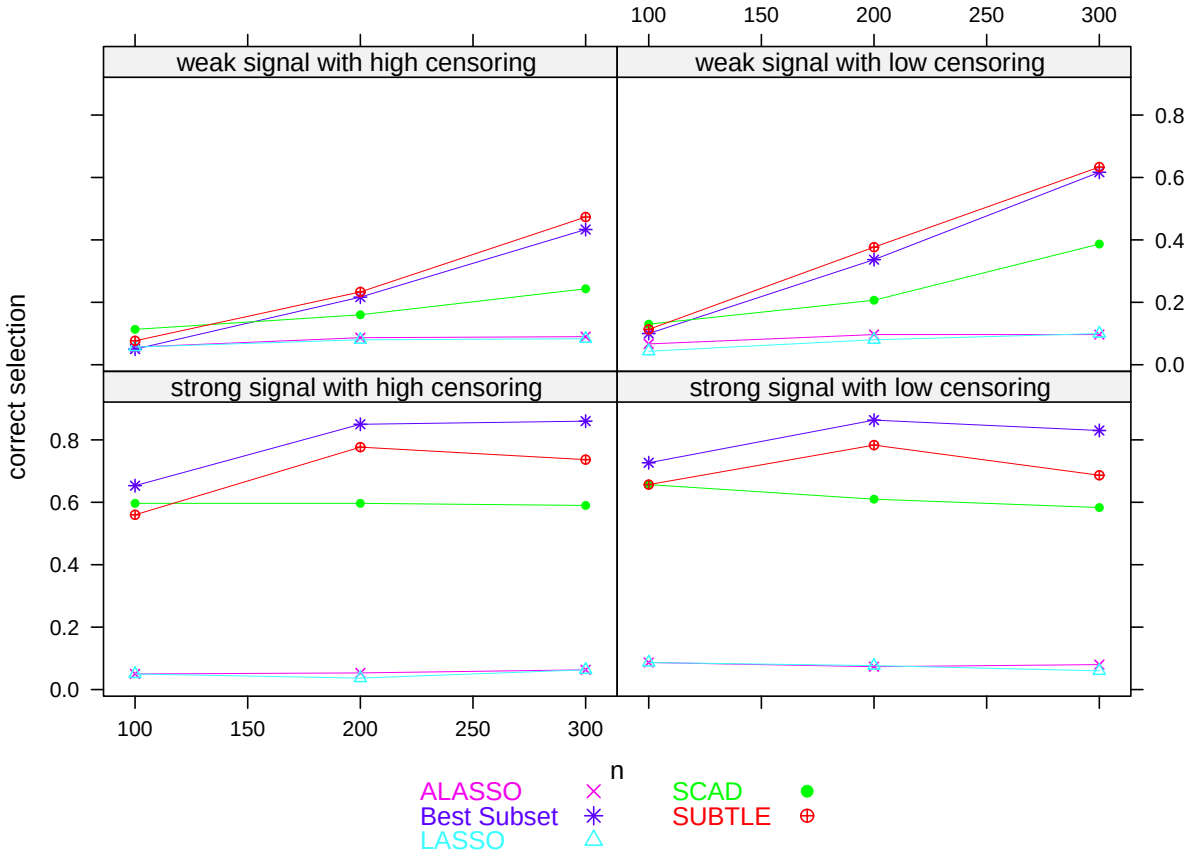

Figure 4.2: Correct selection in different settings

Figure 4.2 shows the mean values of correct selection proportions in different settings. According to the figure it is clear that as far as correct selection is considered, subtle method outperforms all the methods when the signal is weak. This implies that, under this setting subtle method performs the best variable selection. Best subset selection shows the best correct selection when the signal is strong. However, subtle uprooting in all four settings shows a similar performance to best subset selection. This figure shows the appropriateness of subtle uprooting method as a variable selection method in all settings.

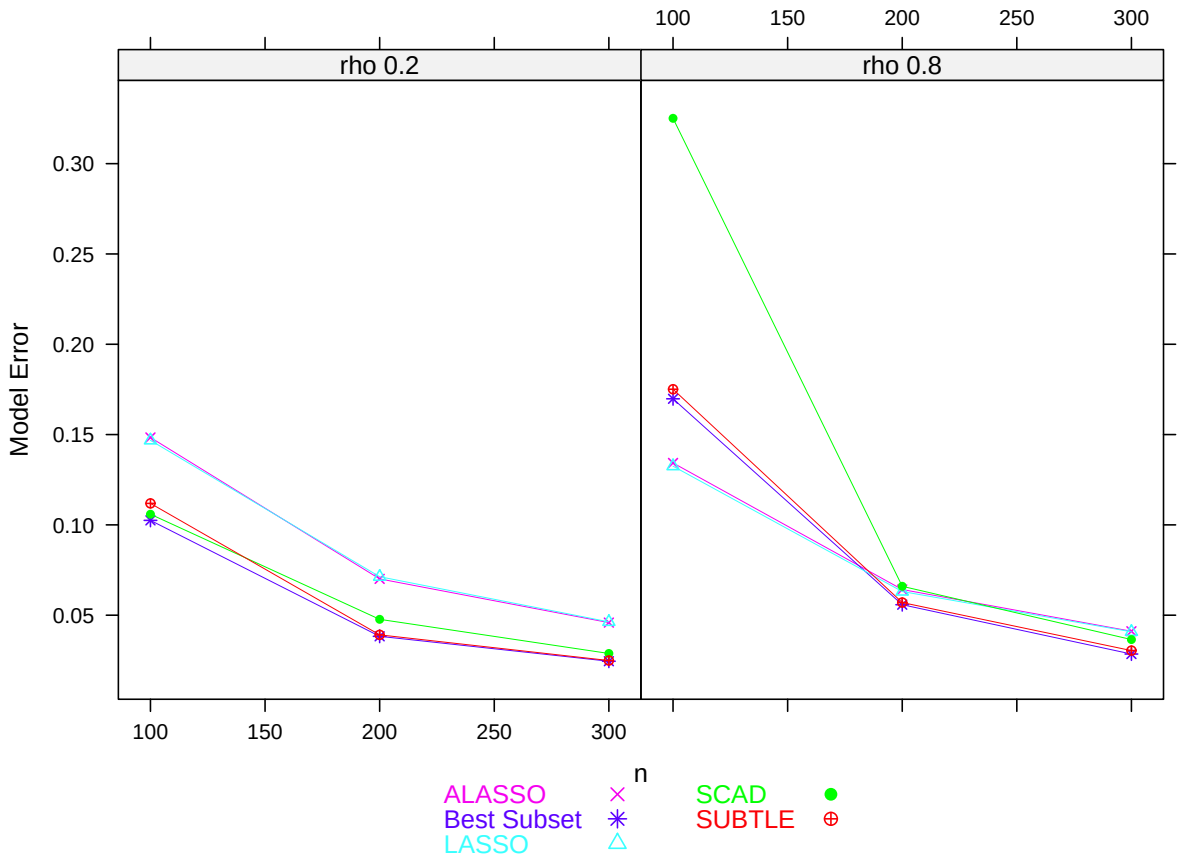

Figure 4.3: Model error for different correlation setting

Figure 4.3 shows the mean values for model error for $\rho = 0.2$ and 0.8. From the figure it can be clearly seen that, when the covariates are highly correlated, overall performance of all methods decreases. However SCAD method shows huge model error in $n = 100$ case. In both situations, Subtle method performs very close to the best subset selection method. This figure shows that, subtle method is capable of performing well even with the correlated data specially with a higher sample size.

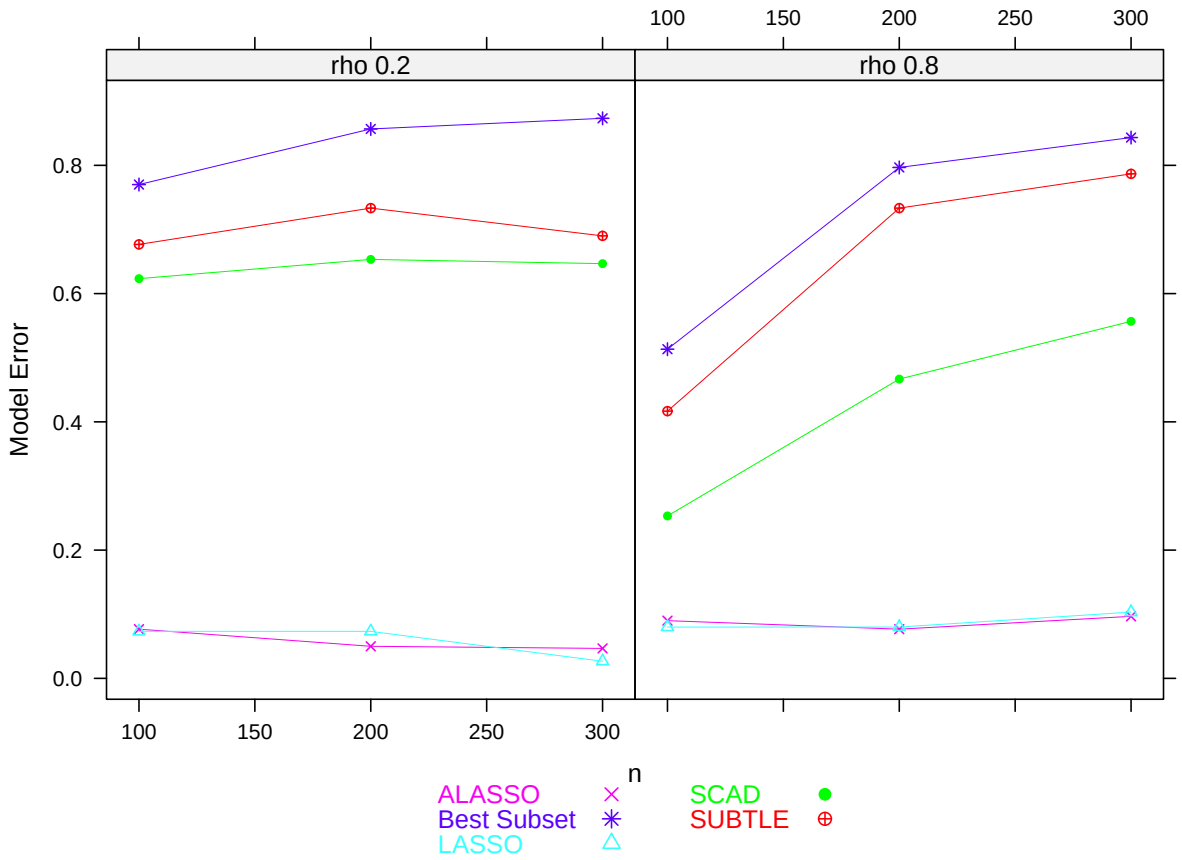

Figure 4.4: Correct selection for different correlation setting

Mean values for correct selection in the presence of high correlation and low correlation are shown in figure 4.4. According to the figure, it is clear that, overall performance is decreased when the covariates are highly correlated. However in both settings, best subset selection performs as the best variable selection method. The second best is subtle method. Although SCAD method performs bit closer to subtle method with less correlated data, there is a considerable discrepancy when the covariates become highly correlated. The correct selection proportions for LASSO and adaptive LASSO methods are very low.

All the simulation results show that subtle method performs similar to the best subset selection under all settings, specially with higher sample sizes. As discussed in chapter 2, best subset selection becomes infeasible in higher dimensions. Therefore from the simula-

tion studies, we can recommend to use subtle method as an approximation of best subset selection method in the higher dimensions.

4.4 Data Example

For further illustration, we consider the well-known PBC data set. Primary biliary cirrhosis (PBC) is a rare but fatal chronic liver disease of unknown cause,with a prevalence of about 50 cases per million population. The available data were collected from the Mayo Clinic trial in PBC of the liver conducted between 1974 and 1984 to compare the drug D-penicillamine (DPCA) with a placebo. A total of 424 PBC patients, referred to Mayo Clinic during that ten year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The first 312 cases in the data set participated in the randomized trial. The additional 112 cases did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival. Six of those cases were lost to follow-up shortly after diagnosis, so there are data on an additional 106 cases as well as the 312 randomized participants. Large number of clinical, biochemical, serologic, and histologic variables were recorded for each of the 312 clinical trial patients.

The data discussed here are important in two respects. First, controlled clinical trials are difficult to complete in rare diseases, and in this case, series of patients uniformly diagnosed, treated, and followed up. The treatment comparison in this trial is more precise than in similar trials having fewer participants and avoids the bias that may arise in comparing a case series to historical controls.

Secondly, the data present an opportunity to study the natural history of the disease. Despite the immunosuppressive properties of DPCA, there are no detectable differences between the distributions of survival times for the DPCA and placebo treatment groups. This suggests that these groups can be combined in studying the association between survival time from randomization and clinical and other measurements. In the early to mid 1980s, the rate of successful liver transplant increased substantially, and transplant

has become an effective therapy for PBC. The Mayo Clinic data set is therefore one of the last, allowing a study of the natural history of PBC in patients who were treated with only supportive care or its equivalent. In this analysis, missing values are removed from the data set, to maintain the consistency with (Tibshirani, 1997). After the removal of missing values there are 276 observations in the data set. Out of this 276, there are 165 censored observations yielding a censoring rate of 59.77%. There are 17 covariates in this data set. The description of the covariates is presented in table 4.21. Data set contains both quantitative and qualitative variables.

Table 4.21: Variables in the PBC dataset

Variable	Description
trt:	D-penicillmain or placebo
age:	in years
sex:	male or female
ascites:	presence of ascites (yes or no)
hepato:	presence of hepatomegaly (yes or no)
spiders:	presence of blood vessel malformations in the skin (yes or no)
edema:	presence of edema (no edema, untreated or successfully treated, edema despite diuretic therapy
bili:	Serum bilirunbin (mg/dl)
chol:	Serum cholesterol (mg/dl)
albumin:	Serum albumin (g/dl)
copper:	urine copper (ug/day)
alk.phos:	alkaline phosphotase (U/liter)
ast	aspartate aminotransferase, once called SGOT (U/ml)
trig:	triglycerides (mg/dl)
platelet:	platelet count
protime:	standardised blood clotting time
stage:	histologic stage of disease

Table 4.22: Parameter estimates for PBC data

Covariate	Full	Subtle	Best subset	SCAD	LASSO	ALASSO
trt	-0.0622	0	0	0	0	0
	(0.1075)	-	-	-	-	-
age	0.3041	0.3283	0.3303	0.3227	0.1505	0.1802
	(0.1225)	(0.1072)	(0.1074)	(0.1065)	(0.1060)	(0.1062)
sex	(0.1204)	0	0	0	0	0
	(0.1025)	-	-	-	-	-
ascites	0.0224	0	0	0	0.0274	0.0241
	(0.0982)	-	-	-	(0.0972)	(0.0976)
hepato	0.0128	0	0	0	0	0
	(0.1257)	-	-	-	-	-
spiders	0.0460	0	0	0	0	0
	(0.1107)	-	-	-	-	-
edema	0.2733	0.2105	0.2222	0.1550	0.1725	0.1814
	(0.1065)	(0.0944)	(0.0939)	(0.0976)	(0.1005)	(0.0991)
bili	0.3681	0.3994	0.3917	0.4539	0.3869	0.3854
	(0.1173)	(0.0887)	(0.0890)	(0.0864)	(0.0934)	(0.0927)
chol	0.1155	0	0	0	0	0
	(0.1043)	-	-	-		-
albumin	-0.2999	-0.2953	-0.2909	-0.3045	-0.2155	-0.2284
	(0.1246)	(0.1101)	(0.1103)	(0.1093)	(0.1182)	(0.1181)
copper	0.2198	0.2515	0.2519	0.2269	0.2421	0.2454
	(0.1033)	(0.0869)	(0.0868)	(0.0891)	(0.0946)	(0.0936)
alk.phos	0.0022	0	0	0	0	0
	(0.0840)	-	-	-	-	-
ast	0.2308	0.2394	0.2483	0.1725	0.0503	0.0837
	(0.1111)	(0.1028)	(0.1025)	(0.1063)	(0.1106)	(0.1091)
trig	(0.0637)	0	0	0	0	0
	(0.0870)	-	-	-	-	-
platelet	0.0840	0	0	0	0	0
	(0.1103)	-	-	-	-	-
protime	0.2344	0.2183	0.2294	0.1478	0.1191	0.1371
	(0.1070)	(0.1029)	(0.1022)	(0.1063)	(0.1051)	(0.1046)
stage	0.3881	0.3720	0.3696	0.3936	0.2171	0.2387
	(0.1498)	(0.1245)	(0.1244)	(0.1247)	(0.1147)	(0.1163)

( ) :standard errors of the estimates

Table 4.22 summarizes the parameter estimates and their standard errors for full model, subtle uprooting, best subset selection, SCAD, LASSO and adaptive LASSO methods. Both subtle uprooting and SCAD methods provide the same variable selection as in best subset selection. Subtle uprooting estimates are very close to best subset selection estimates. Furthermore, the standard errors for subtle estimates are approximately same as best subset estimates. LASSO and adaptive LASSO select the variable ascities, which is not selected under any other method. According to the results, all five methods suggest that the treatment is not significant and hence does not significantly affect the hazard.

Based on the results obtained from subtle uprooting method, the resulting Cox PH model can be written as,

$$\begin{aligned}\lambda_i(t) &= \lambda_0(t)\exp(0.3283 \times \text{age} + 0.2105 \times \text{edema} + 0.3994 \times \text{bili} - 0.2953 \times \text{albumin} \\ &\quad + 0.2515 \times \text{copper} + 0.2394 \times \text{ast} + 0.2183 \times \text{protime} + 0.3720 \times \text{stage})\end{aligned}$$

Table 4.23: Confidence intervals for coefficients of covariates

covariate	coefficient	SE	Lower	Upper	exp(Lower)	exp(Upper)
age	0.3283	0.1072	0.1182	0.5384	1.1255	1.7133
edema	0.2105	0.0944	0.0255	0.3955	1.0258	1.4852
bili	0.3994	0.0887	0.2255	0.5733	1.2530	1.7740
albumin	-0.2953	0.1101	-0.5111	-0.0795	0.5998	0.9236
copper	0.2515	0.0869	0.0812	0.4218	1.0846	1.5247
ast	0.2394	0.1028	0.0379	0.4409	1.0386	1.5541
protime	0.2183	0.1029	0.0166	0.4200	1.0168	1.5219
stage	0.3720	0.1245	0.1280	0.6160	1.1365	1.8515

Table 4.23 provides the confidence intervals for coefficients of covariates and the hazard

rates. Provided in the last two columns of Table are the lower and upper limits for the 95% confidence interval of resulting hazard ratio with one unit increase in each covariate while holding all others fixed.

Chapter 5

Discussion

This chapter include a simple time analysis for five methods, discussion on the robustness of subtle uprooting method against the parameters used in the procedure and some important remarks. Furthermore this entails possible extensions on this work.

It is important to analyze the times spent for each method. For this task, data example is used. Here variable selection and model estimation for PBC data is done using subtle, best subset, SCAD, LASSO and adaptive LASSO methods over 10 runs and the system, user, and elapsed times are recorded. Here 'user time' is the CPU time charged for the execution of user instructions of the calling process. The 'system time' is the CPU time charged for execution by the system on behalf of the calling process. The difference in times since you started the stopwatch is given by 'elapsed time'

Table 5.1: Times consumed by each method

	user	system	elapsed
SUBTLE	2.64	0.00	2.63
Best subset	553.30	0.04	603.35
SCAD	8.15	0.03	8.18
LASSO	0.14	0.00	0.14
ALASSO	0.15	0.00	0.15

Table 5.1 gives the mean values for each time component for each method used. It can be seen that LASSO and adaptive LASSO performs very fast than other methods. Both LASSO and adaptive LASSO methods involve convex optimization while other methods

involve non convex optimization. Since convex optimization is faster, LASSO and adaptive LASSO methods deliver the results faster. Best subset method is very time consuming, due to its discrete nature. Subtle method is the fastest non convex method.

5.1 Robustness of subtle uprooting with "a"

The most attractive property of subtle uprooting against other methods is that it does not require tuning of parameters. This can greatly save the resources one should allocate for the analysis. In this section, the effect of value of a on the performance of subtle uprooting method is studied.

For subtle uprooting method, it is recommended to use $a = 50$. However it is important to confirm that, the parameter a does not need any tuning. For this task a simple simulation is done. 300 data sets are generated with sample size of 300. In the simulation setting $\boldsymbol{\beta} = (-0.7, -0.7, 0, 0, 0, -0.7, 0, 0, 0)^T$, emphirical censoring rate $\approx 25\%$, correlation $= 0.5$ are used. Variable selection and estimation is done using subtle uprooting method for different values for a. Here $a = (20, 30, 40, 50, 60, 70, 80, 90, 100)$ are used.

Table 5.2 summarizes the mean values for model statistics for different values of a. From the table 5.2 WE can observe that, the model statistics are less sensitive to the choice of a. Although the value of a changes in a large range (from 20 to 100), all the model statistics statistics show a very small or no change in their values. Since the model error for different a values does not change from each other, we can argue that the estimation results may not differ considerably. The values for correct selection does not differ greatly either. This helps to conclude that, the variable selection is not greatly affected by different values of a.

Table 5.2: Model Statistics for different a values

a	ME	SIZE	Under	Over	correction selection
20	0.0264	3.4233	0	0.3333	0.6667
30	0.0258	3.3200	0	0.2600	0.7400
40	0.0256	3.2667	0	0.2200	0.7800
50	0.0257	3.2433	0	0.2033	0.7967
60	0.0261	3.2367	0	0.2000	0.8000
70	0.0266	3.2367	0	0.2000	0.8000
80	0.0271	3.2400	0	0.2000	0.8000
90	0.0272	3.2233	0	0.1833	0.8167
100	0.0278	3.2367	0	0.1933	0.8067

Table 5.3 summarizes the median estimates for subtle uprooting for different values of a. Table shows that the median values of the estimates are not or very less affected by the choice of a. Table 5.2 and 5.3 shows the robustness of the subtle uprooting method against the value of parameter a.

Table 5.3: Model Statistics for different a values

a	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$
20	-0.7172	-0.7011	0	0	0	-0.7045	0	0	0
30	-0.7169	-0.7017	0	0	0	-0.7046	0	0	0
40	-0.7169	-0.7017	0	0	0	-0.7020	0	0	0
50	-0.7169	-0.7017	0	0	0	-0.7020	0	0	0
60	-0.7169	-0.7009	0	0	0	-0.7046	0	0	0
70	-0.7177	-0.7008	0	0	0	-0.7043	0	0	0
80	-0.7177	-0.7001	0	0	0	-0.7043	0	0	0
90	-0.7186	-0.7000	0	0	0	-0.7024	0	0	0
100	-0.7189	-0.6994	0	0	0	-0.6996	0	0	0

5.2 Robustness of subtle uprooting with "ϵ"

This section analyses the robustness of subtle uprooting method with ϵ. In chapter 3, it is recommended to use $\epsilon = 1 \times 10^{-4}$. For this task, 300 data sets are generated with sample size of 300. In the simulation setting $\boldsymbol{\beta} = (-0.7, -0.7, 0, 0, 0, -0.7, 0, 0, 0)^T$, emphirical censoring rate $\approx$ 25%, correlation = 0.5 are used. Variable selection and estimation is done using subtle uprooting method considering different values for ϵ. Here $\epsilon = (0, 10^{-2}, 10^{-4}, 10^{-6})$ are used.

Table 5.4: Model Statistics for different ϵ values

ϵ	ME	SIZE	Under	Over	correction selection
0	0.0253	3.2600	0	0.2233	0.7767
0.01	0.0250	3.2600	0	0.2333	0.7667
1.00E-04	0.0253	3.2533	0	0.2233	0.7767
1.00E-06	0.0253	3.2567	0	0.2200	0.7800

Table 5.4 summarizes the mean values of model statistics for different values of ϵ. All the model statistics are not or very less sensitive to the choice of ϵ. Therefore the parameter ϵ in subtle uprooting, does not require tuning.

Table 5.5: Median estimates for different ϵ values

ϵ	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$
0	-0.7097	-0.7056	0	0	0	-0.7056	0	0	0
0.01	-0.7096	-0.7067	0	0	0	-0.7045	0	0	0
1.00E-04	-0.7097	-0.7056	0	0	0	-0.7053	0	0	0
1.00E-06	-0.7098	-0.7056	0	0	0	-0.7053	0	0	0

Table 5.5 summarizes the median values of parameter estimates for different values of ϵ. Median estimate of each parameter is approximately the same over four different values of ϵ. Therefore the parameter ϵ does not need to be tuned for subtle uprooting.

5.3 Performance of LASSO and adaptive LASSO methods

There are several methods available to select the constraint parameter for LASSO and adaptive LASSO methods. For example in glmnet package, we can use the value of lambda which minimizes the mean of cross validation error or the largest value of lambda such that error is within 1 standard error of the minimum. Results may change depending on which method is employed. In this case, we have used is the value of lambda, which minimizes the mean cross validation error. Under this choice of λ, they fail to perform well. However performance of LASSO and adaptive LASSO may improve if the largest value of lambda such that error is within 1 standard error of the minimum is used.

5.4 Summary

Although subset selection methods and shrinkage selection methods suggested in the context of Cox PH model, they have their own strengths and weakness as discussed in chapter 2. Subset selection methods become infeasible in higher dimensions and the available shrinkage methods need tuning of parameters, making the approach expensive and time consuming. This fact motivates to suggest a method, that performs variable selection and estimation simultaneously without further tuning of parameters. Hence the 'Subtle uprooting' is suggested for Cox PH models. The strengths of both subset selection and shrinkage methods are capitalized in Subtle uprooting. It is related to BIC criteria since $logK$ is used as a parameter.

BIC criteria works best with strong signals, higher sample sizes and lower censoring rates. Therefore in the simulation results, subtle uprooting performs well in this setting. In all the settings subtle uprooting method outperforms SCAD, LASSO and adaptive LASSO methods in the higher sample sizes. In many situations best subset selection produces best performance, subtle method produces very close results.

Overall, best subset selection, subtle uprooting and SCAD performs better than LASSO and adaptive LASSO in both variable selection and estimation. In the data example result, subtle uprooting gives approximately closer values to best subset method. Furthermore, under all simulation settings, subtle method converges to best subset selection when the sample size increases. As discussed in chapter 2, best subset selection suffers from its discrete nature and becomes infeasible in higher dimensions. Hence, subtle method can be used as an approximate for best subset selection. This will allow the researcher to obtain better results with less resources.

5.5 Future Work

Simulation study suggested that subtle uprooting method perform well in the context of Cox PH models. Furthermore the results for subtle method and best subset methods become very close with the increasing sample size. Therefore subtle uprooting method deserves future work. Covariates may change values after the study starts and are known as time dependent covariates. Time dependent covariates is very common in real world data sets. It is extremely important to make this distinction since the methods of analyses differ substantially for time dependent covariates. Hence it is very useful to extend the method for time dependent covariates. In this case $\mathbf{x}_i$ will get replaced with $\mathbf{x}_i(t)$ and the partial likelihood need to be altered according to that. There fore the Cox PH model can be written as,

$$\lambda_i(t|\mathbf{x}(t)) = \lambda_0(t)\, \exp(\boldsymbol{\beta}^T \mathbf{x}_i(t)) \tag{5.1}$$

Sometimes, researcher may have to stratify on a covariate. Stratification fits a different baseline hazard function for each stratum. This is commonly used when the proportional hazard assumption is violated for a covariate. It will be interesting to apply this method to such a situation. Simulation study shows that, although subtle method performs fairly with correlated data, there is a reduction of the performance when the correlation is intro-

duced. Therefore it is another important extension to address. This will help to improve the method so that it can handle correlated data and provide better variable selection and estimation.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on, 19* (6), 716-723.

Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics, 6*(1), 76-90.

Cox, D. R. (1972). Regression models and life tables. *JR stat soc B, 34* (2), 187-220.

Cox, D. R. (1975). Partial likelihood. *Biometrika, 62* (2), 269-276.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of statistics, 32*(2), 407-499.

Fan, J., and Li, R. (2002). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96* (456), 1348-1360.

Fan, J., and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics, 30* (1), 74-99.

Furnival, G. M., and Wilson, R. W. (1974). Regressions by leaps and bounds. *Technometrics, 16* (4), 499-511.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics, 7*(3), 397-416.

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.

Kim, J., Kim, Y., and Kim, Y. (2008). A gradient-based optimization algorithm for lasso. *Journal of Computational and Graphical Statistics, 17*(4).

Li, D. H., and Fukushima, M. (2001). On the global convergence of the BFGS method for nonconvex unconstrained optimization problems. *SIAM Journal on Optimization, 11*(4), 1054-1064.

Meinshausen, N., and Bhlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 1436-1462.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics, 6* (2), 461-464.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for Coxs proportional hazards model via coordinate descent. *Journal of Statistical Software, 39*(5), 1-13.

Sohn, I., Kim, J., Jung, S. H., and Park, C. (2009). Gradient lasso for Cox proportional hazards model. *Bioinformatics, 25*(14), 1775-1781.

Su, X. (2013). Variable Selection via Subtle Uprooting. Under revision, Journal of Computational and Graphical Statistics.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (*Methodological*), 267-288.

Tibshirani, R. (1996). The lasso method for variable selection in the Cox model. *Statistics in medicine, 16*,(4), 385-395.

Volinsky, C. T., and Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics, 56*(1), 256-262.

Wu, Y., and Liu, Y. (2009). Variable selection in quantile regression. *Statistica Sinica, 19*(2), 801.

Zhang, H. H., and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika, 94*(3), 691-703.

Zou, H., and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics, 36*(4), 1509.

Appendix A

```
#loading required packages
library(MASS)
library(survival)
library(survival)
library(SIS)
library(matrixStats)
library(glmnet)
# ====================================
# FUNCTION rdat() GENERATES DATA SETS
# ====================================
rdat <- function(n=50, beta =c(-0.35, -.35, 0, 0, 0, -.35, 0, 0, 0),
rho=.5, censor.control=1)
{
p <- length(beta)
# GENERATE X
mu <- rep(0, p)
S <- matrix(1, p, p)
for (i in 1:p){
for (j in 1:p){
S[i, j] <- rho^(abs(i-j))
}
}
# print(S)
```

```
X <- mvrnorm(n = n, mu=mu, Sigma=S, tol=1e-6, empirical=F)

# GENERATE SURVIVAL TIME AND CENSORING TIME
rate <- exp(X%*%beta)
T0 <- rexp(n, rate)
C0 <- rexp(n, rate)
time <- pmin(T0, (C0*censor.control))
status <- sign(T0 <= (C0*censor.control))

# OUTPUT THE DATA
dat <- data.frame(cbind(id=1:n, time, status, X, T0, C0))
colnames(dat) <- c("id", "time", "status", paste("x", 1:p, sep=""),
"true.time", "true.cencor")
return(list(dat=dat, beta.true=beta, S=S))
}
# ====================================================
# FUNCTION install() INSTALLS PACKAGES AFTER CHECKING
# ====================================================
install <- function(pkgs, repos="http://cran.us.r-project.org"){
n.pkgs <- length(pkgs)
check <- !(is.element(pkgs, installed.packages()[,1]))
pkgs0 <- pkgs[check]
if (length(pkgs0)>1) install.packages(pkgs0, repos=repos)
}
# ============================
# OBTAINING THE PARTIAL LOGLIK
# ============================
control0 <- coxph.control(iter.max = 0.5, outer.max = 0.5)
```

```
loglik.pen <- function(beta, time, status, X,
lambda, epsilon=0, a, details=F)
{
# THE PENALTY
beta.eps <- sign(beta)* pmax(abs(beta)-epsilon, 0)
w <- tanh(a*beta.eps^2)
beta.prime <- beta*w

# LOG-LIKELIHOOD
if (details) print(X%*%beta.prime)
L <- coxph(formula=Surv(time, status)~ offset(X%*%beta.prime),
control=control0)$loglik

# THE OBJECTIVE FUNCTION
L.pen <- -2*L + lambda* sum(w)
return(L.pen)
}
#=========================================================================
# MODIFIED BFGS FOR NONCONVEX MINIMIZATION (LI AND Fukushima, 2001 JCAM)
#=========================================================================
#
# REFERENCES:
# D.H. Li and M. Fukushima, On the global convergence of the BFGS method
#for nonconvex unconstrained optimization problems,SIAM J.Optim.,11 (2001),
#1054-1064.
#=========================================================================
```

```
# ====================================================================
# FUNCTION MBFGS() IMPLEMENTS MODIFIED BFGS (LI AND FUKUSHIMA, 2001)
# ====================================================================
norm2 <- function(x) sqrt(sum(x^2))  # EUCLIDEAN NORM

MBFGS <- function(beta0, fn, gr=NULL, ...,
c0=.4, s0=0.5, B0=NULL, details=F,
nrun.max=100,  epsilon=.Machine$double.eps^0.25)
{
f <- function(beta) fn(beta, ...)
# COMPUTE THE GRADIENT EITHER NUMERICALLY OR USING FORMULA IF GIVEN
if (!is.null(gr)) {gradient <- function(beta) gr(beta, ...)}
else {
# install.packages("numDeriv")
require(numDeriv); # help(package="numDeriv")
gradient <- function(beta) grad(func=f, x=beta, method="Richardson")
}
p <- length(beta0)
k <- 0
beta.k <- beta0
g.k <- gradient(beta.k)
if (is.null(B0)) B.k <- diag(p)
else B.k <- B0
while (k < nrun.max && max(abs(g.k)) > epsilon) {
if (details) print(cbind(k=k, f=f(beta.k)))
if (details) print(beta.k)
d.k <- solve(a=B.k, b=-g.k)
i <- 1
```

```
s.i <- s0^i
while (f(beta.k + s.i*d.k) > f(beta.k) + c0 *s.i*sum(g.k*d.k)){
s.i <- s0^i
i <- i+1
}
# if (details) print(paste("Steps in Backtracking:", i, sep=""))
beta.diff <- s.i*d.k
g.diff <- gradient(beta.k + s.i*d.k) - g.k
r.k <- 1 + max(0, - sum(g.diff*beta.diff)/norm2(beta.diff))
u.k <- g.diff + r.k* norm2(g.diff)* beta.diff
k <- k+1
C1.tmp <- (B.k%*% beta.diff %*% t(beta.diff) %*% B.k)
c1.tmp <- as.vector(t(beta.diff)%*%B.k %*% beta.diff)
B.k <- B.k - C1.tmp/c1.tmp + u.k%*%t(u.k)/sum(u.k*beta.diff)
beta.k <- beta.k + s.i*d.k
g.k <- gradient(beta.k)
}
convergence <- ifelse(max(abs(g.k)) > epsilon, FALSE, TRUE)
out <- as.list(NULL)
out$f.min <- f(beta.k); out$beta.min <- beta.k
out$nrun <- k-1; out$convergence <- convergence
return(out)
}
# ======================================
# Method1: VARIABLE SELCTION VIA SUBTLE UPROOTING
# =================1====================
cox.VS <- function(formula=Surv(time, status)~., data,
method.beta0="MLE", beta0=NULL, theta0=1,
```

```
# INTITAL VALUES FOR BETA; theta0 IS THE PENALTY PARAMETER FOR RIDGE REGRESSION
method="BIC", lambda0=2,
# OPTIONS FOR LAMBDA
epsilon0=1e-4, a0=50, SCALE=T,
maxit.global = 300,  maxit.local = 100,
rounding.digits = 4,
zero=sqrt(.Machine$double.eps),
details=F)
{
formula0 <- formula

# PREPARE time, status, AND X MATRIX
# -------------------------------------------
if (missing(data))  data <- environment(formula)
     Call <- match.call()
      indx <- match(c("formula", "data"), names(Call), nomatch = 0)

     temp <- Call[c(1, indx)]
     temp[[1]] <- as.name("model.frame")

     if (is.R()) m <- eval(temp, parent.frame())
     else m <- eval(temp, sys.parent())
     Terms <- terms(m)

var_names <- all.vars(formula) # extract the variable names from the fomula
time <- data[,var_names[1]]
status <- data[,var_names[2]]
```

```
     attr(Terms, "intercept") <- 1
     X <- model.matrix(Terms, m)

if (is.R()) {
       assign <- lapply(attrassign(X, Terms)[-1], function(x) x - 1)
         xlevels <- .getXlevels(Terms, m)
     }
     else {
        assign <- lapply(attr(X, "assign")[-1], function(x) x -1)
        xvars <- as.character(attr(Terms, "variables"))
        xvars <- xvars[-attr(Terms, "response")]
        if (length(xvars) > 0) {
            xlevels <- lapply(m[xvars], levels)
            xlevels <- xlevels[!unlist(lapply(xlevels, is.null))]
            if (length(xlevels) == 0)
                xlevels <- NULL
        }
        else xlevels <- NULL
}
     X <- X[, -1, drop = F]
Xnames <- colnames(X)
p <- NCOL(X); n <- NROW(X)
if (details) print(head(X))

X <- X[order(time), ]
status <- status[order(time)]
time <- sort(time)
```

```
# SCALE THE DATA OR NOT
if (SCALE) X <- scale(X, center = TRUE, scale = TRUE)

# OBTAIN INITIAL VALUES FOR BETA
# ----------------------------------
if (method.beta0=="MLE") {beta0 <- coxph(formula0, data=data)$coef}
else if (method.beta0=="ridge") {
form <- as.formula(paste("Surv(time, status) ~ ridge(",
c(paste(Xnames, ", ", sep="", collapse="")), "theta=", theta0, ")", sep=""))
beta0 <- as.vector(coxph(form, data=data)$coef)
}
else if (!is.null(beta0)) beta0 <- beta0
else beta0 <- rep(0, p)
if (details) print(beta0)

# COMPUTE LAMBDA
# --------------------
if (method=="AIC") lambda <- 2
else if (method=="BIC") { K <- sum(status==1); lambda <- log(K)}
else if(!is.null(lambda0)) lambda <- lambda0
else stop("Please provide a value for lambda.")

# OPTIMIZATION OF THE OBJECTIVE FUNCTION
# ----------------------------------------
# THE PENALIZED PARTIAL LOG-LIKELIHOOD AND GRADIENT
fun <- loglik.pen
grad <- NULL
```

```
# OPTIMIZATION USING SIMULATED ANNEALING, FOLLOWED BY BFGS
opt.fit1 <- optim(par=beta0, fn=fun, gr = grad,
       method = "SANN", control = list(maxit=maxit.global, trace=F, reltol=zero),
time=time, status=status, X=X, lambda=lambda, epsilon=epsilon0, a=a0, details=F)
beta1 <- opt.fit1$par; #
if (details) print(beta1)
opt.fit2 <- optim(par=beta1, fn=fun, gr = grad,
method = "BFGS", control = list(maxit=maxit.local, trace=F, reltol=zero),
time=time, status=status, X=X, lambda=lambda, epsilon=epsilon0, a=a0, details=F)
beta2 <- opt.fit2$par
if (details) print(beta2)
min.Q <- opt.fit2$value
# CHECK CONVERGENCE
converge <- ifelse(opt.fit2$convergence==0, T, F)
if (!(converge)) {
if (details) print("M-BFGS is used!!")
fit.MBFGS <- MBFGS(beta0=beta2, fn=fun, gr=grad,
c0=.4, s0=0.5, B0=NULL, details=details,
nrun.max=100,  epsilon=zero,
time=time, status=status, X=X, lambda=lambda, epsilon=epsilon0, a=a0, details=F)
beta2 <- fit.MBFGS$beta.min
min.Q <- fit.MBFGS$f.min
if (details) print(fit.MBFGS$convergence)
}

# PREPARE THE OUTPUT
beta.eps <- sign(beta2)*pmax(abs(beta2)-epsilon0, 0)
w.hat <- tanh(a0*beta.eps^2)
```

```
beta.hat <- beta2*(w.hat)
}
# ======================================
# VARIABLE SELCTION VIA best subset
# ===============2=====================
terms_form<-names(beta.true)
formula_coxph<-as.formula(paste(c("Surv(time, status)~", terms_form),
collapse=" + "))
fit_coxph_full <- coxph(formula_coxph, data=dat)
n0<-length(dat$status[dat$status==1])
fit.step <- step(fit_coxph_full, direction="both", k=log(n0),trace=0)#BIC
# ======================================
# VARIABLE SELCTION VIA SCAD
# ===============3=====================
X_scad<-dat[,4:12]
X_scad<-scale(X_scad)
Y_scad<- dat[, 2:3]
time_scad<-Y_scad[,1]
status_scad<-Y_scad[,2]

#BIC method is used for tuning
cox.result1=SIS(data=list(x=X_scad,time=time_scad,status=status_scad),model='cox',
tune.method="BIC",vartype=0)
beta.hat_scad<-cox.result1$SIScoef
# ======================================
# VARIABLE SELCTION VIA LASSO
# ===============4=====================
y_lasso<-cbind(time=dat$time,status=dat$status)
```

```
#glmnet needs data in matrix format

x_lasso<-as.matrix(dat[,4:12])
x_lasso<-scale(x_lasso)

cv.cox_2<-cv.glmnet(x_lasso,y_lasso,family = "cox")
#peform the cross validation with default

lasso_cox_2<- glmnet(x_lasso,y_lasso, family = "cox")
beta.hat_lasso <- coef(lasso_cox_2, s =cv.cox_2$lambda.min)[,1]
#uses lambda which gives the minimum cross validation mean error
# ======================================
# VARIABLE SELCTION VIA ALASSO
# ================5=====================
beta.full.abs <- abs(fit_coxph_full$coef)
beta_diag<-diag(beta.full.abs)
x_adp_lasso<-x_lasso%*%beta_diag
colnames(x_adp_lasso)<-terms_form

#Then apply atandard LASSO to new data
cv.cox_alasso_2<-cv.glmnet(x_adp_lasso,y_lasso,family = "cox")

#peform the cross validation with default
lasso_cox_alasso_2<- glmnet(x_adp_lasso,y_lasso, family = "cox")
beta.hat_adp_lasso <- coef(lasso_cox_alasso_2, s =cv.cox_alasso_2$lambda.min)[,1]
beta.hat_adp_lasso <- beta.hat_adp_lasso*beta.full.abs
# ========================================
# FUNCTION assess() COMPUTES ME, SIZE, ETC.
```

```
# ==========================================
assess <- function(beta.hat, beta.true, S0){
ME <- as.vector(t(beta.hat -beta.true)%*%S0%*%(beta.hat-beta.true))
  size <- sum(beta.hat!=0)
underfitting <- ifelse(sum(beta.hat[which(beta.true !=0)]==0) >0, 1, 0)
overfitting <- ifelse(sum(beta.hat[which(beta.true !=0)]==0)==0 &&
sum(beta.hat[which(beta.true ==0)]!=0) >0, 1, 0)
correct.selection <- ifelse(sum((beta.hat==0)==(beta.true==0))==length(beta.true),
1, 0)
return(c(ME, size, underfitting, overfitting, correct.selection))
}
# ==========================================
# COMPUTES SE
# ==========================================
compute_se<-function(beta,data){
beta_nonzero<- beta[ beta!=0]
terms<-names(beta_nonzero)
formula_se<-as.formula(paste(c("Surv(time, status)~0", terms), collapse=" + "))
fit_sd_2<-coxph(formula_se,
              data=dat,init=beta_nonzero,control=coxph.control(iter.max = 0))
se_beta_hat<-data.frame(se=sqrt(diag(vcov(fit_sd_2))))
se_beta_temp<-data.frame(row.names= c("x1" ,"x2", "x3", "x4" ,"x5" ,"x6" ,
"x7" ,"x8" ,"x9"))
se_beta_hat<-merge(se_beta_temp,se_beta_hat,by="row.names",all.x=T)[,-1]
return(se_beta_hat)
}
```

Curriculum Vita

Chalani Wijayasinghe was born on January 17, 1987. The third daughter of Lekamalage Wijayasinghe and Sunethra Hemamala. She entered The University of Colombo, Sri Lanka, in the Spring of 2006 and graduated in Special degree in Statistics with Computer Science in 2010. Just after the graduation she joined the Department of Statistics at University of Colombo as a temporary instructor. After two years of working, she started her graduate studies at University of Texas at El Paso, as a Statistics Masters graduate student.

In the fall of 2012, she entered the Graduate School of The University of Texas at El Paso. While pursuing a master's degree in Statistics she worked as a Teaching and Research Assistant.

Permanent address: 342/ C, Mahawaththa Road, Himbutana, Mulleriyawa New Town.
Colombo, Sri Lanka