Fall 2010

Grouped variable selection in high dimensional partially linear additive Cox model

Li Liu
University of Iowa

GROUPED VARIABLE SELECTION IN HIGH DIMENSIONAL PARTIALLY LINEAR ADDITIVE COX MODEL

by

Li Liu

An Abstract

Of a thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Biostatistics
in the Graduate College of
The University of Iowa

December 2010

Thesis Supervisor: Professor Jian Huang

ABSTRACT

In the analysis of survival outcome supplemented with both clinical information and high-dimensional gene expression data, traditional Cox proportional hazard model fails to meet some emerging needs in biological research. First, the number of covariates is generally much larger the sample size. Secondly, predicting an outcome with individual gene expressions is inadequate because a gene's expression is regulated by multiple biological processes and functional units. There is a need to understand the impact of changes at a higher level such as molecular function, cellular component, biological process, or pathway. The change at a higher level is usually measured with a set of gene expressions related to the biological process. That is, we need to model the outcome with gene sets as variable groups and the gene sets could be partially overlapped also.

In this thesis work, we investigate the impact of a penalized Cox regression procedure on regularization, parameter estimation, variable group selection, and nonparametric modeling of nonlinear effects with a time-to-event outcome.

We formulate the problem as a partially linear additive Cox model with high-dimensional data. We group genes into gene sets and approximate the nonparametric components by truncated series expansions with B-spline bases. After grouping and approximation, the problem of variable selection becomes that of selecting groups of coefficients in a gene set or in an approximation. We apply the group Lasso to obtain an initial solution path and reduce the dimension of the problem and then update the whole solution path with the adaptive group Lasso. We also propose a generalized group lasso method to provide more freedom in specifying the penalty and excluding covariates from being penalized.

A modified Newton-Raphson method is designed for stable and rapid computation. The core programs are written in the C language. An user-friendly R interface is implemented to perform all the calculations by calling the core programs.

We demonstrate the asymptotic properties of the proposed methods. Simulation studies are carried out to evaluate the finite sample performance of the proposed procedure using several tuning parameter selection methods for choosing the point on the solution path as the final estimator. We also apply the proposed approach on two real data examples.

Abstract Approved: ______________________________
Thesis Supervisor

____
Title and Department

____
Date

GROUPED VARIABLE SELECTION IN HIGH DIMENSIONAL PARTIALLY LINEAR ADDITIVE COX MODEL

by

Li Liu

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Biostatistics
in the Graduate College of
The University of Iowa

December 2010

Thesis Supervisor: Professor Jian Huang

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Li Liu

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Biostatistics at the December 2010 graduation.

Thesis Committee: ______________________________
Jian Huang, Thesis Supervisor

____
Joseph Cavanaugh

____
Kathryn Chaloner

____
Kung-Sik Chan

____
Michael P. Jones

ACKNOWLEDGEMENTS

I am greatly indebted to all the faculty and staff in the the departments of biostatistics and biostatistics for their wonderful teaching, encouragement, guidance, and support in the past. Especially, I would like to thank my advisor Professor Jian Huang for guiding me to the field of high-dimensional data analysis and penalized regression, and for being supportive and flexible while helping me to shape my thesis work. This work would be impossible without his support. I would like to thank Professor Mike Jones for introducing me to the field of survival analysis and for his guidance on collaborative data analysis and statistical consulting work with numerous faculty in nursing, health science research, medical science, and epidemiology. I would also like to thank Professor Kung-Sik Chan for introducing me to asymptotic statistics, Professor Joe Cavanaugh for introducing me to categorical analysis and model selection, Professor Kathryn Chaloner for sharing with me her insights on statistical research. My special thanks go to Terry Kirk for her excellent administrative work and her passion with students. They are all my role models in various aspects of life.

I would also like to thank my husband Ye Liu for encouraging and supporting me to pursue what I want to do, and my family members for their understanding and support.

ABSTRACT

In the analysis of survival outcome supplemented with both clinical information and high-dimensional gene expression data, traditional Cox proportional hazard model fails to meet some emerging needs in biological research. First, the number of covariates is generally much larger the sample size. Secondly, predicting an outcome with individual gene expressions is inadequate because a gene's expression is regulated by multiple biological processes and functional units. There is a need to understand the impact of changes at a higher level such as molecular function, cellular component, biological process, or pathway. The change at a higher level is usually measured with a set of gene expressions related to the biological process. That is, we need to model the outcome with gene sets as variable groups and the gene sets could be partially overlapped also.

In this thesis work, we investigate the impact of a penalized Cox regression procedure on regularization, parameter estimation, variable group selection, and nonparametric modeling of nonlinear effects with a time-to-event outcome.

We formulate the problem as a partially linear additive Cox model with high-dimensional data. We group genes into gene sets and approximate the nonparametric components by truncated series expansions with B-spline bases. After grouping and approximation, the problem of variable selection becomes that of selecting groups of coefficients in a gene set or in an approximation. We apply the group Lasso to obtain an initial solution path and reduce the dimension of the problem and then update the whole solution path with the adaptive group Lasso. We also propose a generalized group lasso method to provide more freedom in specifying the penalty and excluding covariates from being penalized.

A modified Newton-Raphson method is designed for stable and rapid computation. The core programs are written in the C language. An user-friendly R interface is implemented to perform all the calculations by calling the core programs.

We demonstrate the asymptotic properties of the proposed methods. Simulation studies are carried out to evaluate the finite sample performance of the proposed procedure using several tuning parameter selection methods for choosing the point on the solution path as the final estimator. We also apply the proposed approach on two real data examples.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1
INTRODUCTION

High-throughput gene expression profiling technologies generate an enormous amount of data. High-dimensional gene expression data are increasingly used for modeling various clinical outcomes to facilitate disease diagnosis, disease prognosis, and prediction of treatment outcome(Yasrebi et al. 2009).

Modeling gene expression data imposes a few challenges onto the traditional statistical methods. First, the number of covariates is generally much larger than the sample size. Secondly, predicting an outcome with individual gene expressions is inadequate because a gene's expression is regulated by multiple biological processes and functional units. There is a need to understand the impact of changes at a higher level such as molecular function, cellular component, biological process, or pathway. The change at a higher level is usually measured with a set of gene expressions related to the biological process. That is, we need to model the outcome with gene sets as variable groups and the gene sets could be partially overlapped also. Thirdly, some clinical effects are known to be nonlinear.

We formulate this gene expression problem with a time-to-event outcome into a high-dimensional partially linear additive Cox model. Penalized regression methods are effective ways to deal with high-dimensional data and have been applied successfully to parametric models such as linear model and logistic regression.

This chapter first describes a gene expression dataset collected in two independent clinical trials, then introduces the Cox PH model, partially linear additive

Cox model, penalized regression, and related work in the literature.

1.1 Motivating example and test dataset

The thesis work is motivated by gene expression data. Norway/Stanford breast tumor data is a gene expression dataset with survival outcome and supplemental clinical information such as age at diagnosis and tumor stage. In Chapter 5, the procedure developed in this thesis work is applied to this dataset to illustrate the capacity of the method. PBC data is a relatively small dataset provided in the R package *survival.* This dataset is used to debug the computational implementation of the method and to illustrate the various concepts throughout the thesis.

1.1.1 Norway/Stanford breast tumors data

A total of 115 advanced breast cancer patients were enrolled into two independent studies as described in Geisler et al. (2001) and Geisler et al. (2003). The dataset includes 12,793 gene features and a set of clinical information such as survival outcome, age at diagnosis, Estrogen receptor status, tumor grade, and so on.

We group the gene features into gene oncology (GO) related biological processes to facilitate the modeling of the survival outcome. 3,776 of the gene features mapped into 825 GO biological processes, i.e., gene sets. The remaining 9,017 gene features cannot be mapped into any gene set and thus are not excluded from the statistical model in this thesis work. Each gene set includes multiple gene features

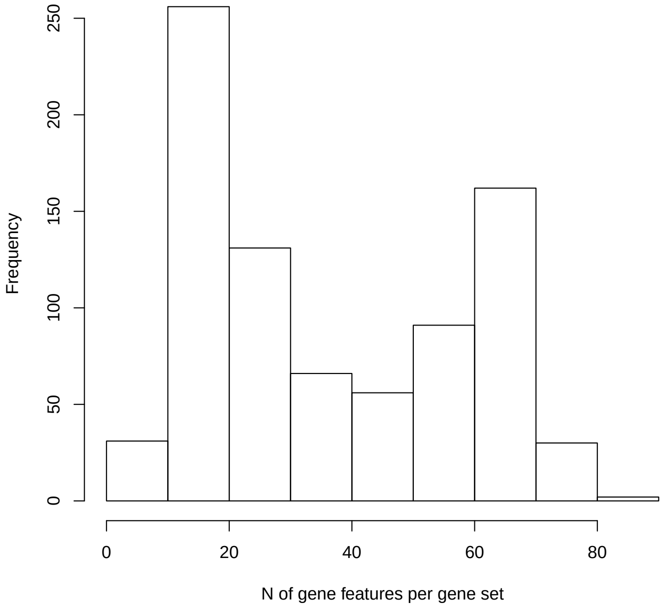

Figure 1.1: Histogram of gene sets by the number of gene features that a gene set includes.

as shown by the histogram in 1.1. Each gene feature can be mapped into multiple gene sets as shown in table 1.1. As a result, there are 30,388 gene features in total in the 825 gene sets.

The dataset can be extracted from the R package *survJamDa.data* or downloaded from the National Center for Biotechnology Information (NCBI) website. R package *GSA* is used to map the gene features into GO gene sets.

1.1.2 PBC data

Mayo clinic conducted a placebo controlled trial on the theraputical effect of the drug D-penicillamine on patients with primary biliary cirrhosis (PBC) of liver from 1974 to 1984. 312 PBC patients were randomized into the study. The measurements listed in Table 1.2 are available for all 312 participants without missing values. 185 were censored by the end of the study.

Except the 2 outcome variables, *time* and *status*, the remaining 14 variables in Table 1.2 are included as covariates in a penalized Cox model. According to Aalen et al. (2007), most of the continuous variables have nonlinear effects. A B-spline with 6 degrees of freedom is used to approximate the nonparametric effects of the 6 continuous variables, *age*, *alb*, *alkphos*, *bili*, *protime*, and *sgot*. The 6 columns of the B-spline basis matrix corresponding to each continuous variable are grouped together. Categorical variables, *stage* and *edtrt*, have more than two levels and their respective dummy variables are grouped together. Dichotomous variables, *ascites*, *edema*, *hepmeg*, *sex*, *spiders*, and *trt*, are one variable per group.

Table 1.1: Frequency table of gene features by the number of gene sets that a gene feature is mapped into.

Gene sets	N(%) of gene features
1	77 (21)
2	542 (14)
3	451 (12)
4	298 (8)
5	260 (7)
6	215 (6)
7	156 (4)
8	130 (3)
9	124 (3)
10	92 (2)
11-20	453 (12)
21-30	148 (4)
31-100	115 (3)
101-800	15 (0)

Table 1.2: Description of variables in the PBC data.

Variable name	Description
age	in years
alb	serum albumin
alkphos	alkaline phosphotase
ascites	presence of ascites
bili	serum bilirunbin
edema	presence of edema
edtrt	0 no edema, 0.5 untreated or successfully treated
	1 unsuccessfully treated edema
hepmeg	enlarged liver
time	number of days between registration and the earlier of death
	transplantion, or study analysis in July, 1986
protime	standardised blood clotting time
sex	1 = male
sgot	liver enzyme (now called AST)
spiders	blood vessel malformations in the skin
stage	histologic stage of disease (needs biopsy)
status	status at endpoint, 0/1/2 for censored, transplant, dead
trt	1 for control and 2 for treatment

The penalized model with this PBC dataset is described in the computational implementation section in Chapter 2. All the illustrative plots throughout the thesis report are based on this dataset.

1.2 Cox proportional hazard (PH) model

Let $X = (x_1, x_2, \cdots, x_p)'$ be a vector of p covariates measured at or before time 0 on individuals under a study, and let T be a corresponding absolutely continuous failure time. In the Cox PH model, the hazard function is assumed to be (Kalbfleisch and Prentice 2002)

$$\begin{aligned} \lambda(t|X) &= \lim_{h \to 0^+} P(t \leq T < t + h \mid T \geq t, X)/h \\ &= \lambda_0(t)\exp(X'\beta) \qquad (1.1) \end{aligned}$$

where λ_0 is an arbitrary unspecified baseline hazard function and β is a vector of regression coefficients of length p.

Suppose there are n observations in a study and the i^{th} observation is (Y_i, X_i, Δ_i). $Y_i = \min(T_i, C_i)$ and $\Delta_i = 1(T_i \leq C_i)$, where T_i is the survival time or failure time, C_i is the censoring time, and X_i is a vector of p covariates. Let $t_1 < t_2 < \cdots < t_k$ be the k distinct failure times and assume that there are no tied failure times, under the independent right censoring assumption and Cox proportional hazard assumption, the partial likelihood for β is

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(X_i'\beta)}{\sum_{j \in R(t_i)} \exp(X_j'\beta)} \right\}^{\Delta_i}$$

where $R(t) = \{k; Y_k \geq t\}$ is the set of items at risk of failure at time t^-, just prior to time t.

When there are ties among the failure times, the partial likelihood can be adjusted in various ways. Let $D(t) = \{k; Y_k = t \text{ and } \Delta_k = 1\}$ be the set of tied failures at time t and suppose there are d_j failures at time t_j. Efron(1997) suggested the following correction for ties

$$L(\beta) = \prod_{i=1}^{k} \frac{\prod_{l \in D(t_i)} \exp(X_l'\beta)}{\prod_{r=0}^{d_i-1}\{\sum_{j \in R(t_i)} \exp(X_j'\beta) - \frac{r}{d_i}\sum_{l \in D(t_i)} \exp(X_l'\beta)\}}$$

The log partial likelihood function is

$$\begin{aligned} \ell(\beta) &= \log\{L(\beta)\} \\ &= \sum_{i=1}^{k}\Big\{\sum_{l \in D(t_i)} X_l'\beta - \sum_{r=0}^{d_i-1}\log\Big\{\sum_{j \in R(t_i)} \exp(X_j'\beta) - \\ &\quad \frac{r}{d_i}\sum_{l \in D(t_i)} \exp(X_l'\beta)\Big\}\Big\} \end{aligned} \tag{1.2}$$

1.3 Partially linear additive Cox model

The partially linear additive Cox model is an extension of the linear additive Cox model which allows flexible modeling of covariate effects semiparametrically (Huang 1999). Assume $Z = (X, W) \in R^p \times R^q$ is the set of baseline covariates. In the partially linear additive Cox model, the hazard function takes the following form,

$$\lambda(t|X, W) = \lambda_0(t)\exp(X'\beta + \sum_{i=1}^{q}\phi_i(W_i)) \tag{1.3}$$

where $\phi_1, \cdots, \phi_q$ are smooth functions.

Numerous work has been done to reduce the dimensionality of a functional form such as $\phi_i(\cdot)$. Stone(1986) first proposed using a polynomial spline to study a fully nonparametric additive Cox model. Lu (2007) used splines to approximate the mean function of a counting process and found that the estimation is consistent and the convergence rate is faster than $n^{\frac{1}{3}}$. Huang (1999) studied the estimation efficiency when some nuisance covariates were estimated using splines. He found that although the estimated convergence rate for the spline approximated components is slower than root n, the convergence rate for the components of interest is efficient and in the order of root n.

1.4 LARS algorithm and penalized methods

Efron et al. (2004) described the least angle regression (LARS) algorithm to estimate parameters and to select an optimal model in the classic linear regression setting $Y = X\beta + \epsilon$. The LARS algorithm builds up the estimation $\hat{\mu} = X\hat{\beta}$ in successive steps. In each step, it adds the covariate that has the largest correlation coefficient with the residuals into the model. The regression at each step is proceeded along the equiangular direction defined by the covariates in the model until a new covariate has the same correlation coefficient with the residuals as those covariates already in the model. If there are p covariates in total, only p steps are needed to computer the solution path and the path is piecewise linear.

The LARS algorithm motivated a large amount of research in the penalized regression field in terms of finding the solution path.

Breiman (1995) demonstrated that the traditional subset selections are discontinuous, computation intensive when the number of covariates is large, unsatisfactory in terms of prediction accuracy and stability, and proposed the non-negative garrotte, a penalized method, which is shown to be more accurate and stable.

Tibshirani (1996) proposed a penalization method called lasso. The lasso estimation of β is

$$\hat{\beta}^{LASSO}(\lambda) = \arg\min_{\beta}\{-l(\beta) + \lambda\|\beta\|_1\} \tag{1.4}$$

where λ is a tuning parameter and $\|\cdot\|_1$ is the l_1-norm. Under the l_1 penalty, some elements of β are estimated to be zero to achieve a sparse solution.

Efron et al. (2004) also studied how to modify the LARS algorithm to find the solution path for the lasso penalized linear model. The major modification is to allow a covariate to drop out of the model when the sign of its estimate changes.

Tibshirani (1997) looked into the lasso penalized Cox proportional hazard model and found that this method shrinks the regression coefficients and produces a sparse soltion. Furthermore, his simulation studies demonstrated that the lasso is more accurate than stepwise method in terms of variable selection..

Bakin (1999) proposed a generalized penalty term for selecting prespecified groups of variables, which is called group lasso.

Huang et al.(2006) studied a group bridge penalized method which is capable of both bi-level variable selection in the linear regression setting. Huang et al. (2010) studied the adaptive group lasso method in nonparametric additive models.

Xie (2007) studied the property of the smoothly clipped absolute deviation penalty imposed on the linear regression model and on the accelerated failure time model and demonstrated the oracle property achieved by the method.

There are various form of penalties studied by other researchers too. Fan and Li (2001) studied the SCAD penalty. Zou and Hastie (2005) studied regularization and variable selection via the elastic net. The ridge regression is the first penalized method proposed by Hoerl and Kennard (1970).

Yuan and Lin (2006) studied the model selection and estimation properties with grouped variables in linear regression framework. Three methods for grouped variable estimation and selection, the grouped least angle regression, the group lasso, and the group non-negative garrotte, were compared. They concluded that the group lasso enjoys excellent performance but its solution path is not piecewise linear and therefore requires intensive computation in large-scale problems. Meier et al. (2008) studied the group lasso for logistic regression. They proposed the block co-ordinate descent algorithm to solve the problem and claimed that it was efficient.

Nardi, Y. and Rinaldo, A. (2008) established estimation and model selection consistency, prediction and estimation bounds and persistence for the group-lasso estimator and model selector proposed by Yuan and Lin (2006) for least squares problems where the covariates have a natural grouping structure. They considered the case of a fixed-dimensional parameter space with increasing sample size and the double asymptotic scenario where the model complexity changes with the sample size.

Zhang and Lu (2007) studied the adaptive lasso in the Cox proportional hazards model framework. Their adaptive lasso penalty was based on parameter estimates from the regular Cox model and thus is unsuitable for high-dimensional data. Furthermore, the optimal tuning parameter was determined using the generalized cross-validation instead of finding it along a solution path. They also showed that adaptive lasso penalized method has the oracle properties under a proper choice of the tuning parameter.

1.5 Outline of the thesis

In Chapter 2, we study the properties and computational issues in the group lasso penalized partially linear additive Cox model. An algorithm to calculate the solution path is proposed and implemented in R. AIC, BIC, and cross validation (CV) are adapted separately to select an optimal model and its corresponding optimal tuning parameter λ. The implementation works not only in the traditional setting of $p < n$, but also in the high-dimensional setting when $p > n$. Asymptotic results indicate that group lasso estimator is root-n consistent under some regularity conditions. It can reduce the dimension of a Cox regression problem, however, it is inconsistent in terms of variable selection.

Based on the work on the group lasso, we investigate the properties of the adaptive group lasso method on the same Cox regression framework in Chapter 3. Generally speaking, we first apply group lasso on a Cox regression problem to get an initial optimal model with reduced dimension. We then refine the resulting model with adaptive group lasso to improve model fitting and promote sparsity. The method is demonstrated to be not only estimation consistent but also variable selection consistent. That is, it possesses oracle properties.

Both group lasso and adaptive group lasso estimates are biased to trade for sparsity despite they are asymptotically unbiased. There is a need to estimate the unbiased effect of some covariates and to put different penalty on different covariate group. A generalized group lasso method is proposed, studied, and implemented based on the previous work in Chapter 4.

In Chapter 5, a numerical study is conducted to compare the three methods and the generalized group lasso method is applied to analyze a gene expression dataset. Some discussion and potential future work conclude the thesis in Chapter 6.

CHAPTER 2
GROUP LASSO PENALIZED PARTIALLY LINEAR ADDITIVE COX REGRESSION

2.1 Approximation of nonlinear effects in a partially linear additive Cox model

Truncated B-spline approximation is utilized to estimate a nonlinear effect. Assume ϕ is a smooth function defined on $[a, b]$ with $a > -\infty$ and $b < \infty$. Assume there exists a finite set of knots, $\{\eta_i\}_{i=0}^{K}$, such that $a = \eta_0 < \eta_1 < \eta_2 < \cdots < \eta_K = b$. Thus, $\{\eta_i\}_{i=0}^{K}$ partitions $[a, b]$ into K subintervals. Let S be the set of polynomial splines of order $l \geq 1$ defined on $[a, b]$, such that $\forall s \in S$, s is a polynomial of order l in $[\eta_{i-1}, \eta_i]$, $\quad \forall 1 \leq i \leq K$ and s is $l - 2$ times continuously differentiable on $[a, b]$ if $l \geq 2$. According to Schumaker (1981), there exists a set of local basis functions on $[a, b]$, $\{B_t, t = 1, \cdots, K + l\}$, for S. That is, $\forall s \in S$, there exists $\{b_t \in R, t = 1, \cdots, K + l\}$, such that

$$s(\cdot) = \sum_{t=1}^{K+l} b_t B_t(\cdot).$$

With a set of suitably chosen knots $\{\eta_i\}_{i=0}^{K}$ and l for S, and under some suitable smoothness conditions on ϕ, ϕ can be well approximated by a function in S.

A B-spline is a spline function that has minimal support with respect to a given degree, smoothness, and domain partition. By induction, it can be proved that the sum of the B-spline basis functions for a particular value of the covariate is unity. This is known as the partition of unity property of the basis functions

(Boor 2001). Since the intercept part of a Cox model is absorbed in the baseline function, we exclude the last column of the basis matrix for each covariate to avoid the model identification problem. The rationale is

$$\begin{aligned} s(\cdot) &= \sum_{t=1}^{m+1} b_t B_t(\cdot) \\ &= \sum_{t=1}^{m} b_t B_t(\cdot) + b_{m+1} B_{m+1}(\cdot) \\ &= \sum_{t=1}^{m} b_t B_t(\cdot) + b_{m+1} \left(1 - \sum_{t=1}^{m} B_t(\cdot) \right) \\ &= \sum_{t=1}^{m} (b_t - b_{m+1}) B_t(\cdot) + b_{m+1} \\ &= \sum_{t=1}^{m} b_t^* B_t(\cdot) + b_{m+1} \end{aligned}$$

where $m = K + l - 1$. That is, we can use all the B-spline basis functions except the last one to approximate a covariate's nonlinear effect.

As to the partially linear additive Cox model introduced above, we need to find a set of suitable basis $\{B_{j,t} \in R, t = 1, \cdots, m_j\}$ and the corresponding coefficients $\{b_{j,t} \in R, t = 1, \cdots, m_j\}$ to approximate each functional form $\phi_j(\cdot)$.

Let $W_{j,b} = [B_{j,1}(W_j), B_{j,2}(W_j), \cdots, B_{j,m}(W_j)]$ be the basis matrix of W_j, $W_b = [W_{1,b}, W_{2,b}, \cdots, W_{q,b}]$ be the basis matrix of W, $b = \{b_{j,k}; 1 < k < m_j, 1 < j < q\}$, $Z = (X, W_b)$, and $\theta = (\beta, b)$, then the partial likelihood for the partially linear additive Cox model can be approximated by

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(Z_i'\theta)}{\sum_{j \in R(t_i)} \exp(Z_j'\theta)} \right\}^{\Delta_i}$$

This is in the same form as the partial likelihood for the Cox PH model. We can incorporate Efron's method on top of this approximation to deal with tied failure times. In the remaining of this thesis, we still use X to represent the complete set of covariates instead of Z

Partially linear additive Cox model is equivalent in form to a regular Cox PH model once the basis matrix for a nonlinear effect to be modeled nonparametrically is generated. R function *bs* can be used to generate the basis matrix for a covariate.

However, the set of coefficients corresponding to the basis matrix of a nonlinear effect needs to be grouped together for the purpose of variable selection. This kind of grouping is treated the same way as the grouping of genes into a gene set in this thesis work.

2.2 Group lasso penalized Cox regression

In group lasso penalized regression, the objective is to minimize the following function over β

$$\begin{gathered} Q_n(\beta, \lambda) = -\frac{1}{n}\ell_n(\beta) + \lambda \sum_{j=1}^{K} \sqrt{p_j}\|\beta_{(j)}\| \\ \hat{\beta}_n(\lambda) = \arg\min_{\beta} Q_n(\beta, \lambda) \end{gathered} \tag{2.1}$$

where $Q_n(\beta, \lambda)$ is the objective function and $\ell_n(\beta)$ is the partial log likelihood from the partially linear additive Cox model as introduced in the previous section.

The covariates are grouped into K groups. $\beta_{(j)}$ is the regression coefficient for covariate group j, where $j = 1, \cdots, K$. Thus, $\beta = (\beta'_{(1)}, \cdots, \beta'_{(K)})'$. $\| \cdot \|$ is the L_2 norm.

λ is the tuning parameter. Intuitively, it controls the trade-off between the estimation bias and the estimation variance. In a traditional Cox regression setting with $n > p$, if $\lambda = 0$, then there is no estimation bias and the estimation variance is relative large. As λ increases, the estimation bias increases and the estimation variance decreases. When λ is large enough, all the parameter estimates are zeros and their variances are zeros also.

The minimization problem in (2.1) is the Lagrangian form of the following optimization problem,

$$\begin{aligned} &\min_{\beta} -\ell_n(\beta) \\ s.t. \quad &\sum_{j=1}^{K} \sqrt{p_j}\|\beta_{(j)}\| \leq t \end{aligned} \tag{2.2}$$

Where t is the upper bound of the constraints on β.

$-n^{-1}\ddot{\ell}_n(\beta)$ is symmetric and thus nonnegative definite. It is positive definite when $n > p$ or when the covariates are linearly independent. $\|\beta_{(j)}\|$ is convex $\forall j \in \{1, 2, \cdots, K\}$, thus $\sum_{j=1}^{K} \sqrt{p_j}\|\beta_{(j)}\|$ is convex. As a result, the objective function in (2.1) to be minimized is strictly convex and finding the group lasso estimation in (2.1) is a strictly convex optimization problem. As a result, any local optimal point is also the global optimal point (Boyd and Vandenberghe 2004).

2.3 KKT conditions

The Karush-Kuhn-Tucker (KKT) conditions for finding the minimizer in (2.1) are

$$\begin{cases} -\frac{1}{n}\dot{\ell}_{n(j)}(\beta) + \frac{\lambda\sqrt{p_j}\beta_{(j)}}{\|\beta_{(j)}\|} = 0 & \forall \beta_{(j)} \neq 0 \\ \|\frac{1}{n}\dot{\ell}_{n(j)}(\beta)\| \leq \lambda\sqrt{p_j} & \forall \beta_{(j)} = 0 \end{cases} \tag{2.3}$$

where $\dot{\ell}_{n(j)}(\beta)$ denotes $\partial\ell_n(\beta)/\partial\beta_{(j)}$.

The KKT conditions are not only necessary but also sufficient. That is, the solution to the conditions in (2.3) is the minimizer for both (2.1) and (2.2). We call the collection of variable groups corresponding to those $\hat{\beta}_{n(j)} \neq 0$ the active set A and that corresponding to $\hat{\beta}_{n(j)} = 0$ the non-active set $\bar{A}$.

The above KKT conditions provide a lot of information about the group lasso estimates. In order for a covariate group to have nonzero estimate, i.e., to be in the active set A, it has to meet the second KKT condition. This explains why the group lasso can produce zero estimates for some covariate groups and thus reduce the dimensionality of a regression problem. The first KKT condition implies $\|\dot{\ell}_{n(j)}(\hat{\beta}_n)\|/\sqrt{p_j} = \lambda, \quad \forall\hat{\beta}_{n(j)} \neq 0$. The first KKT condition allows us to find the penalized estimates once the active set is known.

If $p < n$ and the set of covariates has full column rank, we can define

$$\begin{aligned}
\lambda_{max} &= \inf\{\lambda; \quad \hat{\beta}_n(\lambda) = 0\} \\
&= \max_j \frac{\|\dot{\ell}_{n(j)}(0)\|}{n\sqrt{p_j}} \\
t_{max} &= \sum_{j=1}^{K} \sqrt{p_j}\|\hat{\beta}_{n(j)}(0)\|
\end{aligned}$$

where $\dot{\ell}_{n(j)}(0)$ includes the components of the score function $\dot{\ell}_n(\beta)$ at $\beta = 0$ corresponding to the j^{th} grouped variable.

Figure 2.1 illustrates the relationship between λ and t. As λ decreases from λ_{max} to 0, the penalized estimate for β changes continuously from 0 to the regular Cox model estimate. Furthermore, t increases from 0 to t_{max} when λ decreases

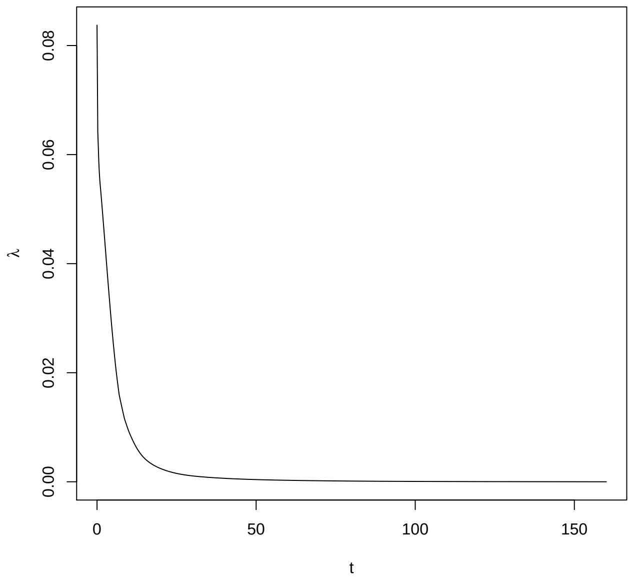

Figure 2.1: Plot of $\lambda - t$ with PBC data.

from λ_{max} to 0. That is, the constraint in (2.2) is tight when $t \in [0, t_{max}]$. Since t is a strictly decreasing function of λ when $\lambda \in [0, \lambda_{max}]$, the β estimate can also be denoted as a function of t, $\hat{\beta}_n(t)$. We call both $\hat{\beta}_n(\lambda)$ where λ decreases from λ_{max} to 0 and $\hat{\beta}_n(t)$ where t increases from 0 to t_{max} the solution path for the group lasso penalized model.

When $p > n$ or the set of covariates does not have full rank, then λ need to be a positive number to obtain a unique parameter estimation.

2.4 Modified Newton-Raphson method

Newton-Raphson method generally has quadratic convergence rate and is preferred when the number of parameters is not too large. When the number of parameters is large, block co-ordinate descent algorithm proposed by Meier et al. (2008) can be implemented to speed up the process. In the loop of updating one block at a time, the algorithm is essentially New-Raphson at each step.

Applying the Newton-Raphson method to our problem requires the calculation of the following first and second order derivatives of the objective function in (2.1) corresponding to the active set A.

$$\begin{aligned} \dot{Q}_{nA}(\beta,\lambda) &= -\frac{1}{n}\dot{\ell}_{nA}(\beta) + \left[\frac{\lambda\beta_1\sqrt{p_1}}{\|\beta_1\|},\cdots,\frac{\lambda\beta_{(j)}\sqrt{p_j}}{\|\beta_{(j)}\|}\right]_A \\ \ddot{Q}_{nA}(\beta,\lambda) &= -\frac{1}{n}\ddot{\ell}_{nA}(\beta) + diag\left(\frac{\lambda\sqrt{p_j}}{\|\beta_{(j)}\|}\left(I_{p_j\times p_j} - \frac{\beta_{(j)}\beta'_{(j)}}{\|\beta_{(j)}\|^2}\right)\right)_A \end{aligned} \tag{2.4}$$

where $(\cdot)_A$ is a subset operation, which selects only the components that are in the active set A.

It is important that $\ddot{Q}_{nA}(\beta,\lambda)$ is positive definite in order to have a unique solution for a give λ. If there is only one variable group in the variable group, then $diag\Big(\frac{\lambda\sqrt{p_j}}{\|\beta_{(j)}\|}\Big(I_{p_j\times p_j} - \frac{\beta_{(j)}\beta'_{(j)}}{\|\beta_{(j)}\|^2}\Big)\Big)_A$ has a rank of $A^* - 1$ with A^* denoting the number of covariates in the active set. The only zero eigenvalue of this matrix has an eigenvector of β_A. If β_A is not an eigenvector of $\ddot{\ell}_{nA}(\beta)$ corresponding to a zero eigenvalue, then $Q_{nA}(\beta,\lambda)$ is positive definite. This condition is easy to meet when the group size is larger then one. With similar argument, we know that it is easy for $Q_{nA}(\beta,\lambda)$ to be positively definite if most of the group sizes are greater than one. This also explains why the number of covariates in the active sets could be larger than the sample size in the gene expression data example described in Chapter 5.

Similar to LARS, the active set grows as λ decreases from λ_{max} to 0. When

$\lambda = \lambda_{max}$ the, active set is empty. If we let λ decreases along a set of fine grid points between λ_{max} and 0, the covariates will enter into the active set gradually according to the criteria specified in the KKT conditions. At the same time, covariate groups might also move out of the active set. When the grid points are densely placed, we expect that the change in the active set and the change in the β estimate are small between two consecutive λ values. Also, the previous β estimate is an excellent initial guess for finding the next β estimate. The active set corresponding to the previous λ value is a close estimate of the current active set. In addition, the second KKT condition provides a criterion to evaluate whether a covariate group should enter into or move out of the active set.

There are a couple of reasons that we need to modify the Newton-Raphson method to make it work in our computational implementation.

The first challenge of applying the Newton-Raphson method to solve the convex minimization problem in (2.2) is that the estimates of some variable groups, could be zero. For those zero estimates, their first and second derivatives do not exist. We don't know the exact set of non-active covariates in advance even if we search along the solution path. During the iterations of finding a solution for a specific λ, the estimate for a non-active covariate group could be approximately zero before it the iteration converges.

The second challenge of applying the Newton-Raphson method is the possible overflow of floating point in computational implementation. The overlow of floating point is mainly caused by extremely small $\|\beta_{(j)}\|$ for $j \in A$. When λ is slightly smaller than λ_{max}, a small change in β or equivalently in t will lead to a dramatic change in λ. This is evident in the plot of the relationship between λ and t in Figure 2.1. That means that some components of the β estimate could be extremely small.

The solution is that we set $\|\beta_{(j)}\| = \epsilon$ whenever $\|\beta_{(j)}\| < \epsilon$ with a prespecified $\epsilon > 0$. This modification also makes sure that any previous β estimate won't cause any problem as an initial guess for finding the next β estimate and that the Newton-Raphson algorithm is robust.

Define

$$\begin{aligned} \dot{Q}^*_{nA}(\beta, \lambda) &= -\frac{1}{n}\dot{\ell}_{nA}(\beta) + \left[\frac{\lambda\beta_1\sqrt{p_1}}{\|\beta_1\|^*}, \cdots, \frac{\lambda\beta_{(j)}\sqrt{p_j}}{\|\beta_{(j)}\|^*}\right]_A \\ \ddot{Q}^*_{nA}(\beta, \lambda) &= -\frac{1}{n}\ddot{\ell}_{nA}(\beta) + diag\left(\frac{\lambda\sqrt{p_j}}{\|\beta_{(j)}\|^*}\left(I_{p_j\times p_j} - \frac{\beta_{(j)}\beta'_{(j)}}{\|\beta_{(j)}\|^{*2}}\right)\right)_A \end{aligned} \tag{2.5}$$

where $\|\beta_{(j)}\|^* = \max(\|\beta_{(j)}\|, 10^{-9})$. $(\cdot)_A$ is a subset operation, which selects only the components that are in the active set A.

After the modification, $\ddot{Q}^*_{nA}(\beta, \lambda)$ is still a symmetric matrix, Cholesky decomposition is used to speed up the calculation of its inverse. If an initial guess of β is $\hat{\beta}^0_n$, then the update of β estimate is

$$\hat{\beta}_{nA}(\lambda) = \hat{\beta}^0_{nA} - \ddot{Q}^{*-1}_{nA}(\hat{\beta}^0_n, \lambda)\dot{Q}^*_{nA}(\hat{\beta}^0_n, \lambda) \tag{2.6}$$

2.5 Algorithm to find the penalized solution path

A set of M grid points $\{\lambda_i\}_{i=0}^M$ equally spaced on the log scale within $[0, \lambda_{max}]$ is chosen to find the penalized solution path. If $n > p$, the last λ value is chosen to be zero to obtain the regular Cox estimate.

$$\begin{cases} \lambda_1 & = \lambda_{max} \\ \lambda_i & = \frac{\lambda_{i-1}}{c} \quad \forall i \in [2, M] \\ \lambda_M & = 0 \qquad \text{if } n > p \end{cases} \tag{2.7}$$

In the above expression, $c > 1$ and c could be modified to space the grid points differently. A small λ value could cause the computation to be unstable when the number of covariates is larger than the sample size.

The following steps are implemented in R to find the solution path:

- Initialize: set $A = \emptyset$, $\hat{\beta}_n = 0$, and $i = 2$.
- (1) Calculate $\|\dot{\ell}_{n(j)}(\hat{\beta}_n)/n\|$ corresponding to each variable group, $j = 1, \cdots, K$.
- For each $j \notin A$, if $\|\dot{\ell}_{n(j)}(\hat{\beta}_n)/n\| > \lambda_i \sqrt{p_j}$, then variable group j is selected into the active set A. That is,

  $$A = A \cup \{j; \quad \|\frac{1}{n}\dot{\ell}_{n(j)}(\hat{\beta}_n)\| > \lambda_i \sqrt{p_j}, j \notin A\}$$

  The parameter estimate for those newly added variable group is set to be zero.
- (2) Update $\hat{\beta}_n$ using the formula in (2.6) until convergence. Note that only the components in the active set A are updated.
- Check for the dropping of variable groups. For each variable group $j \in A$, set $\hat{\beta}_n^0 = \hat{\beta}_n$ and again set $\hat{\beta}_{n(j)}^0 = 0$. If $\|\dot{\ell}_{n(j)}(\hat{\beta}_n^0)/n\| < \lambda_i \sqrt{p_j}$, drop variable

group j from the active set, set $\hat{\beta}_{n(j)} = 0$, and go to step (2). If there is no variable group to be dropped, continue.

- Check for the adding of variable groups. For each variable group $j \notin A$, if $\|\dot{\ell}_{n(j)}(\hat{\beta}_n)/n\| > \lambda_i \sqrt{p_j}$, then add variable group j to the active set and go to step (2). Otherwise continue. If the total number of covariates in the active set is greater than n, stop.

- Save current result and set $i = i + 1$. Go to step (1) until $i > M$.

Figure 2.2 illustrates a group lasso penalized solution path from modeling the PBC data. The model details are described in the computational implementation in this chapter.

2.6 Model selection - find an optimal tuning parameter λ

AIC, BIC, and cross-validation are adapted to find the optimal λ values along the solution path. Since there is a penalty term in the group lasso penalized Cox model, the effective number of parameters is less than the number of parameters in the active set when $\lambda \in (0, \lambda_{max})$. Moody (1992) proposed the following formula to calculate the effective number of parameters in a penalized model.

$$d.f.(\lambda) = trace\{-\frac{1}{n}\ddot{\ell}_{nA}(\hat{\beta}_n(\lambda))\ddot{Q}^{*-1}_{nA}(\hat{\beta}_n(\lambda), \lambda)\}$$

where $\hat{\beta}_n(\lambda)$ is the penalized estimate of β and $\ddot{\ell}_{nA}(\hat{\beta}_n(\lambda))$ is the negative Hessian matrix of the Cox partial likelihood which corresponds to the active set A and is evaluated at $\hat{\beta}_n(\lambda)$. $\ddot{Q}_{nA}(\hat{\beta}_n(\lambda), \lambda)$ is the second derivative of the Lagrangian

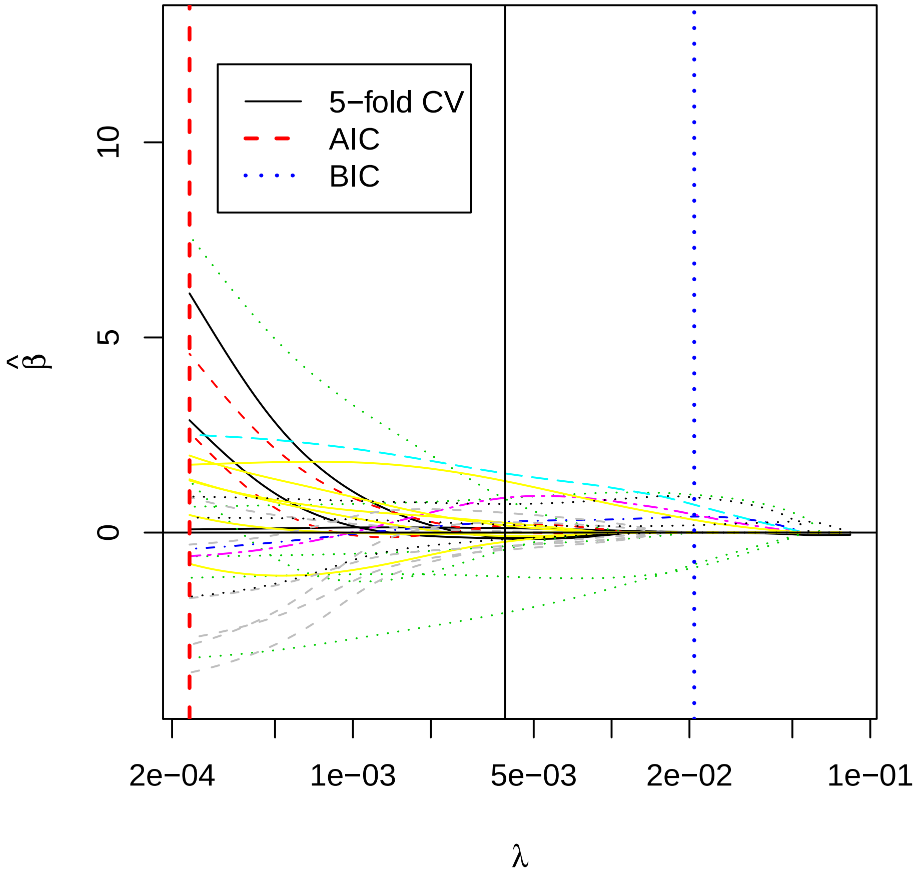

Figure 2.2: Plot of a group lasso penalized solution path with the PBC data. The horizontal axis is the the tuning parameter λ. The vertical lines are optimal models chosen by AIC, BIC, and cross validation. The estimates for each variable group are plotted using the same line type.

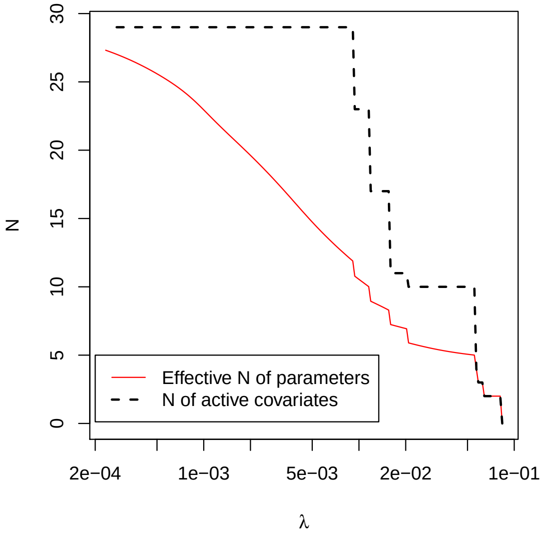

Figure 2.3: Plot of the effective N of parameters and the N of active covariates with the PBC data.

function which corresponds to the active set A and is evaluated at $(\hat{\beta}_n(\lambda), \lambda)$.

Figure 2.3 is a plot of the effective number of parameters and the number of active covariates with the PBC data.

At each λ value along the solution path, the modified variable selection criteria are

$$\begin{aligned} AIC(\lambda) &= -2\ell_n(\hat{\beta}_n(\lambda))/n + 2d.f.(\lambda)/n \\ BIC(\lambda) &= -2\ell_n(\hat{\beta}_n(\lambda))/n + 2d.f.(\lambda)\log(n)/n \end{aligned}$$

k-fold cross validation is also implemented in R to select an optimal λ value. Generally speaking, in a k-fold cross validation, the modeling data is randomly divided into k subsets with approximately equal sample size. Each time, one subset is left out as the test set while the remaining k-1 subsets are used as the training set to estimate β. The cross validation error for the subset is naturally chosen to be the negative log likelihood. The overall cross validation error is the sum of the errors across the k subsets divided by the total number of observations in the original dataset. Put it in a mathematical form, the k-fold cross validation error can be expressed as,

$$CV_k(\lambda) = -\sum_{i=1}^{k} \ell_n^i(\hat{\beta}_{n(-i)}(\lambda))/n \tag{2.8}$$

$\hat{\beta}_{n(-i)}(\lambda)$ is the penalized estimate for β at λ with the i^{th} subset taken out as the test set and the remaining $k-1$ subsets kept as the training set. $\ell_n^i(.)$ is the log partial likelihood for the i^{th} subset.

With each of the three model selection methods, the goal is to choose an optimal λ that minimizes the corresponding criterion. It is well known that AIC tends to choose a relatively more complicated model than BIC does and a more complicated model is generally associated with a smaller λ value. Cross validation generally performs better because it is a data driven method, however it is extremely computation intensive.

It is worthwhile to note that the popular leave-one-out cross validation methods is asymptotically equivalent to the AIC model selection criterion. Shao (1993) pointed out that both AIC and k-fold cross-validation are asymptotically inconsistent in the sense that the probability of selection the model with the best predictive ability does not converge to 1 as the total number of observations $n \to \infty$. He also proved that a leave-n_v-out cross validation with is asymptotically

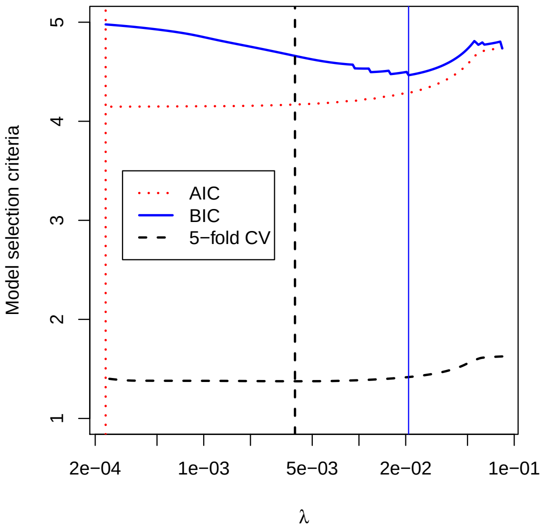

Figure 2.4: Plot of the three model selection criteria with the PBC data.

consistent if $n_v/n \to 1$ and $n_v \to \infty$ as $n \to \infty$.

Figure 2.4 plots the three model selection criteria with the PBC data. It illustrates the different optimal λ values chosen by the three model selection criteria.

2.7 Asymptotic Results

In this section, we list the asymptotic properties of the group lasso estimator for the Cox PH model. The results are derived with the number of covariates being fixed while the sample size increases. The B-spline approximation of a covariate's nonparametric effect is not considered here.

2.7.1 Set up

The objective function in the group lasso penalized Cox PH model is

$$Q_n(\beta, \lambda_n) = -\frac{1}{n}\ell_n(\beta) + \lambda_n \sum_{j=1}^{K} \sqrt{p_j}\left\|\beta_{(j)}\right\|$$

where $Q_n(\beta, \lambda_n)$ is the objective function to be minimized over the model parameter β with a given λ_n and n is the sample size. $\ell_n(\beta)$ is the partial likelihood from a Cox PH model.

λ_n refers to an optimal tuning parameter which changes with the sample size. In the computational implementation, a tuning parameter is usually prespecified and an optimal tunning parameter is generally chosen according to some criterion such as AIC, BIC, cross validation, or generalized cross validation. However, in the proofs of the asymptotic results, λ_n is assumed to converge at a rate in addition to the regularity conditions for Cox PH model.

The model parameter β is decomposed into K vectors $\beta_{(j)}, j = 1, 2, \cdots, K$, which correspond to the K covariate groups, respectively. A covariate group could include only one covariate or multiple covariates. The number of covariates in group j is denoted by $p_j, j = 1, 2, \cdots, K$, and $\|\cdot\|$ is the L_2 norm.

The true model parameter is denoted as β^0. Without loss of generality, we assume the coefficients of the first S, $S \leq K$, covariate groups are nonzero and the remaining are zero. That is, $\beta^0_{(j)} \neq 0$ for $j = 1, 2, \cdots, S$, and $\beta^0_{(j)} = 0$ for $j = S+1, S+2, \cdots, K$. The set of nonzero covariate groups are denoted as A and we use the subscript $_{(A)}$ to represent it.

The group lasso estimator for a Cox PH model with an optimal tuning parameter λ_n is defined as

$$\hat{\beta}_n(\lambda_n) = \arg\min_{\beta} Q_n(\beta, \lambda_n)$$

2.7.2 Preparation

The proofs of the asymptotic results stated in this section builds up on two asymptotic results given by Andersen and Gill (1982). According to Theorem 3.2 in Andersen and Gill (1982), the following two results hold given the regularity conditions (A) - (D).

$$\begin{aligned} n^{-1/2}\dot{\ell}_n(\beta^0) &\xrightarrow{p} N(0, \Sigma) \\ -\frac{1}{n}\ddot{\ell}_n(\beta^*) &\xrightarrow{p} \Sigma \qquad\qquad \text{for any random } \beta^* \xrightarrow{p} \beta^0 \end{aligned}$$

where β^0 is the true model parameter and $\dot{\ell}_n(\beta^0)$ and $\ddot{\ell}_n(\beta^*)$ are the first and the second order derivatives of $\ell_n(\beta)$, i.e., the score function and the Hessian matrix, evaluated at β^0 and β^* respectively. Σ is the positive definite Fisher information matrix and $\Sigma \triangleq E\{-\ddot{\ell}(\beta^0)\}$. $\Sigma = E\dot{\ell}(\beta^0)^{\otimes 2}$ holds also because the model is assumed to be true.

Theorem 2.1. *(Consistency) Under regularity conditions (A) - (D) in Andersen and Gill (1982), if $\lambda_n \to 0$ as $n \to 0$, then there exists a minimizer $\hat{\beta}_n(\lambda_n)$ of $Q_n(\beta, \lambda_n)$ such that $\hat{\beta}_n(\lambda_n) \xrightarrow{p} \beta^0$ as $n \to \infty$.*

Proof of Theorem 2.1

Proof. Theorem 5.7 in Van der Vaart(1998) can be applied to prove the theorem. However, a slightly different approach is adopted in this proof.

As in Van der Vaart and Wellner (1996), We only need to show that $\forall a > 0$,

$$Pr\{\sup_{\beta:\|\beta-\beta^0\|=a} Q_n(\beta,\lambda_n) > Q_n(\beta^0,\lambda_n)\} \to 1 \tag{2.9}$$

This implies that with probability goes to 1, $Q_n(\beta,\lambda_n)$ has a local minima in the ball $\{\beta : \|\beta-\beta^0\| < a\}$ for a given λ_n. As discussed in the previous section, the local minima is also a global minima since the objective function is concave.

Thus, we have

$$\left[\sup_{\beta:\|\beta-\beta^0\|=a} Q_n(\beta,\lambda_n) > Q_n(\beta^0,\lambda_n)\right] \subseteq \left[\left\|\hat{\beta}_n(\lambda_n)-\beta^0\right\| < a\right]$$

where $[\cdot]$ represents the event set which satisfies the condition inside the bracket. Therefore,

$$Pr\left\{\left\|\hat{\beta}_n(\lambda_n)-\beta^0\right\| < a\right\} \geq Pr\left\{\sup_{\beta:\|\beta-\beta^0\|=a} Q_n(\beta,\lambda_n) > Q_n(\beta^0,\lambda_n)\}\right\}$$

which implies,

$$Pr\left\{\left\|\hat{\beta}_n(\lambda_n)-\beta^0\right\|\right\} \to 1$$

Now, we will show that equation (2.9) holds.

$$\begin{aligned}
& Q_n(\beta,\lambda_n) - Q_n(\beta^0,\lambda_n) \\
= \quad & -\frac{1}{n}\left(\ell_n(\beta)-\ell_n(\beta^0)\right) + \sum_{j=1}^{K}\lambda_n\sqrt{p_j}\left(\|\beta_{(j)}\| - \|\beta^0_{(j)}\|\right) \\
\geq \quad & -n^{-1/2}\left(n^{-1/2}\frac{\partial\ell_n(\beta^0)}{\partial\beta}\right)'(\beta-\beta^0) + (\beta-\beta^0)'\left(n^{-1}\frac{\partial^2\ell_n(\beta^0)}{\partial\beta^2}\right)(\beta-\beta^0) \\
& +n^{-1}o_p(\left\|\beta-\beta^0\right\|^2) - \lambda_n\sum_{j=1}^{K}\sqrt{p_j}\left\|\beta-\beta^0\right\| \\
\geq \quad & -n^{-1/2}O_p(1)\left\|\beta-\beta^0\right\| + (\beta-\beta^0)'\left(\Sigma+o_p(1)\right)(\beta-\beta^0) \\
& +n^{-1}o_p\left(\left\|\beta-\beta^0\right\|^2\right) - \lambda_n\sum_{j=1}^{K}\sqrt{p_j}\left\|\beta-\beta^0\right\|
\end{aligned}$$

Since $\lambda_n \to 0$ as $n \to 0$, the above quantity is dominated by the second term $(\beta - \beta^0)' (\Sigma + o_p(1)) (\beta - \beta^0)$. It is positive since Σ is positive definite. Hence, equation (2.9) holds. This completes the proof.

Theorem 2.2. *(Convergence rate) Under regularity conditions (A) - (D) in Andersen and Gill (1982), if $\lambda_n = O(n^{-1/2})$, then $\left\| \hat{\beta}_n(\lambda_n) - \beta^0 \right\| = O_p\left(n^{-1/2}\right)$*

Proof of Theorem 2.2

Proof. We need to show $\sqrt{n} \left\| \hat{\beta}_n(\lambda_n) - \beta^0 \right\| = O_p(1)$, that is, $\forall \epsilon > 0$, $\exists M > 0$, such that

$$P^* \left\{ \sqrt{n} \left\| \hat{\beta}(\lambda_n) - \beta^0 \right\| > 2^M \right\} < \epsilon \text{ as } n \to \infty$$

where P^* is an outer probability measure.

As in the proof of Theorem 3.2.5 in Van der Vaart and Wellner (1996), for a given integer $M > 0$, partition the the parameter space $\left\{\beta; \sqrt{n} \left\|\beta - \beta^0\right\| \geq 2^M\right\}$ into the "shells" $S_{j,n} = \{\beta; 2^{j-1} < \sqrt{n} \left\|\beta - \beta^0\right\| \leq 2^j\}$ with $j = M+1, M+2, \cdots$.

Thus, for every $\eta > 0$, we have

$$\begin{aligned}
& P^* \left\{ \sqrt{n} \left\| \hat{\beta}_n(\lambda_n) - \beta^0 \right\| > 2^M \right\} \\
= & \sum_{\substack{j>M \\ 2^j \leq \eta\sqrt{n}}} P^* \left\{ 2^{j-1} < \sqrt{n} \left\| \hat{\beta}_n(\lambda_n) - \beta^0 \right\| \leq 2^j \right\} + \\
& P^* \left\{ \sqrt{n} \left\| \hat{\beta}_n(\lambda_n) - \beta^0 \right\| > \eta\sqrt{n} \right\} \\
= & \sum_{\substack{j>M \\ 2^j \leq \eta\sqrt{n}}} P^* \left\{ \hat{\beta}_n(\lambda_n) \in S_{j,n} \right\} + P^* \left\{ \left\| \hat{\beta}_n(\lambda_n) - \beta^0 \right\| > \eta \right\}
\end{aligned}$$

From Theorem 2.1, the second term $\to 0$ as $\eta \to \infty$. As to the first term, since $\hat{\beta}_n(\lambda_n) \in S_{j,n}$ implies

$$\inf_{\beta \in S_{j,n}} Q_n(\beta, \lambda_n) \leq Q_n(\beta^0, \lambda_n)$$

thus,

$$P^* \left\{ \hat{\beta}_n(\lambda_n) \in S_{j,n} \right\} \leq P^* \left\{ \inf_{\beta \in S_{j,n}} \left\{ Q_n(\beta, \lambda_n) - Q_n(\beta^0, \lambda_n) \right\} \leq 0 \right\}$$

Let $h_n \triangleq \sqrt{n}(\beta - \beta^0)$ and r_1 be the smallest eigenvalue of Σ. From the proof of Theorem 2.1, we have

$$\begin{aligned} & Q_n(\beta, \lambda_n) - Q_n(\beta^0, \lambda_n) \\ \geq\ & -n^{-1} \left(n^{-1/2} \frac{\partial \ell_n(\beta^0)}{\partial \beta} \right)' h_n + n^{-1} h_n' \left(n^{-1} \frac{\partial^2 \ell_n(\beta^0)}{\partial \beta^2} \right) h_n - \\ & n^{-1/2} \lambda_n \sum_{j=1}^{K} \sqrt{p_j} \left\| h_n \right\| + n^{-2} o_p(\left\| h_n \right\|^2) \\ \geq\ & -n^{-1} \left(n^{-1/2} \frac{\partial \ell_n(\beta^0)}{\partial \beta} \right)' h_n + n^{-1} r_1 \left\| h_n \right\|^2 - n^{-1/2} \lambda_n \sum_{j=1}^{K} \sqrt{p_j} \left\| h_n \right\| + \\ & n^{-2} o_p(\left\| h_n \right\|^2) \end{aligned}$$

By assumption, $\lambda_n = O(n^{-1/2})$, which means

$$\lambda_n \sqrt{n} \sum_{j=1}^{K} \sqrt{p_j} \left\| h_n \right\| \leq 0.5 r_1 \left\| h_n \right\|^2$$

for any $\left\| h_n \right\| > 2^M$ if we choose n and 2^M large enough. Thus,

$$Q_n(\beta, \lambda_n) - Q_n(\beta^0, \lambda_n) \geq -\frac{1}{n} \left(n^{-1/2} \frac{\partial \ell_n(\beta^0)}{\partial \beta} \right)' h_n + 0.5 n^{-1} r_1 \left\| h_n \right\|^2 + n^{-2} o_p(\left\| h_n \right\|^2)$$

Let r_2 be the largest eigenvalue of Σ and denote $\frac{8\sqrt{(r_2)}}{r_1}$ as r, then we have

$$\begin{aligned}
& P^*\left\{\hat{\beta}_n(\lambda_n) \in S_{j,n}\right\} \\
\leq\ & o(1) + P^*\left\{\inf_{2^{j-1}<\|h_n\|\leq 2^j} -\left(n^{-1/2}\frac{\partial \ell_n(\beta^0)}{\partial\beta}\right)' h_n + 0.5r_1\|h_n\|^2 \leq 0\right\} \\
\leq\ & o(1) + P^*\left\{\sup_{2^{j-1}<\|h_n\|\leq 2^j} \left(n^{-1/2}\frac{\partial \ell_n(\beta^0)}{\partial\beta}\right)' h_n \geq r_1 2^{2j-3}\right\} \\
\leq\ & o(1) + \frac{E\left|\sup_{2^{j-1}<\|h_n\|\leq 2^j}\left(n^{-1/2}\frac{\partial \ell_n(\beta^0)}{\partial\beta}\right)' h_n\right|}{r_1 2^{2j-3}} \\
\leq\ & o(1) + \frac{E^{1/2}\left\{\sup_{2^{j-1}<\|h_n\|\leq 2^j} h_n'\left(n^{-1/2}\frac{\partial \ell_n(\beta^0)}{\partial\beta}\right)^{\otimes 2} h_n\right\}}{r_1 2^{2j-3}} \\
\leq\ & o(1) + \frac{E^{1/2}\left\{\sup_{2^{j-1}<\|h_n\|\leq 2^j} h_n'(\Sigma + O_p(1))h_n\right\}}{r_1 2^{2j-3}} \\
\leq\ & o(1) + \frac{2^j * \sqrt{r_2}}{r_1 2^{2j-3}} \\
\leq\ & o(1) + r2^{-j}
\end{aligned}$$

That is,

$$P^*\left\{\sqrt{n}\left\|\hat{\beta}_n(\lambda_n) - \beta^0\right\| > 2^M\right\} \leq o(1) + \sum_{j=M+1}^{\infty} r2^{-j} \leq o(1) + r2^{-M}$$

This indicates $\sqrt{n}\left\|\hat{\beta}_n(\lambda_n) - \beta^0\right\| = O_p(1)$.

2.7.3 Lack of oracle property

Sparsity means that the estimate of zero coefficients goes to zero with probability tending to 1. The oracle property includes not only sparsity, but also that the estimate of nonzero coefficients reaches the Cramer-Rao lower bound asymptotically as if the set of zero coefficients were known in advance. In the Group Lasso penalized Cox PH model, the estimator is root-n consistent, however it does not have sparsity or oracle property. From the proofs of Theorem 3.3 in Chapter 3, we conclude that a necessary condition for an estimator to be sparse is that the first order derivative of the penalty function with respect to those zero coefficients needs to go to infinity with probability tending to 1.

2.8 Computational implementation of the algorithm

2.8.1 Simplification of the calculations of the log partial likelihood function, score function and negative Hessian matrix

The calculation of the three expressions could be complicated and time-consuming because of their dependence on risk sets and tied-failure groups. Furthermore, we need to evaluate them numerous time while calculating the solution path for the group lasso penalized Cox model. A pre-processing is implemented to better structure the calculation.

First, all the observations are sorted first by survival time in ascending order and then by status indicator with events first and censoring second. After sorting data in this way, each group of tied failures is aggregated together and the risk set for each failure starts from itself till the last observation when there are no tied failures.

Second, for each group of tied failure events, the start index and number of ties are searched and saved to speed up the processing of tie correction.

Third, define the following three qualities $s^*_{0,j}$, $s^*_{1,j}$, and $s^*_{2,j}$.

$$\begin{aligned} s^*_{0,j} &= \frac{\sum_{l\in D(t_i)} \exp(X_l'\beta)}{d_i} \\ s^*_{1,j} &= \frac{\sum_{l\in D(t_i)} \exp(X_l'\beta)X_l}{d_i} \\ s^*_{2,j} &= \frac{\sum_{l\in D(t_i)} \exp(X_l'\beta)X_l X_l'}{d_i} \\ & \forall i \in \{1, 2, \cdots, k\} \quad \forall j \in D(t_i) \end{aligned}$$

The calculation of the three expressions now becomes

$$\begin{aligned} \ell_n(\beta) &= \sum_{i=1}^n \Delta_i \Big\{ X_i'\beta - \log\Big\{ \sum_{j=i}^n s^*_{0,j} \Big\} \Big\} \\ \dot{\ell}_n(\beta) &= \sum_{i=1}^n \Delta_i \Big\{ X_i' - \frac{\sum_{j=i}^n s^*_{1,j}}{\sum_{j=i}^n s^*_{0,j}} \Big\} \\ \ddot{\ell}_n(\beta) &= -\sum_{i=1}^n \Delta_i \Big\{ \frac{\sum_{j=i}^n s^*_{2,j}}{\sum_{j=i}^n s^*_{0,j}} - \frac{\big\{\sum_{j=i}^n s^*_{1,j}\big\}\big\{\sum_{j=i}^n s^*_{1,j}\big\}'}{\big\{\sum_{j=i}^n s^*_{0,j}\big\}^2} \Big\} \end{aligned}$$

With this processing, the only extra work needed for tie correction is just as that described in the third step. Since each group of tied failure events is aggregated together after the sorting in the first step and we've saved the start index and number of each tied failure group in the second step, the correction of ties in the third step is greatly simplified. After the correction of ties, the remaining calculation is the same as the situation without ties. Furthermore, there is no need to search for risk sets in order to calculate the three expressions after the sorting.

2.8.2 R Interface

This R implemention of the grouped variable selection in a high-dimensional partially linear additive Cox model, *gs.coxph*, inherits the interface for R function *coxph*. This interface provides a convenient way to specify the Cox model and the model covariates. With this interface, a categorical variable can be incorporated into the model specification using the *factor* keyword. Similarly, if we need to use B-spline to approximate the nonparamatric effect of a covariate, we can use R function *bs* to obtain the basis matrix.

Four new arguments are added to the interface to make it suitable for the specification of a group lasso penalized Cox model. Two additional arguments for the adaptive group lasso and for the generalized group lasso are also described here to make the description complete.

- *n.freevars* specifies the how many covariate at the beginning of the covariate set are free variables. The default value is zero. More information on free variables is available in Chapter 4. It is zero by default.

- *group.selection*. A vector is used to indicate which variables belong to a group. For example, if there are 16 variables included in a Cox model and 2 free variables and the group specification is *c(3,1,8)*, then covariates 1-2 are free variables, variables 3-5 form the first group, covariate 6 the second group, covariates 7-14 the third group, and the remaining two covariates are one covariate per group. The default value NULL indicates there is no grouping of covariates.

- *lambda* argument. The default value is NULL. It could be specified as one value or a vector of values. If *lambda=NULL*, then 200 λ values are chosen according to the rules specified in (2.7). Otherwise, it will calculate a solution corresponding to the specified λ value or a solution path is it is a vector of values.

- *lambdaRate*. It is equivalent to c in (2.7). The default value is 1.03. It is ignored if *lambda* is not NULL.
- *LassoWGT* specifies the adaptive weight which is described in Chapter 3.
- *debugtype*. The default value is zero for calculating a solution path. If *debugtype = 2*, it performs a 5-fold cross validation and *lambda* needs to be specified too.

```
fit1 = gs.coxph(Surv(time, status) ~ factor(stage) + factor(edtrt)
       + bs(age, df=6) + bs(bili, df = 6) + sex + trt + edema,
       data = pbc1, lambda = NULL, debugtype = 0,
       group.selection = c(3,2,6, 6),
       control = list(iter.max = 50))
```

The above example illustrates how to call the R function *gs.coxph* to calculate the group lasso penalized solution path with the PBC data.

- In *gs.coxph*, the specification of a Cox model is the same as that in *coxph*.
- Data *pbc1* is a subset of *pbc* dataset that comes with the *survival* package. More information about the PBC data is available in Chapter 1. The subset excludes observations 313-418 and a few covariates because of missing values.

- The whole solution path will be calculated since *lambda* is NULL.

- *group.selection* is used to specify covariate groups. The four levels of *stage* are coded by three dummy covariates, which constitute the first group. Similarly, the three levels of *edtrt* are coded by two dummy covariates in the second group. B-spline with 6 degrees of freedom is used to approximate the nonparametric effect of *age*. The 6 columns of the B-spline basis matrix for age form the third group. Similarly, the 6 columns of the B-spline basis matrix for *bili* form the fourth group. The last three covariates, *sex*, *trt*, and *edema*, are one variable per group, which is not explicitly specified in *group.selection*.

To find an optimal λ value using 5-fold cross validation, it takes another functional call to *gs.coxph*. The specification of the function call is as the following.

```
cv1 = gs.coxph(Surv(time, status) ~ factor(stage) + factor(edtrt)
       + bs(age, df=6) + bs(bili, df = 6) + sex + trt + edema,
       data = pbc1, lambda = fit1$Lambdas, debugtype = 2,
       group.selection = c(3,2,6, 6),
       control = list(iter.max = 50))
```

Most of the specifications are similar to those in the first function call to *gs.coxph* except the specifications of *lambda* and *debugtype*. Here *lamba* is set to be *fit1$Lambdas*, where *fit1* is an R dataset returned by the first function call and *Lambdas* is a component of *fit1* and contains a set of λ grid points where the solution path are located on. *debugtype* is set to 2 to require a 5-fold cross validation. It returns the cross validation errors along the solution path.

2.8.3 Optimization of R implementation

Rewriting R functions in C dramatically speeds up the calculation as it can reduce computing time and optimize memory usage.

- R offers a user friendly interface at the cost of computation efficiency. It is well known that R is extremely slow when there are more than three levels of nested **for** loops. The calculation of the three Cox expressions, the partial log-likelihood, the score function, and the Hessian matrix takes about 99% of the computation time when it is implemented in R. Rewriting the calculation of the three Cox expressions in C speeds up the computation by about 400 times.

- Memory usage is another bottleneck for modeling high-dimensional survival data. R function interface passes data by value, which means it will make a temporary copy for each input data at the beginning of a function call. However, C functions that can be called inside R via R function **.Call** to pass data by reference, which means that all the input datasets are directly accessed via pointers.

- Since the negative Hessian matrix is symmetric, only the lower triangular part of the matrix is saved during calculation. This reduces the memory usage almost by half.

- A general practice in R and C mixed programming, which is also followed in our implementation, is to implement the input part and the output part in R and put the computational intensive part in C. Thus, we can make use of the various packages available to create high quality graphs and summarize the results with little effort. By putting the computation intensive part in C, the computing time and memory requirement can be dramatically reduced. Furthermore, a lot of R functions have their corresponding API version available to be called inside C.

2.9 Conclusion

In this chapter, the properties and computational issues of the group lasso method with high-dimensional partially linear additive Cox model are studied. A method to find the solution path is proposed and implemented in R. AIC, BIC, and cross validation are adapted to select an optimal λ value. The group lasso estimate is demonstrated to be root-n consistent under some regularity conditions, However, it does not have oracle property.

Although group lasso penalized model gives out some zero estimates, it tends to select a large model which includes zero effects. That is, group lasso is not model selection consistent. An improved method is introduced next.

CHAPTER 3
ADAPTIVE GROUP LASSO PENALIZED COX REGRESSION

In this Chapter, we study the adaptive group lasso variables selection in high-dimensional partially linear additive Cox PH model. This method is closely related to the group lasso method. The computational implementation of the method builds up on the implementation of group lasso. The strength of this method is that it can produce a sparse solution and possesses oracle properties. This strength is what the group lasso method lacks of.

3.1 Adaptive group lasso

In the group lasso method, the penalty is in the following form,

$$\lambda \sum_{j=1}^{K} \sqrt{p_j} \left\| \beta_{(j)} \right\|$$

The penalty is continuous in β. A disadvantage of continuous penalty is that it does not discriminate zero coefficient and nonzero coefficient efficiently. This is the reason that group lasso does not possess oracle property.

The ideal adaptive group lasso penalty is

$$\lambda \sum_{j=1}^{K} \sqrt{p_j} I\{\beta_{(j)} \neq 0\} \tag{3.1}$$

The penalty is discontinuous and jumps at zero. As proved in the theoretical results section in this chapter, the adaptive group lasso penalty encourages estimates of coefficients to be zero and enjoys oracle property.

3.1.1 Approximation of the adaptive group lasso penalty

The adaptive group lasso penalty is discontinuous, which increases the complexity of computational implementation significantly. An approximation of the penalty could result in a penalty that is continuous and yet preserve the advantages of adaptive group lasso. The approximation is to incorporate a known consistent estimate of β, denoted as $\tilde{\beta}$, into the penalty as follows,

$$\sum_{j=1}^{K} \frac{\lambda\sqrt{p_j}}{\left\|\tilde{\beta}_{(j)}\right\|} \left\|\beta_{(j)}\right\| \tag{3.2}$$

Compared to the group lasso penalty, the above approximated adaptive group lasso penalty penalizes small coefficients more and the large coefficients less. This approximated method lies in somewhere between the group lasso method and the ideal adaptive group lasso method. It is asymptotically equivalent to the ideal adaptive group lasso method. All the thesis work on adaptive group lasso uses the approximated penalty in expression (3.2). From now on, we refer to expression (3.2) whenever we talk about the adaptive group lasso method.

It is continuous which makes the implementation of the method relatively easier and yet preserves the properties of adaptive group lasso when n is large. A slight modification of the group lasso implementation will be able to solve the adaptive group lasso penalized method, if there is a consistent estimate of β known in advance.

From the asymptotic properties of the group lasso, we know that the group lasso estimate is root-n consistent, thus the group lasso estimate can be used as $\tilde{\beta}$ in the approximation of the adaptive group lasso penalty in expression (3.2).

Three choices of $\tilde{\beta}$ are considered. Two of them are group lasso estimates and one is the parameter estimate from a regular Cox PH model. The first one is the optimal group lasso estimate chosen by cross validation, which is a constant. The optimal group lasso estimate chosen by other variable selection criteria such as AIC or BIC will work as well. The second one is the group lasso estimate with its tuning parameter equal to that in expression (3.2). This weight varies as the tuning parameter λ in expression (3.2) changes. When the number of covariates is less than the sample size and a regular Cox PH model can be fitted on the data, the regular Cox estimate can be used to substitute $\tilde{\beta}$ also. This is the third choice and is also a constant with respect to λ in (3.2).

It is worthy to note that the second choice of $\tilde{\beta}$ is somewhat unsatisfactory as it is not always a consistent estimate of β as λ changes. If a group coefficient, $\tilde{\beta}_{(j)}$, is zero, then $1/\left\|\tilde{\beta}_{(j)}\right\| = \infty$ and the covariate group j would be excluded from the adaptive group lasso penalized model. Figure 3.1 illustrates three different solution paths.

3.1.2 Objective function and KKT conditions

Similar to group lasso method, in adaptive group lasso the objective is to minimize the following function,

$$\Phi_n(\beta, \lambda) = -\frac{1}{n}\ell_n(\beta) + \sum_{j=1}^{K} \frac{\lambda\sqrt{p_j}}{\left\|\tilde{\beta}_{(j)}\right\|} \left\|\beta_{(j)}\right\| \tag{3.3}$$

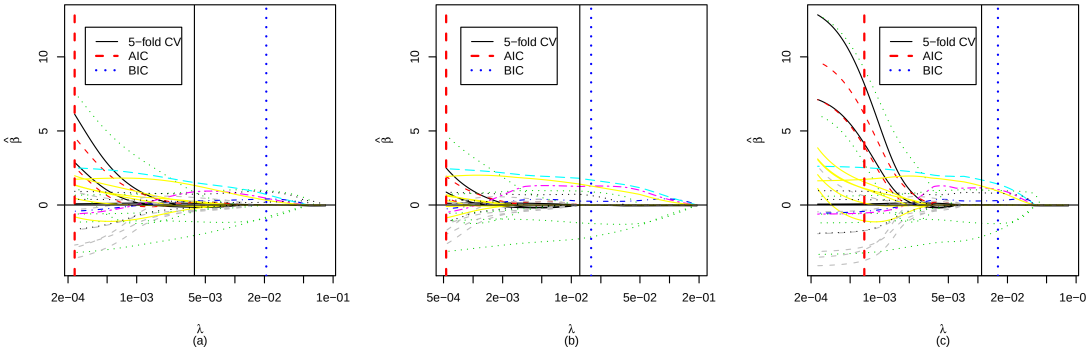

Figure 3.1: Solution paths. (a) - Group lasso, (b) - Adaptive group lasso with $\tilde{\beta}$ from the optimal model chosen by cross validation, (c) - Adaptive group lasso using the whole group lasso solution path as $\tilde{\beta}$.

where $\ell(\beta)$ is the partial likelihood of partially linear additive Cox PH model using B-spline to approximate the nonlinear effects as introduced in the previous chapter.

For a given λ, the adaptive group lasso estimate $\hat{\beta}_n(\lambda)$ is defined as

$$\hat{\beta}_n(\lambda) = \arg\min_{\beta} \Phi_n(\beta, \lambda)$$

By taking the first-order derivative on the objective function $\Phi_n(\beta, \lambda)$, we have derived the following KKT conditions for finding the minimizer of $\Phi_n(\beta, \lambda)$ over β.

$$\begin{cases} \beta_{(j)} = 0 & \forall \tilde{\beta}_{(j)} = 0 \\ -\frac{1}{n}\dot{\ell}_{n(j)}(\beta) + \frac{\lambda\sqrt{p_j}}{\|\tilde{\beta}_{(j)}\|}\frac{\beta_{(j)}}{\|\beta_{(j)}\|} = 0 & \forall \beta_{(j)} \neq 0 \text{ and } \tilde{\beta}_{(j)} \neq 0 \\ \|\frac{1}{n}\dot{\ell}_{n(j)}(\beta)\| \leq \frac{\lambda\sqrt{p_j}}{\|\tilde{\beta}_{(j)}\|} & \forall \beta_{(j)} = 0 \text{ and } \tilde{\beta}_{(j)} \neq 0 \end{cases} \tag{3.4}$$

The first condition is a special case of the third condition since we define $\frac{1}{0} = \infty$.

It is listed separately just to make it explicit. In addition to the first condition, the third condition forces some components of β to be zero. Again the KKT conditions are not only necessary but also sufficient. The computational implementation of the adaptive group lasso method is based on how to solve the above KKT conditions.

The minimization problem in (4.3) is equivalent to the following optimization problem,

$$\begin{aligned} & \min_{\beta} -\ell_n(\beta) \\ s.t. \quad & \sum_{j=1}^{J} \frac{\sqrt{p_j}}{\|\tilde{\beta}_{(j)}\|} \|\beta_{(j)}\| \leq t \end{aligned} \tag{3.5}$$

where t is the upper bound of the constraints on β and reflects the relative magnitude of β.

3.1.3 Selection of the tuning parameter λ

The maximum meaningful λ, λ_{max}, is defined as the minimum λ that gives out an adaptive group lasso estimate of all zeros. That is,

$$\begin{aligned} \lambda_{max} &= \inf\{\lambda; \quad \hat{\beta}_n(\lambda) = 0\} \\ &= \max_{j} \frac{\|\dot{\ell}_{n(j)}(0)\| \|\tilde{\beta}_{(j)}\|}{n\sqrt{p_j}} \end{aligned}$$

The minimum of λ, λ_{min}, can be zero if the sample size is larger than the number of covariates in the model and none of the covariates is a linear combination of other covariates. Otherwise, λ_{min} can be chosen to be a small positive number.

As λ decreases from λ_{max} to λ_{min}, the adaptive group lasso estimate $\hat{\beta}_n(\lambda)$ increases and forms a solution path.

Correspondingly, the meaningful range of t, (t_{min}, t_{max}), is defined as the following,

$$\begin{aligned} t_{min} &= \sum_{j=1}^{K} \frac{\sqrt{p_j}}{\left\| \tilde{\beta}_{(j)} \right\|} \hat{\beta}_{n(j)}(\lambda_{max}) \\ &= 0 \\ t_{max} &= \sum_{j=1}^{K} \frac{\sqrt{p_j}}{\left\| \tilde{\beta}_{(j)} \right\|} \hat{\beta}_{n(j)}(\lambda_{min}) \\ &= \sum_{j=1}^{K} \frac{\sqrt{p_j}}{\left\| \tilde{\beta}_{(j)} \right\|} \|\hat{\beta}_{n(j)}(0)\| \qquad \text{if } \lambda_{min} = 0 \end{aligned}$$

where $\dot{\ell}_{n(j)}(0)$ includes the items of the score function $\dot{\ell}_n(\beta)$ evaluated at $\beta = 0$ that correspond to the j^{th} variable group.

There is a one to one relationship between λ and t, that is, t increases strictly as λ decreases from λ_{max} to λ_{min}.

Similar to the group lasso method, modified AIC, BIC, and cross-validation are implemented to find the optimal tuning parameter λ.

3.2 Asymptotic Results

In this section, we list the asymptotic properties of the adaptive group lasso estimator for the Cox PH model. The results are derived with the number of covariates being fixed while the sample size increases. The B-spline approximation of a covariate's nonparametric effect is not considered here.

3.2.1 Set up

The objective function in the adaptive group lasso penalized Cox PH model with tuning parameter λ_n is

$$\Phi_n(\beta, \lambda_n) = -\frac{1}{n}\ell_n(\beta) + \lambda_n \sum_{j=1}^{K} \frac{\sqrt{p_j}}{\tilde{\beta}_{(j)}} \left\|\beta_{(j)}\right\|$$

where $\Phi_n(\beta, \lambda_n)$ is the objective function to be minimized over the model parameter β with a given λ_n and n is the sample size. $\ell_n(\beta)$ is the partial likelihood from a Cox PH model.

λ_n refers to an optimal tuning parameter which changes with the sample size. In the computational implementation, a tuning parameter is usually prespecified and an optimal tunning parameter is generally chosen according to some criterion such as AIC, BIC, cross validation, or generalized cross validation. However, in the proofs of the asymptotic results, λ_n is assumed to converge at a rate.

The model parameter β is decomposed into K vectors $\beta_{(j)}, j = 1, 2, \cdots, K$, which corresponds to the K covariate groups. A covariate group could include only one covariate or multiple covariates. The number of covariates in group j is denoted by $p_j, j = 1, 2, \cdots, K$, and $\|\cdot\|$ is the L_2 norm.

The true model parameter is denoted as β^0. Without loss of generality, we assume the coefficients of the first S, $S \leq K$, covariate groups are nonzero and the remaining are zero. That is, $\beta^0_{(j)} \neq 0$ for $j = 1, 2, \cdots, S$, and $\beta^0_{(j)} = 0$ for $j = S+1, S+2, \cdots, K$. The set of nonzero covariate groups are denoted as A and we use the subscript $_{(A)}$ to represent it.

The adaptive group lasso estimator for a Cox PH model with an optimal tuning parameter λ_n is defined as

$$\hat{\beta}_n(\lambda_n) = \arg\min_{\beta} \Phi_n(\beta, \lambda_n)$$

3.2.2 Preparation

The proofs of the asymptotic results stated in this section builds up on two asymptotic results given by Andersen and Gill (1982). According to Theorem 3.2 in Andersen and Gill (1982), the following two results hold given the regularity conditions (A) - (D).

$$\begin{aligned} n^{-1/2}\dot{\ell}_n(\beta^0) &\xrightarrow{p} N(0,\Sigma) \\ -\frac{1}{n}\ddot{\ell}_n(\beta^*) &\xrightarrow{p} \Sigma \qquad\qquad \text{for any random } \beta^* \xrightarrow{p} \beta^0 \end{aligned}$$

where β^0 is the true model parameter and $\dot{\ell}_n(\beta^0)$ and $\ddot{\ell}_n(\beta^*)$ are the first and the second order derivatives of $\ell_n(\beta)$, i.e., the score function and the Hessian matrix, evaluated at β^0 and β^* respectively. Σ is the positive definite Fisher information matrix and $\Sigma \triangleq E\{-\ddot{\ell}(\beta^0)\}$. $\Sigma = E\dot{\ell}(\beta^0)^{\otimes 2}$ holds also because the model is assumed to be true.

Theorem 3.1. *(Consistency) Under regularity conditions (A) - (D) in Andersen and Gill (1982), if $\lambda_n \to 0$ as $n \to 0$, then there exists a minimizer $\hat{\beta}_n(\lambda_n)$ of $\Phi_n(\beta,\lambda_n)$ such that $\hat{\beta}_n(\lambda_n) \xrightarrow{p} \beta^0$ as $n \to \infty$.*

Proof of Theorem 3.1

Proof. Theorem 5.7 in Van der Vaart(1998) can be applied to prove the theorem. However, a slightly different approach is adopted in this proof.

As in Van der Vaart and Wellner (1996), We only need to show that $\forall a > 0$,

$$Pr\{\sup_{\beta:\|\beta-\beta^0\|=a} \Phi_n(\beta,\lambda_n) > \Phi_n(\beta^0,\lambda_n)\} \to 1 \qquad (3.6)$$

This implies that with probability goes to 1, $\Phi_n(\beta,\lambda_n)$ has a local minima in the ball $\{\beta : \|\beta - \beta^0\| < a\}$ for a given λ_n. In addition, $E\{\Phi_n(\beta^0, \lambda_n)\} \rightarrow E\{-n^{-1}\ell_n(\beta)\}$ as $\lambda_n \rightarrow 0$ and $E\{-n^{-1}\ell_n(\beta)\}$ is a strictly convex function with a well separated global minimizer β^0. Thus, we have

$$\left[\sup_{\beta:\|\beta-\beta^0\|=a} \Phi_n(\beta, \lambda_n) > \Phi_n(\beta^0, \lambda_n)\right] \subseteq \left[\left\|\hat{\beta}_n(\lambda_n) - \beta^0\right\| < a\right]$$

where $[\cdot]$ represents the event set which satisfies the condition inside the bracket. Therefore,

$$Pr\left\{\left\|\hat{\beta}_n(\lambda_n) - \beta^0\right\| < a\right\} \geq Pr\left\{\sup_{\beta:\|\beta-\beta^0\|=a} \Phi_n(\beta, \lambda_n) > \Phi_n(\beta^0, \lambda_n)\}\right\}$$

which implies,

$$Pr\left\{\left\|\hat{\beta}_n(\lambda_n) - \beta^0\right\|\right\} \rightarrow 1$$

Now, we will show that equation (3.6) holds. Since $\tilde{\beta}$ is a consistent estimate of β^0, thus we have,

$$\left\|\tilde{\beta}_{(j)}\right\| I\{\beta^0_{(j)} \neq 0\} > \frac{1}{2}\left\|\beta^0_{(j)}\right\| \qquad \text{with probability tending to 1}$$

and there exists a $c > 0$ such that,

$$\min_{j=1}^{S}\left\|\tilde{\beta}_{(j)}\right\| > \frac{1}{2}\min_{j=1}^{S}\left\|\beta^0_{(j)}\right\| > c$$

$$\begin{aligned}
& \Phi_n(\beta, \lambda_n) - \Phi_n(\beta^0, \lambda_n) \\
= \; & -\frac{1}{n}\left(\ell_n(\beta) - \ell_n(\beta^0)\right) + \sum_{j=1}^{K}\frac{\lambda_n\sqrt{p_j}}{\left\|\tilde{\beta}(j)\right\|}\left(\left\|\beta_{(j)}\right\| - \left\|\beta^0_{(j)}\right\|\right) \\
\geq \; & -n^{-1/2}\left(n^{-1/2}\frac{\partial\ell_n(\beta^0)}{\partial\beta}\right)'(\beta - \beta^0) + (\beta - \beta^0)'\left(n^{-1}\frac{\partial^2\ell_n(\beta^0)}{\partial\beta^2}\right)(\beta - \beta^0)
\end{aligned}$$

$$
\begin{aligned}
& +n^{-1}o_p(\left\|\beta-\beta^0\right\|^2)+\lambda_n\sum_{j=1}^{S}\frac{\sqrt{p_j}}{\left\|\tilde{\beta}(j)\right\|}\left(\|\beta_{(j)}\|-\|\beta^0_{(j)}\|\right)\\
\geq\ & -n^{-1/2}O_p(1)\left\|\beta-\beta^0\right\|+(\beta-\beta^0)'\left(\Sigma+o_p(1)\right)(\beta-\beta^0)\\
& +n^{-1}o_p\left(\left\|\beta-\beta^0\right\|^2\right)-\lambda_n\sum_{j=1}^{S}\frac{\sqrt{p_j}}{c}\left\|\beta-\beta^0\right\|
\end{aligned}
$$

Since $\lambda_n \to 0$ as $n \to 0$, the above quantity is dominated by the second term $(\beta-\beta^0)'\left(\Sigma+o_p(1)\right)(\beta-\beta^0)$. It is positive since Σ is positive definite. Hence, equation (3.6) holds. This completes the proof.

Theorem 3.2. *(Convergence rate) Under regularity conditions (A) - (D) in Andersen and Gill (1982), if $\lambda_n = O(n^{-1/2})$, then $\left\|\hat{\beta}(\lambda_n)-\beta^0\right\| = O_p\left(n^{-1/2}\right)$*

Proof of Theorem 3.2

Proof. We need to show $\sqrt{n}\left\|\hat{\beta}_n(\lambda_n)-\beta^0\right\| = O_p(1)$, that is, $\forall\epsilon > 0$, $\exists M > 0$, such that

$$P^*\left\{\sqrt{n}\left\|\hat{\beta}_n(\lambda_n)-\beta^0\right\| > 2^M\right\} < \epsilon \text{ as } n \to \infty$$

where P^* is an outer probability measure.

As in the proof of Theorem 3.2.5 in Van der Vaart and Wellner (1996), for a given integer $M > 0$, partition the the parameter space $\left\{\beta; \sqrt{n}\left\|\beta-\beta^0\right\| \geq 2^M\right\}$ into the "shells" $S_{j,n} = \{\beta; 2^{j-1} < \sqrt{n}\|\beta-\beta^0\| \leq 2^j\}$ with $j = M+1, M+2, \cdots$.

Thus, for every $\eta > 0$, we have

$$\begin{aligned}
& P^*\left\{\sqrt{n}\left\|\hat{\beta}_n(\lambda_n) - \beta^0\right\| > 2^M\right\} \\
= & \sum_{\substack{j>M \\ 2^j \le \eta\sqrt{n}}} P^*\left\{2^{j-1} < \sqrt{n}\left\|\hat{\beta}_n(\lambda_n) - \beta^0\right\| \le 2^j\right\} + \\
& P^*\left\{\sqrt{n}\left\|\hat{\beta}_n(\lambda_n) - \beta^0\right\| > \eta\sqrt{n}\right\} \\
= & \sum_{\substack{j>M \\ 2^j \le \eta\sqrt{n}}} P^*\left\{\hat{\beta}_n(\lambda_n) \in S_{j,n}\right\} + P^*\left\{\left\|\hat{\beta}_n(\lambda_n) - \beta^0\right\| > \eta\right\}
\end{aligned}$$

From Theorem 3.1, the second term → 0 as $\eta \to \infty$. As to the first term, since $\hat{\beta}_n(\lambda_n) \in S_{j,n}$ implies

$$\inf_{\beta \in S_{j,n}} \Phi_n(\beta, \lambda_n) \le \Phi_n(\beta^0, \lambda_n)$$

thus,

$$P^*\left\{\hat{\beta}_n(\lambda_n) \in S_{j,n}\right\} \le P^*\left\{\inf_{\beta \in S_{j,n}} \left\{\Phi_n(\beta, \lambda_n) - \Phi_n(\beta^0, \lambda_n)\right\} \le 0\right\}$$

Let $h_n \triangleq \sqrt{n}(\beta - \beta^0)$ and r_1 be the smallest eigenvalue of Σ. From the proof of Theorem 3.1, we have

$$\begin{aligned}
& \Phi_n(\beta, \lambda_n) - \Phi_n(\beta^0, \lambda_n) \\
\ge & -n^{-1}\left(n^{-1/2}\frac{\partial \ell_n(\beta^0)}{\partial \beta}\right)' h_n + n^{-1}h_n'\left(n^{-1}\frac{\partial^2 \ell_n(\beta^0)}{\partial \beta^2}\right) h_n \\
& -n^{-1/2}\lambda_n \sum_{j=1}^{S} \frac{\sqrt{p_j}}{c}\|h_n\| + n^{-2}o_p(\|h_n\|^2) \\
\ge & -n^{-1}\left(n^{-1/2}\frac{\partial \ell_n(\beta^0)}{\partial \beta}\right)' h_n + n^{-1}r_1\|h_n\|^2 - n^{-1/2}\lambda_n \sum_{j=1}^{S} \frac{\sqrt{p_j}}{c}\|h_n\| \\
& +n^{-2}o_p(\|h_n\|^2)
\end{aligned}$$

By assumption, $\lambda_n = O(n^{-1/2})$, which means

$$\lambda_n \sqrt{n} \sum_{j=1}^{S} \frac{\sqrt{p_j}}{c} \left\| h_n \right\| \leq 0.5 r_1 \left\| h_n \right\|^2$$

for any $\|h_n\| > 2^M$ if we choose n and 2^M large enough. Thus,

$$\Phi_n(\beta, \lambda_n) - \Phi_n(\beta^0, \lambda_n) \geq -n^{-1} \left(n^{-1/2} \frac{\partial \ell_n(\beta^0)}{\partial \beta} \right)' h_n + 0.5 n^{-1} r_1 \left\| h_n \right\|^2 + n^{-2} o_p(\left\| h_n \right\|^2)$$

Let r_2 be the largest eigenvalue of Σ and denote $8\sqrt{(r_2)}/r_1$ as r, then we have

$$\begin{aligned}
& P^* \left\{ \hat{\beta}_n(\lambda_n) \in S_{j,n} \right\} \\
\leq \;& o(1) + P^* \left\{ \inf_{2^{j-1} < \|h_n\| \leq 2^j} - \left(n^{-1/2} \frac{\partial \ell_n(\beta^0)}{\partial \beta} \right)' h_n + 0.5 r_1 \left\| h_n \right\|^2 \leq 0 \right\} \\
\leq \;& o(1) + P^* \left\{ \sup_{2^{j-1} < \|h_n\| \leq 2^j} \left(n^{-1/2} \frac{\partial \ell_n(\beta^0)}{\partial \beta} \right)' h_n \geq r_1 2^{2j-3} \right\} \\
\leq \;& o(1) + \frac{E \left| \sup_{2^{j-1} < \|h_n\| \leq 2^j} \left(n^{-1/2} \frac{\partial \ell_n(\beta^0)}{\partial \beta} \right)' h_n \right|}{r_1 2^{2j-3}} \\
\leq \;& o(1) + \frac{E^{1/2} \left\{ \sup_{2^{j-1} < \|h_n\| \leq 2^j} h_n' \left(n^{-1/2} \frac{\partial \ell_n(\beta^0)}{\partial \beta} \right)^{\otimes 2} h_n \right\}}{r_1 2^{2j-3}} \\
\leq \;& o(1) + \frac{E^{1/2} \left\{ \sup_{2^{j-1} < \|h_n\| \leq 2^j} h_n' (\Sigma + o_p(1)) h_n \right\}}{r_1 2^{2j-3}} \\
\leq \;& o(1) + \frac{2^j * \sqrt{r_2}}{r_1 2^{2j-3}} \\
\leq \;& o(1) + r 2^{-j}
\end{aligned}$$

That is,

$$P^* \left\{ \sqrt{n} \left\| \hat{\beta}_n(\lambda_n) - \beta^0 \right\| > 2^M \right\} \leq o(1) + \sum_{j=M+1}^{\infty} r 2^{-j} \leq o(1) + r 2^{-M}$$

This indicates $\sqrt{n} \left\| \hat{\beta}_n(\lambda_n) - \beta^0 \right\| = O_p(1)$.

Theorem 3.3. *(Oracle property) Assume the regularity conditions (A) - (D) in Andersen and Gill(1982) and assume there exist $\eta > 0$ and $q \in (0.5, 1)$ such that $\lambda_n n^q \to \eta$ as $n \to \infty$. From Theorem 3.2, we know $\hat{\beta}_n(\lambda_n)$ is a root-n consistent minimizer of $\Phi_n(\beta, \lambda_n)$.*

(a) (Sparsity) For $\forall j \in \bar{A}$, i.e., $\beta_{(j)} = 0$, let $\hat{\beta}_{n(j)}(\lambda_n)$, a sub vector of $\hat{\beta}_n(\lambda_n)$, be the estimator of $\beta_{(j)}$, then $\hat{\beta}_{n(j)}(\lambda_n) = 0$ with probability tending to 1.

(b) (Asymptotic Normality) Let $\hat{\beta}_{n(A)}(\lambda_n)$, a sub vector of $\hat{\beta}_n(\lambda_n)$, be the estimator of all $\{\beta_{(j)}; j \in A\}$, then $\sqrt{n}(\hat{\beta}_{n(A)}(\lambda_n) - \beta_{(A)}) \xrightarrow{D} N(0, \Sigma^{-1}_{(A)})$.

Proof of Theorem 3.3

Proof. First, we prove the sparsity as specified in (a).

Since the KKT conditions derived for the adaptive group Lasso penalized Cox PH model are both necessary and sufficient condition to minimize the objective function, $\left[\hat{\beta}_{n(j)}(\lambda_n) = 0\right]$ equals to $\left[n^{-1} \left\|\dot{\ell}_{(j)}\left(\hat{\beta}_n(\lambda_n)\right)\right\| \leq \lambda_n \sqrt{p_j} / \left\|\tilde{\beta}_{(j)}\right\|\right]$. We only need to show

$$Pr\left\{n^{-1/2}\left\|\dot{\ell}_{n(j)}(\hat{\beta}_n(\lambda_n))\right\| \geq \lambda_n \sqrt{n}\sqrt{p_j} / \left\|\tilde{\beta}_{(j)}\right\|\right\} \to 0 \text{ as } n \to 0$$

Apply Taylor's expansion, we have

$$\begin{aligned} n^{-1/2}\dot{\ell}_{n(j)}(\hat{\beta}_n(\lambda_n)) &\triangleq n^{-1/2}\frac{\partial \ell_n(\hat{\beta}_n(\lambda_n))}{\partial(\beta_{(j)})} \\ &= n^{-1/2}\dot{\ell}_{n(j)}(\beta^0) + \frac{1}{n}\left(\frac{\partial \dot{\ell}_{n(j)}(\beta^*)}{\partial \beta}\right)' \sqrt{n}\left(\hat{\beta}_n(\lambda_n) - \beta^0\right) \\ &= O_p(1) - O_p(1)O_p(1) \\ &= O_p(1) \end{aligned}$$

where β^* is on the line segment between $\hat{\beta}_n(\lambda_n)$ and β^0.

Whereas

$$\lambda_n\sqrt{n}\sqrt{p_j}/\left\|\tilde{\beta}_{(j)}\right\| = \lambda_n\sqrt{n}\sqrt{p_j}/O(n^{-1/2}) = \lambda_n n\sqrt{p_j}O(1) \to \infty \text{ as } n \to 0$$

Thus the probability of $\left[n^{-1/2}\left\|\dot{\ell}_{n(j)}(\hat{\beta}_n(\lambda_n))\right\| \geq \lambda_n\sqrt{n}\sqrt{p_j}/\left\|\tilde{\beta}_{(j)}\right\|\right]$ tends to 0. That is, $\hat{\beta}_{n(j)}(\lambda_n) = 0$ with probability tending to 1.

Secondly, we prove the asymptotic normality.

From the root-n consistency and sparsity property we have just demonstrated, there exists a root-n consistent estimator $\hat{\beta}_n(\lambda_n) = (\hat{\beta}_{n(A)}(\lambda_n)', 0)'$ that satisfies the first KKT condition. That is,

$$\begin{aligned}
0 &= \sqrt{n}\frac{\partial\Phi_n(\hat{\beta}_n(\lambda_n), \lambda_n)}{\partial\beta_{(A)}} \\
&= n^{-1/2}\dot{\ell}_{n(A)}(\beta^0) + \frac{1}{n}\frac{\partial\dot{\ell}_{n(A)}(\beta^*)}{\partial\beta_{(A)}}\sqrt{n}\left(\hat{\beta}_{n(A)}(\lambda_n) - \beta^0_{(A)}\right) \\
&\quad +\lambda_n\sqrt{n}\left(\frac{\beta^0_{(1)}\sqrt{p_1}}{\left\|\tilde{\beta}_{(1)}\right\|\left\|\beta^0_{(1)}\right\|}, \cdots, \frac{\beta^0_{(S)}\sqrt{p_S}}{\left\|\tilde{\beta}_{(S)}\right\|\left\|\beta^0_{(S)}\right\|}\right)' \\
&\quad +\lambda_n\sqrt{n}\text{diag}\left\{\frac{\sqrt{p_1}}{\left\|\tilde{\beta}_{(1)}\right\|\left\|\beta^*_{(1)}\right\|}\left(I_{p_1\times p_1} - \frac{\beta_{(1)}\beta'_{(1)}}{\|\beta^*_{(1)}\|^2}\right), \cdots,\right. \\
&\quad \left.\frac{\sqrt{p_S}}{\left\|\tilde{\beta}_{(S)}\right\|\left\|\beta^*_{(S)}\right\|}\left(I_{p_S\times p_S} - \frac{\beta_{(S)}\beta'_{(S)}}{\|\beta^*_{(S)}\|^2}\right)\right\}\left(\hat{\beta}_{n(A)}(\lambda_n) - \beta^0_{(A)}\right) \\
&= N(0, \Sigma_{(A)}) + o_p(1) - (\Sigma_{(A)} + o_p(1))\sqrt{n}\left(\hat{\beta}_{n(A)}(\lambda_n) - \beta^0_{(A)}\right) \\
&\quad +o(1)O_p(1) + o(1)O(1)o_p(n^{-1/2}) \\
&= N(0, \Sigma_{(A)}) + o_p(1) - \Sigma_{(A)}\sqrt{n}\left(\hat{\beta}_{n(A)}(\lambda_n) - \beta^0_{(A)}\right)
\end{aligned}$$

From the above equation, we have

$$\sqrt{n}\left(\hat{\beta}_{n(A)}(\lambda_n) - \beta^0_{(A)}\right) \xrightarrow{D} N(0, \Sigma^{-1}_{(A)})$$

3.3 Computational implementation

The implementation of the adaptive group lasso method builds on the implementation of the group lasso method. There are a few major additions and modifications.

- Allow input of $\tilde{\beta}$ and can automatically identify which type of $\tilde{\beta}$ is provided. If the input is a constant type, it will calculate the range of λ and select grid points approximately equally spaced on the log scale. If the input is a varying type, it will also look for the input of an λ vector associating with $\tilde{\beta}$ and use it as the grid points.

- Take overflow of floating point into consideration. Since computer does not have exact zeros and the inverse of a small number could cause the overflow of floating point, a cutoff value $\epsilon > 0$ is chosen for $\tilde{\beta}$. If $\left\|\tilde{\beta}_{(j)}\right\| < \epsilon$, then $\left\|\tilde{\beta}_{(j)}\right\|$ is set to 0 and the estimate for the covariate group j will be set to zero according to the first KKT condition. In the implementation, ϵ is chosen to be $1e-10$. This slight modification prevents potentially unintentional interruption of computation. However, this modification might have an impact on the sparsity property if the choice of ϵ is not small enough. The appropriate choice of ϵ depends on the magnitude of the high-dimensional problem, that is, both the sample size and the number of covariates.

- The algorithm for updating the active set is modified correspondingly to match the new KKT conditions.

- The updating algorithm for finding the adaptive group lasso estimate is modified according to the new objective function and its first- and second-order derivatives.

In general, there are four steps to follow in order to find the optimal adaptive group lasso estimate with our computational implementation.

- In step one, a group lasso regression is fit on the data to find the group lasso solution path.
- In step 2, modified AIC, modified BIC, and cross validation criteria are calculated along the solution path to find the optimal tuning parameter and thus the optimal group lasso estimate corresponding to each of the three variable selection criteria.

  Either the whole group lasso solution path, or one of the optimal group lasso estimates, or the regular Cox PH model estimate can be chosen as $\tilde{\beta}$ for the adaptive group lasso model.
- In step three, an adaptive group lasso regression is fit to find the adaptive group lasso solution path.
- In step four, modified AIC, modified BIC, and cross validation criteria are calculated along the solution path to find the optimal tuning parameters and thus the optimal adaptive group lasso estimates.

  More information on the interface of the implementation are provided in the next chapter on generalized group lasso method as both the group lasso and the adaptive group lasso are special cases of the generalized group lasso method.

CHAPTER 4
GENERALIZED GROUP LASSO PENALIZED COX REGRESSION

Both group lasso method and adaptive group lasso method produces sparse and consistent estimate of the model parameters. However, both estimates are somewhat biased towards zero in the finite sample setting. In reality, we might want to know the unbiased effects of some covariates such as treatment. We might also want to put different penalty weights for different covariate groups. In this chapter, we extend the adaptive group lasso method in high-dimensional partially linear additive Cox PH model to reflect these two requirements. The computational implementation of the method builds up on the implementations of the adaptive group lasso and the group lasso.

4.1 Generalized group lasso method

In the adaptive group lasso, the penality is

$$\sum_{j=1}^{K} \frac{\lambda\sqrt{p_j}}{\left\|\tilde{\beta}_{(j)}\right\|} \left\|\beta_{(j)}\right\| \tag{4.1}$$

In the generalized group lasso, we assume there is no penalty on the first F covariates and the remaining covariates are still grouped into K sets. We also use w_j to substitute $\tilde{\beta}_{(j)}$, now the penalty becomes,

$$\sum_{j=1}^{K} \frac{\lambda\sqrt{p_j}}{w_j} \left\|\beta_{(j)}\right\| \tag{4.2}$$

where $w_j > 0, \forall j = 1, 2, \cdots, K$.

In both group lasso and adaptive group lasso, $\beta = (\beta'_{(1)}, \cdots, \beta'_{(K)})'$. However, $\beta = (\beta_1, \cdots, \beta_F, \beta'_{(1)}, \cdots, \beta'_{(K)})'$ in generalized group lasso.

4.1.1 Objective function and KKT conditions

The objective is to minimize the following function,

$$\Phi_n(\beta, \lambda) = -\frac{1}{n}\ell_n(\beta) + \sum_{j=1}^{K} \frac{\lambda\sqrt{p_j}}{w_j} \left\|\beta_{(j)}\right\| \tag{4.3}$$

where $\ell(\beta)$ is the partial likelihood of partially linear additive Cox PH model using B-spline to approximate the nonlinear effects as introduced in the previous chapter.

For a given λ, the adaptive group lasso estimate $\hat{\beta}_n(\lambda)$ is defined as

$$\hat{\beta}_n(\lambda) = \arg\min_{\beta} \Phi_n(\beta, \lambda)$$

By taking the first-order derivative on the objective function $\Phi_n(\beta, \lambda)$, we have derived the following KKT conditions for finding the minimizer of $\Phi_n(\beta, \lambda)$ over β.

$$\begin{cases} \frac{\partial \ell_n(\beta)}{\partial \beta_j} = 0 & \forall j = 1, \cdots, F \\ -\frac{1}{n}\frac{\partial \ell_n(\beta)}{\partial \beta_{(j)}} + \frac{\lambda\sqrt{p_j}}{w_j}\frac{\beta_{(j)}}{\|\beta_{(j)}\|} = 0 & \forall \beta_{(j)} \neq 0 \text{ and } w_j \neq 0 \\ \beta_{(j)} = 0 & \forall w_j = 0 \\ \|\frac{1}{n}\frac{\partial \ell_n(\beta)}{\partial \beta_{(j)}}\| \leq \frac{\lambda\sqrt{p_j}}{w_j} & \forall \beta_{(j)} = 0 \text{ and } w_j \neq 0 \end{cases} \tag{4.4}$$

The KKT conditions are not only necessary but also sufficient to minimize the objective function.

4.2 Asymptotic Results

Similar to the proofs in the previous chapter, we have the same asymptotic results on the generalized group lasso method with an additional assumption on $w_1, \cdots, w_K$.

The addition assumption is that $\min_{j=1}^{K} w_j > 0$ or $w_1, \cdots, w_K$ are consistent estimates of $\left\|\beta^0_{(1)}\right\|, \cdots, \left\|\beta^0_{(K)}\right\|$

Theorem 4.1. *(Consistency) Under regularity conditions (A) - (D) in Andersen and Gill (1982), if $\lambda_n \to 0$ as $n \to 0$, then there exists a maximizer $\hat{\beta}_n(\lambda_n)$ of $\Phi_n(\beta, \lambda_n)$ such that $\hat{\beta}_n(\lambda_n) \xrightarrow{p} \beta^0$ as $n \to \infty$.*

Theorem 4.2. *(Convergence rate) Under regularity conditions (A) - (D) in Andersen and Gill (1982), if $\lambda_n = O(n^{-1/2})$, then $\left\|\hat{\beta}(\lambda_n) - \beta^0\right\| = O_p\left(n^{-1/2}\right)$*

Theorem 4.3. *(Oracle property) Assume the regularity conditions (A) - (D) in Andersen and Gill(1982) and assume there exist $\eta > 0$ and $q \in (0.5, 1)$ such that $\lambda_n n^q \to \eta$ as $n \to \infty$. From Theorem 3.2, we know $\hat{\beta}_n(\lambda_n)$ is a root-n consistent estimate of β^0.*

(a) (Sparsity) For $\forall j \in \bar{A}$, i.e., $\beta_{(j)} = 0$, let $\hat{\beta}_{n(j)}(\lambda_n)$, a sub vector of $\hat{\beta}_n(\lambda_n)$, be the estimator of $\beta_{(j)}$, then $\hat{\beta}_{n(j)}(\lambda_n) = 0$ with probability tending to 1.

(b) (Asymptotic Normality) Let $\hat{\beta}_{n(A)}(\lambda_n)$, a sub vector of $\hat{\beta}_n(\lambda_n)$, be the estimator of all $\{\beta_j; j = 1, \cdots, F\}$ and $\{\beta_{(j)}; \beta_{(j)} \neq 0\}$, then $\sqrt{n}(\hat{\beta}_{n(A)}(\lambda_n) - \beta_{(A)}) \xrightarrow{D} N(0, \Sigma^{-1}_{(A)})$.

4.3 Bias and standard error

The bias and standard error can be estimated via the following estimation

$$\dot{\ell}_n(\hat{\beta}_n(\lambda_n)) = \dot{\ell}_n(\beta^0) + \ddot{\ell}_n(\beta^*)\left(\hat{\beta}_n(\lambda_n) - \beta^0\right)$$

where β^* is on the line segment between $\hat{\beta}_n(\lambda_n)$ and β^0.

Since $n^{-1}\dot{\ell}_n(\beta^0) \xrightarrow{a.s.} 0$, $n^{-1}\ddot{\ell}_n(\beta^*) \xrightarrow{p} \Sigma$ which is a positive definite matrix, and $\ddot{\ell}_n(\beta^*)$ can be approximated by $\ddot{\ell}_n(\hat{\beta}_n(\lambda_n))$, thus the bias of $\hat{\beta}_n(\lambda_n)$ can be estimated by

$$\left(\ddot{\ell}_n(\hat{\beta}_n(\lambda_n))\right)^{-1}\left(\dot{\ell}_n(\hat{\beta}_n(\lambda_n))\right)$$

It also implies that the estimate for a free variable is unbiased.

In a parametric regression without a penalty, we have $\dot{\ell}_n(\hat{\beta}_n(\lambda_n)) \equiv 0$. In our generalized group lasso penalized regression, $\dot{\ell}_n(\hat{\beta}_n(\lambda_n))$ is determined jointly by the penalty term and the optimal tuning parameter λ_n, which is independent of any random variables. Thus, the covariance of $\hat{\beta}_n(\lambda_n)$ can be approximated by

$$\left(\ddot{\ell}_n(\hat{\beta}_n(\lambda_n))\right)^{-1} Cov\left(\dot{\ell}_n(\beta^0)\right)\left(\ddot{\ell}_n(\hat{\beta}_n(\lambda_n))\right)^{-1}$$

According to Lin and Wei (1989), we can estimate $Cov\left(\dot{\ell}_n(\beta^0)\right)$ with $\sum_{i=1}^n W_i(\hat{\beta}_n(\lambda_n))^{\otimes 2}$, where

$$W_i(\beta) = \Delta_i\left\{X_i - \frac{\sum_{j=i}^n s^*_{1,i}}{\sum_{j=i}^n s^*_{0,i}}\right\} - \sum_{j=1}^{i}\frac{\Delta_j s^*_{0,i}}{n\sum_{k=j}^n s^*_{0,k}}\left\{X_i - \frac{\sum_{k=j}^n s^*_{1,k}}{\sum_{k=j}^n s^*_{0,k}}\right\} - \frac{1}{n}\dot{\ell}_n(\hat{\beta}_n(\lambda_n))$$

Thus the variance of $\hat{\beta}_n(\lambda_n)$ can be estimated by

$$\left(\ddot{\ell}_n(\hat{\beta}_n(\lambda_n))\right)^{-1}\sum_{i=1}^n W_i(\hat{\beta}_n(\lambda_n))^{\otimes 2}\left(\ddot{\ell}_n(\hat{\beta}_n(\lambda_n))\right)^{-1}$$

When $\ddot{\ell}_n(\hat{\beta}_n(\lambda_n))$ is not positive definite, it can be replaced by $\ddot{\Phi}_n(\hat{\beta}_n(\lambda_n), \lambda_n)$ where the derivative is with respect to the first parameter.

4.4 Computational implementation

The implementation of the generalized group lasso method builds on that of the adaptive group lasso method. As introduced at the beginning of this chapter, there are a couple of modifications.

- Allow covariates in the model whose coefficients are not penalized.
- Allow the input the weight $w_j, j = 1, \cdots, K$, and identify the type of weight being used.
- Modify the algorithm to update the active set according to the new KKT conditions.
- Update the modified Newton-Raphson method to reflect the new KKT conditions.

Below is an example of R code to fit a generalized group lasso method using the PBC data as introduced in Chapter 1.

```
################
# Group lasso
################

# Solution path
    fit1 = gs.coxph( Surv(time, status) ~ trt + bs(age, df=6)
         + sex + edema + bs(bili, df = 6) + bs(protime, df = 6)
         + bs(albumin, df = 6) + spiders +  ascites, data=pbc1,
         lambda = NULL, debugtype = 0, n.freevars= 1,
         group.selection = c(6,1,1,6,6,6,1,1),
         control = list(iter.max = 50))

# Cross validation
    set.seed(175445)
    cv1 = gs.coxph( Surv(time, status) ~ trt + bs(age, df=6)
         + sex + edema + bs(bili, df = 6) + bs(protime, df = 6)
         + bs(albumin, df = 6) + spiders +  ascites, data=pbc1,
         lambda = fit1$Lambdas, debugtype = 2,
         n.freevars= 1, group.selection = c(6,1,1,6,6,6,1,1),
         control = list(iter.max = 50))

# Optimal tuning parameters
    nobs = NROW(pbc1)
    cvs1 = apply(cv1, 1, sum)/nobs
    ind1 = which.min(cvs1)
    aic1 = 2*fit1$Objectives1 + 2*fit1$Ds/nobs
    bic1 = 2*fit1$Objectives1 + 2*fit1$Ds*log(nobs)/nobs
    aind1 = which.min(aic1)
    bind1 = which.min(bic1)

################
# Adaptive group lasso - Type 1
################

# Solution path
    fit2 = gs.coxph( Surv(time, status) ~ trt + bs(age, df=6)
         + sex + edema + bs(bili, df = 6) + bs(protime, df = 6)
         + bs(albumin, df = 6) + spiders +  ascites, data=pbc1,
         lambda = NULL, LassoWGT = fit1$Betas[ind1,], debugtype = 0,
         n.freevars= 1, group.selection = c(6,1,1,6,6,6,1,1),
         control = list(iter.max = 50))

# Cross validation
    set.seed(11010013)
    cv2 = gs.coxph( Surv(time, status) ~ trt + bs(age, df=6)
         + sex + edema + bs(bili, df = 6) + bs(protime, df = 6)
         + bs(albumin, df = 6) + spiders +  ascites, data=pbc1,
        lambda = NULL, LassoWGT = fit1$Betas[ind1,], debugtype = 2,
        n.freevars= 1, group.selection = c(6,1,1,6,6,6,1,1),
        control = list(iter.max = 50))

# Optimal tuning parameters
    cvs2 = apply(cv2, 1, sum)/nobs
    ind2 = which.min(cvs2)
    aic2 = 2*fit2$Objectives1 + 2*fit2$Ds/nobs
    bic2 = 2*fit2$Objectives1 + 2*fit2$Ds*log(nobs)/nobs
    aind2 = which.min(aic2)
    bind2 = which.min(bic2)


################
# Adaptive group lasso - Type 2
################

# Solution path
    fit3 = gs.coxph( Surv(time, status) ~ trt + bs(age, df=6)
         + sex + edema + bs(bili, df = 6) + bs(protime, df = 6)
         + bs(albumin, df = 6) + spiders +  ascites, data=pbc1,
        lambda = NULL,
        LassoWGT = list(Lambdas = fit1$Lambdas, Betas = fit1$Betas),
        debugtype = 0, n.freevars= 1,
        group.selection = c(6,1,1,6,6,6,1,1),
        control = list(iter.max = 50))

# Cross validation
    set.seed(11010017)
    cv3 = gs.coxph( Surv(time, status) ~ trt + bs(age, df=6)
         + sex + edema + bs(bili, df = 6) + bs(protime, df = 6)
         + bs(albumin, df = 6) + spiders +  ascites, data=pbc1,
        lambda = NULL,
        LassoWGT = list(Lambdas = fit1$Lambdas, Betas = fit1$Betas),
        debugtype = 2, n.freevars= 1,
        group.selection = c(6,1,1,6,6,6,1,1),
        control = list(iter.max = 50))

# Optimal tuning parameters
    cvs3 = apply(cv3, 1, sum)/nobs
    ind3 = which.min(cvs3)
    aic3 = 2*fit3$Objectives1 + 2*fit3$Ds/nobs
    bic3 = 2*fit3$Objectives1 + 2*fit3$Ds*log(nobs)/nobs
    aind3 = which.min(aic3)
    bind3 = which.min(bic3)
```

In the above R code, *gs.coxph* is the R function to complete all the major computations in the following steps.

1. Fit a group lasso penalized partially linear additive Cox PH model with the first covariate *trt* not being penalized.

2. Calculate the variable selection criteria including modified AIC, modified BIC, and cross validation along the solution path.

3. Using the optimal tuning parameter chosen by cross validation and the same Cox model from the group lasso method, calculate an adaptive group lasso penalized solution path.

4. Calculate the three variable selection criteria along the adaptive group lasso solution path.

5. Using the whole group lasso solution path and the corresponding grid points on λ, calculate another type of adaptive group lasso penalized solution path.

6. Calculate the three variable selection criteria along this adaptive group lasso solution path.

4.5 Concluding remarks

In this thesis work, we have studied and implemented step-by-step the generalized group lasso method in the high dimensional partially linear additive Cox Model. The method works in both the $p < n$ setting and the $p > n$ setting. It allows both free covariates and penalized covariates coexisting in one model. The penalized covariates can be classified into groups based on the nature of the covariate structure with each covariate group including one or multiple covariates.

Each covariate group can have different weights based on modeling objectives and on pre-existing knowledge of the covariates groups.

We prove that the generalized group lasso method in a Cox PH model enjoys estimation consistency and variable selection consistency. The computational implementation of the generalized group lasso in R environment is also described.

CHAPTER 5
NUMERICAL STUDIES

In this chapter, simulations are conducted to study the parameter estimation and variable selection properties of the three methods, the group lasso, the adaptive group lasso, and the generalized group lasso. The generalized group lasso method is then applied to a gene expression dataset, the Norway/Stanford breast tumor dataset.

5.1 Simulation study

A simulation study is carried out to evaluate the parameter estimation and model selection performance of the three penalized methods. The underlying hazard function is

$$\lambda(t|X, w_1, \beta) = \exp(X\beta + \log(w_1))$$

where the first five elements of β are $(-0.5, 1, 1.5, 0.2, -1.5)$ and the remaining 45 elements of β are zeros. All elements in X, $X1 - X50$, are independently sampled from $Unif\,[0, 1]$ and w_1 is sampled from $Unif\,[0.1, 1.1]$. In addition, X and w_1 are independently sampled. About 2% of the data is right censored with the censoring time being independent of the survival outcome.

5.1.1 Study I with a sample size of 200

200 datasets with a sample size of 200 are simulated. For each dataset, two proportional partially linear additive Cox models are fitted. Model 1 includes the

5 nonzero covariates, $X1-X5$, the 45 zero covariates $X6-X50$, and the B-spline approximation of the nonparametric effect of w_1. The B-spline approximation is chosen to have 6 columns of basis matrix, which forms a covariate group. $X1-X50$ are one variable per group. Model 2 is similar to Model 1 except that $X1$ is a free variable which does not contribute to the penalty. A group lasso penalized method is applied to each model. An optimal group lasso estimate is selected with 5-fold cross validation. Then, this optimal group lasso estimate functions as $\tilde{\beta}$ in the adaptive group lasso. Again, an optimal adaptive group lasso estimate is selected with 5-fold cross validation. The average parameter estimation bias and error are reported in Table 5.1 and the percentage of time that a variable is selection into an optimal model is reported in Table 5.2.

From Table 5.1, we know that the penalized estimates for nonzero effects are all biased towards zero. The adaptive lasso estimates have smaller bias and larger variance compared to the group lasso estimates. the bias of $X1$ is close to zero when it is a free variable, i.e., not penalized.

In Table 5.2, the probability of a nonzero effect being selected into the optimal model is decreased with the adaptive group lasso method. The decrease is ignorable for covariates with relatively larger effects and is noticeable for covariates with relatively smaller effects. However, with the adaptive group lasso method, the probability of a zero effect being selected into the optimal model is also smaller.

Table 5.1: Average parameter estimation bias and error: sample size = 200, N of repetitions = 200.

	$X1$ is penalized		$X1$ is a free variable	
Covariates	Group Lasso	Adaptive Group Lasso	Group Lasso	Adaptive Group Lasso
$X1(-0.5)$	0.28(0.21)	0.17(0.34)	0.04(0.26)	-0.04(0.30)
$X2(1.0)$	-0.43(0.25)	-0.12(0.33)	-0.43(0.26)	-0.09(0.34)
$X3(1.5)$	-0.46(0.28)	-0.06(0.31)	-0.45(0.28)	-0.02(0.33)
$X4(0.2)$	-0.14(0.15)	-0.10(0.22)	-0.14(0.15)	-0.10(0.24)
$X5(-1.5)$	0.45(0.29)	0.04(0.32)	0.45(0.28)	0.00(0.32)
$X6 - X50(0)$	0.00(0.10)	0.00(0.15)	0.00(0.10)	0.00(0.16)

Table 5.2: Percentage of time that a variables is selected into the optimal model: sample size = 200, N of repetitions = 200.

	$X1$ is penalized		$X1$ is a free variable	
Covariates	Group Lasso	Adaptive Group Lasso	Group Lasso	Adaptive Group Lasso
$X1(-0.5)$	75	59		
$X2(1.0)$	99	99	99	99
$X3(1.5)$	100	100	100	100
$X4(0.2)$	38	23	40	28
$X5(-1.5)$	100	100	100	100
$w_1(logw_1)$	100	100	100	100
$X6 - X50(0)$	28	15	28	16

Table 5.3: Average parameter estimation bias and error: sample size = 600, N of repetitions = 200.

	$X1$ is penalized		$X1$ is a free variable	
Covariates	Group Lasso	Adaptive Group Lasso	Group Lasso	Adaptive Group Lasso
$X1(-0.5)$	0.16(0.16)	0.08(0.20)	0.05(0.16)	-0.01(0.17)
$X2(1.0)$	-0.17(0.16)	-0.03(0.17)	-0.18(0.16)	-0.01(0.17)
$X3(1.5)$	-0.21(0.16)	-0.02(0.16)	-0.21(0.16)	0.00(0.16)
$X4(0.2)$	-0.09(0.12)	-0.09(0.15)	-0.09(0.12)	-0.08(0.16)
$X5(-1.5)$	0.21(0.16)	0.02(0.15)	0.21(0.15)	0.00(0.16)
$X6 - X50(0)$	0.00(0.08)	0.00(0.08)	0.00(0.07)	0.00(0.08)

5.1.2 Study II with a sample size of 600

Simulation study I is repeated with a sample size of 600 to examine the parameter estimation and variable selection properties of the three method. The results are reported in Tables 5.3 and 5.4.

The patterns observed in Study I are also found in Study II. Comparing the two studies with different sample sizes, we can see that as the sample size increases, the estimation biases and errors decrease and the probabilities of selecting the nonzero effects and excluding the zero effects increase. The simulation results support the theoretic results in the previous chapters.

Table 5.4: Percentage of time that a variables is selected into the optimal model: sample size = 600, N of repetitions = 200.

	$X1$ is penalized		$X1$ is a free variable	
Covariates	Group Lasso	Adaptive Group Lasso	Group Lasso	Adaptive Group Lasso
$X1(-0.5)$	97	93		
$X2(1.0)$	100	100	100	100
$X3(1.5)$	100	100	100	100
$X4(0.2)$	71	47	70	51
$X5(-1.5)$	100	100	100	100
$w_1(log w_1)$	100	100	100	100
$X6 - X50(0)$	22	11	23	11

5.1.3 Nonparametric estimation of a nonlinear effect

Same as in the simulation study II, a data set with a sample size of 600 is generated. The same group lasso and adaptive group lasso penalized models with $X1$ being a free covariate are fitted. The nonparametric estimations of $w1$'s effect and the true effect $\log(w1)$ are plotted in Figure 5.1. We can see that within a relatively larger range, the adaptive group lasso estimated effect is close to the true underlying effect. However, at both end of the range of $w1$, the estimation error increases dramatically.

5.2 Real data example

In this section, the group lasso penalized grouped variable selection method is applied to the Norway/Stanford Breast Tumors dataset.

The survival model includes ER status and tumor grade greater than 3 as free variables, age is modeled nonparametrically using B-spline approximation with 6 degrees of freedom, and the 825 gene sets. The effects of age and the 825 gene sets are penalized as covariate groups.

A group lasso method is first fitted on the dataset and the 5-fold cross validation is then performed to select the optimal model. Age is excluded from the optimal model. 6 out of the 825 gene sets are selected into the optimal model. The total number of gene features in the 6 gene sets is 155, and the number of unique genes in the 6 sets is 130.The gene sets and the number of gene features mapped into each gene set are listed in Table 5.5.

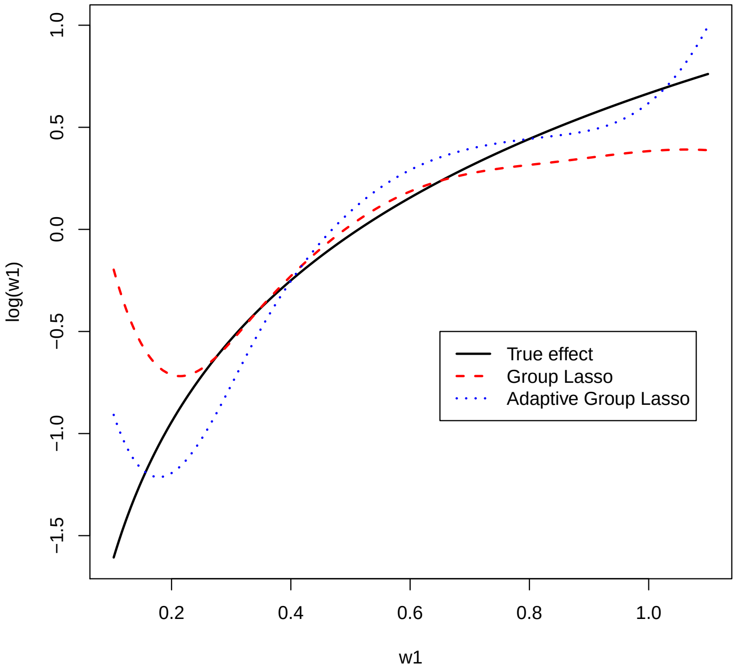

Figure 5.1: Effects of $w1$ true effect, estimated effect with group lasso method, and estimated effect with adaptive group lasso method. All the effects are normalized to have a zero mean.

Table 5.5: The 6 gene sets chosen by the group lasso method

Gene set name	N of genes features
G PROTEIN SIGNALING ADENYLATE	10
EPITHELIAL CELL DIFFERENTIATION	7
DEFENSE RESPONSE	54
PROTEIN TARGETING TO MITOCHONDRION	11
PROTEIN POLYUBIQUITINATION	10
CYCLASE INHIBITING PATHWAY	
RESPONSE TO WOUNDING	63

The gene features in each of the 6 gene sets are listed below in the same order as in Table 5.5.

- CHRM5 DRD2 HR NP NPY NPY1R OPRK1 PRL RGS1 T
- EHF ELF3 F3 KRT4 MB T UPK1A
- ADM ADORA1 AFAP1 AFAP1L2 AHSG AIF1 ALB AR BLNK C2 C5 CAD CADM1 CCR5 CCR6 CD4 CD40 CD48 CDC2 CFP CLEC5A COLEC12 CR2 CX3CR1 DCDC2 F11 F11R F2 FAP GAL HR KL KLRC2 KLRC3 LGALS3 LGALS3BP LILRA2 LILRB2 LILRB3 LTB MST1 MST1R NCF2 RAC1 S100A7 S100A8 S100A9 T TPST1 TYR TYROBP VEZF1 VPS45 WFDC1

- F3 MFN2 MIPEP T TIMM17A TIMM17B TIMM23 TIMM44 TOMM34 TRNT1 TSPO
- AMFR DDB2 ERCC8 HUWE1 STUB1 T TRAF6 UBE2D1 UBE2V2 UBE3C
- ACHE ADORA1 AIF1 AIP AR C3AR1 CBL CCL3 CCR5 CD36 CD4 CD40 CDO1 CR2 CTGF CX3CR1 CXCL1 CXCL10 DCBLD2 F10 F11 F11R F12 F13A1 F2 F2R F2RL2 F2RL3 F5 F7 F8 GC GGCX GNA12 GP9 HNF4A HR IL13 IL1A L3MBTL4 LTB MB MMRN1 NFE2 NFE2L1 REL RELA S100A8 S100A9 SERPINE1 T TF TFPI TG TGFB1 TGFB2 TH THBD TNF TNFAIP6 TPST1 VPS45 XCL1

We then apply the adaptive group lasso method on the dataset. The model includes the 2 free variables and the 6 gene sets selected by group lasso. The group lasso estimate acts as the adaptive weight. The 5-fold cross validation selects the first 3 gene sets as listed in the above table into the final model. The three biological processes identified in the final model are related to the effectiveness of treatment, tumor progression, and tumor-inflammatory response in the advanced breast cancer patients as reported by Filardo et al. (2002), Hudson et al (2001), and Aaroe et al. (2010)

The number of active variables and the number of active variable groups along the group lasso penalized solution path are plotted in Figure 5.2. The penalized solution paths are plotted in Figure 5.3. The effective number of parameters and the variable selection criteria are plotted in Figures 5.4 and 5.5.

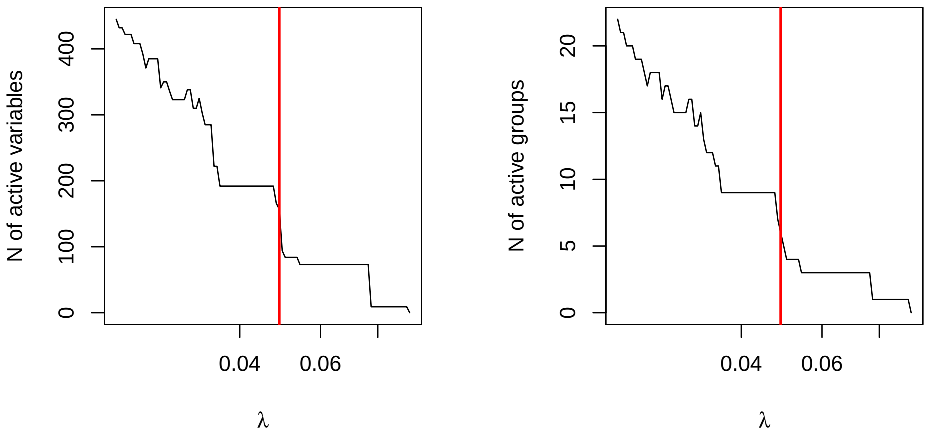

Figure 5.2: Number of active variables and Number of active variable groups along the group lasso penalized solution path. The vertical lines indicate the optimal model chosen by the 5-fold cross validation.

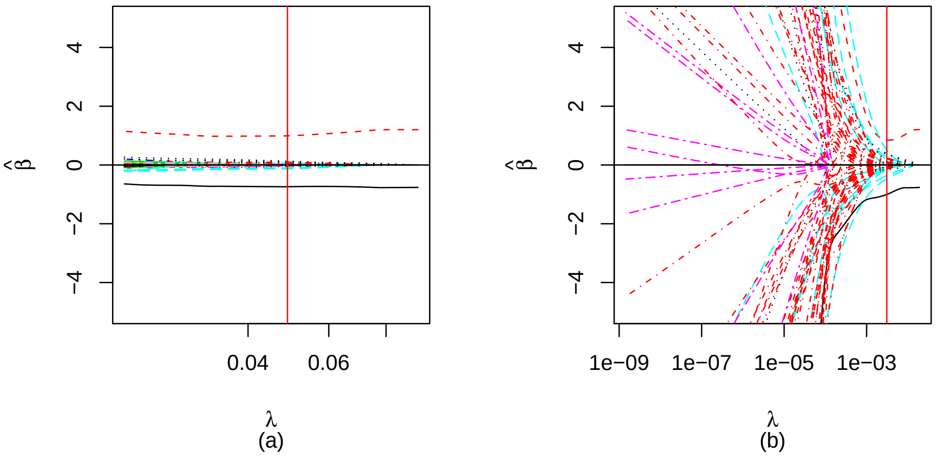

Figure 5.3: Penalized solution paths. (a)-Group lasso, (b)-Adaptive group lasso. The vertical line indicates the optimal model chosen by 5-fold cross validation.

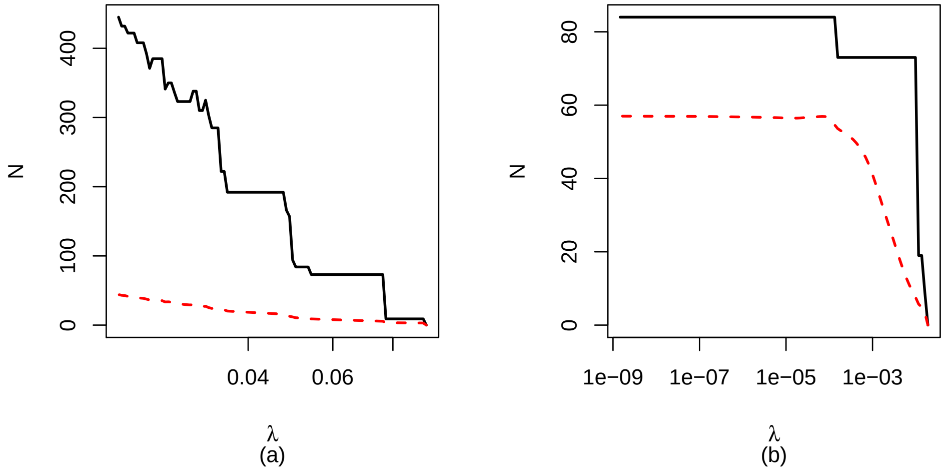

Figure 5.4: Effective number of parameters in solid curve and number of active covariates in dashed curve.(a)-Group lasso, (b)-Adaptive group lasso.

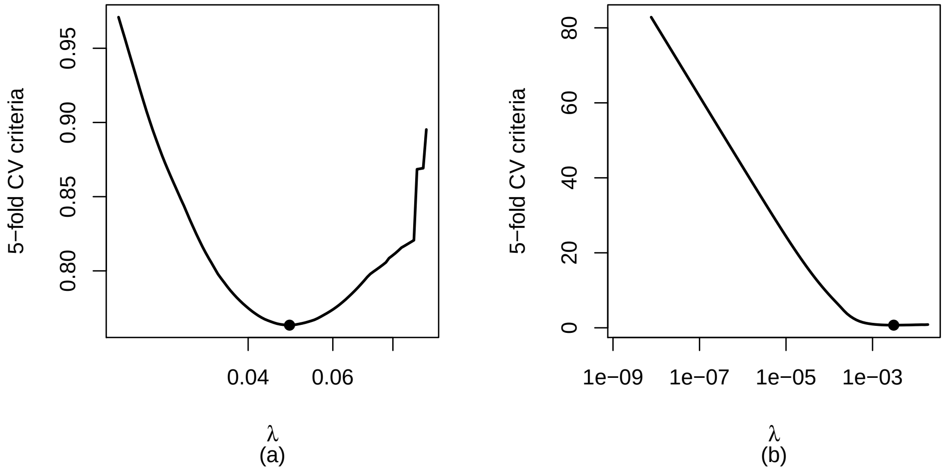

Figure 5.5: Plot of 5-fold CV criteria along the solution path. The solid dot indicate the lowest point on the curve, i.e., the optimal model. (a)-Group lasso, (b)-Adaptive group lasso.

CHAPTER 6
DISCUSSION AND FUTURE WORK

With the rapid advance of new technologies, a huge amount of information in various fields are collected, stored, and available for statistical analysis. The goal of statistical analyses is either to enhance the understanding of a biological process, or to improve some business practice, or to show the safety and efficacy profile of an investigational substance targeting at a disease, and so on. These needs push statistical analysis towards high-dimensional data and performing variable selections at grouped-variable level.

Some reports of the work on high-dimensional data analysis exist in the literature. Most of them are focused on linear regressions because of the relative lower burden in computational implementation. Not much work on high-dimensional survival analysis is reported in the literature and none of them studies the whole solution path to understand the nature of the issue.

We study the group lasso penalized method in the high-dimensional partially linear additive Cox model, propose and implement an efficient algorithm to find the solution path and to locate the optimal tuning parameter using modified AIC, modified BIC, and cross validation criteria. We extend the group lasso method to the adaptive group lasso and then to the generalized group lasso. The extension provides more flexibility in the statical modeling of high-dimensional survival data. we demonstrate that adaptive group lasso and generalized group lasso enjoys parameter estimation and variable selection consistency under some regularity conditions.

This work opens the possibility of more future work along the line. For example, we can take the covariate correlation structure into consideration in the penalty. We can also examine the asymptotic properties when the number of covariates goes to infinity as the sample size goes to infinity. The derivation of the asymptotic results assumes a convergence rate on the optimal tuning parameter λ_n, however the optimal tuning parameter is chosen with some variable selection criterion in the computational implementation. There is a possibility to eliminate this discrepancy, as what Shao (1993) did to prove that cross validation is able to find the best predictive model under some conditions in linear regression.

The computational implementation works well on a Thinkpad T400 with 4G of memories when the number of covariates p is less than 6×10^4 and the number of covariates in the active set is less then 500. If there is a need to identify more than 500 nonzero covariates, the inverse of the second order derivative needs to be improved with more efficient algorithm. If $p > 6 \times 10^4$, we might need to give up the R environment and implement the procedure purely with the C language. It is also recommended by others to implement the procedure using the combination of R, MySQL and Python when the number of observations is in the order of millions.

In conclusion, this thesis work extends the penalized regression method in high dimensional survival analysis to enjoy more freedom in terms of allowing free covariates, specifying covariate groups, and incorporating prior knowledge into the penalty. It opens an foundation for more future work in this field.

REFERENCES

[1] Aalen, O. Borgan, O. and Gjessing, H.(2007). Survival and event history analysis: A process point of view. Springer,

[2] Aaroe, J. Lindahl, T. Dumeaux, V. Sobo, S. Tobin, D. Hagen, N. Skaane, P. Lonneborg, A. Sharma, P. and Borresen-Dale, A.L. (2010). Gene expression profiling of peripheral blood cells for early detection of breast cancer. *Breast Cancer Research* , **12**, 1–11.

[3] Abdelmalek, N. N. (1971). Linear ℓ_1 approximation for a discrete point set and ℓ_1 solutions of overdetermined linear equations. *Journal of the Association for Computing Machinery*, **18**, 41–47.

[4] Bakin, S. (1999). Adaptive regression and model selection in data mining problems. *PHD Thesis.* Australian National University, Canberra.

[5] Boyd, S. and Vandenberghe, L. (2004). Convex optimization. Cambridge University Press.

[6] Boor, C.D. (2001). A practical guide to Splines. Springer Verlag, New York.

[7] Breiman, L. (1996). The heuristics of instability in model selection. *Annals of Statistics*, **24**, 2350–2383.

[8] Efron, B. (1977). Efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, **72**, 557–565.

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

[10] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1360.

[11] Filardo, E.J., Quinn, J.A. Frackelton, A.R. and Bland, K.I(2002). Estrogen action via the G protein-coupled receptor, GPR30: stimulation of adenylyl cyclase and cAMP-mediated attenuation of the epidermal growth factor receptor-to-MAPK signaling axis. *Molecular Endocrinology* , **16**, 70–84.

[12] Geisler, S. Lonning, P.E. Aas, T. Johnsen, H. Fluge, O. Haugen, D.F. Lillehaug, J.R. Akslen, L.A. and Borresen-Dale, A.L. (2001). Influence of TP53

gene alterations and c-erbB-2 expression on the response to treatment with doxorubicin in locally advanced breast cancer. *Cancer Research*, **61**, 2505–2512.

[13] Geisler, S. Borresen-Dale, A.L. Johnsen, H. Aas, T. Geisler, J. Akslen, L.A. Anker, G. and Lonning, P.E. (2003). TP53 gene mutations predict the response to neoadjuvant treatment with FUMI in locally advanced breast cancer. *Cancer Research*, **9**, 5582–5588.

[14] Huang, J.(1999). Efficient estimation of the partly linear additive cox model. *Annals of Statistics*, **27**, 1536–1563.

[15] Huang, J. Ma, S. Xie, H. Zhang, C. (2009). A Group Bridge Approach for Variable Selection. *Biometrika*, **96(2)**, 339–335.

[16] Huang, J. Horowitz, J.L. Wei, F. (2009). Variable selection in nonparametric additive models. *Annals of Statistics*, **96(2)**, 339–335.

[17] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

[18] Hudson, D.L. Guy, A.T. Fry, P. O'Hare, M.J. Watt, F.M. and Masters, J.R. (2001). Epithelial Cell Differentiation Pathways in the Human Prostate: Identification of Intermediate Phenotypes by Keratin Expression . *Journal of Histochemistry and Cytochemistry* , **49**, 271–278.

[19] Lu, M.G.(2007). Analysis of panel count data using monotone polynomial splines. Doctoral Dissertation, The University of Iowa, Iowa City, Iowa.

[20] Kalbfleisch, J.D. and Prentice, R.L. (2002). The statistical analysis of failure time data, second edition. John Wiley & Sons, New York.

[21] Lin, D.Y. and Wei, L.J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, **84**, 1074–1078.

[22] Moody, J. E. Hanson, S. J. and Lippmann, R. P.(1992). The effective number of parameters, an analysis of generalization and regularization in nonlinear learning system. *Advances in Neural Information Processing System*, **4**, 847–854.

[23] Nardi, Y. and Rinaldo, A. (2008) On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, **2**, 605–633

[24] Schumaker, L. (1981). Spline Functions: Basic Theory. John Wiley & Sons, New York.

[25] Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, **88**, 486–494.

[26] Stone, C.J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, **14**, 590–606.

[27] Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**, 385–395.

[28] van der Vaart, A. W. (1998). Asymptotic Statistics. Cambridge University Press.

[29] van der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence and empirical processes, with applications to statistics. Springer.

[30] Xie, H.(2007). Regression with smoothly clipped absolute deviation penalty. Doctoral Dissertation, The University of Iowa, Iowa City, Iowa.

[31] Yasrebi. H. Sperisen, P. Praz, V. Bucher, P. (2009). Can survival prediction be improved by merging gene expression data sets? *PLoS ONE*, **4(10)**, e7431.

[32] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, **68**, 49–67.

[33] Zhang, H. and Lu W. (2007). Adaptive lasso for Cox proportional hazards model. *Biometrika*, **94**, 691-703.

[34] Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, **7**, 2541–2563.

[35] Zou, H. and Hastie, T.(2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301–320.