

ABSTRACT

GUO, ZIFANG. Variable Selection and Dimension Reduction for High Dimensional Complex Data. (Under the direction of Dr. Lexin Li and Dr. Wenbin Lu.)

In high dimensional data analysis, there often exist complex data which add further complications to variable selection and/or dimension reduction problems. In this dissertation, we study methods to address three different such complications. In the first part (Chapter 2), we focus on variable selection for censored survival data with high dimensional predictors, and propose a forward stagewise shrinkage and addition method for simultaneous model estimation and variable selection in Cox proportional hazards models. The proposed method carries out an additive stagewise modeling while introducing shrinkage at each iteration. Our proposal extends a popular statistical learning technique, the boosting method, by explicitly performing variable selection and substantially reducing the number of iterations required for algorithm completion. It also inherits the flexible nature of the boosting and is straightforward to extend to nonlinear Cox models. Our numerical analyses demonstrate that the new method enjoys an equally competitive performance as the best players of the existing solutions in Cox models with $p < n$, whereas it achieves a considerably superior performance than the alternative solutions when $p > n$.

In the second part of this dissertation (Chapter 3), we consider situations where the predictors come from different groups. It is often desirable to incorporate such prior group information into dimension reduction procedures to obtain more interpretable and more informative estimates. We propose a groupwise sufficient dimension reduction method which preserves full regression information in the conditional distribution of response given the predictors, while incorporating the group structure in the predictors during

reduction. The proposed method is based on imposing a group structure onto classical central subspace estimators via a direct sum envelope. Simulation studies and real data analysis show that the proposed method achieves competitive performance, in terms of both estimation accuracy and interpretability of the resulting estimator.

In the last part (Chapter 4), we focus on optimal treatment strategy estimation in the survival framework, and propose the use of variable selection method with the kernel machine Cox proportional hazards model in estimating optimal treatment strategy. The proposed method allows nonparametric specification of the baseline covariate effects and leads to improved decision rules by incorporating shrinkage based variable selection method in estimation. Simulation studies are provided to illustrate the empirical performance of the proposed method.

© Copyright 2012 by Zifang Guo

All Rights Reserved

Variable Selection and Dimension Reduction
for High Dimensional Complex Data

by
Zifang Guo

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2012

APPROVED BY:

Dr. Howard D. Bondell

Dr. Brian J. Reich

Dr. Lexin Li
Co-chair of Advisory Committee

Dr. Wenbin Lu
Co-chair of Advisory Committee

DEDICATION

To my beloved family.

BIOGRAPHY

Zifang Guo was born in Tianjin, China in December 1983, and spent the first eighteen years of her life in Tianjin. She graduated from Yaohua High School in 2001 and attended Tsinghua University afterwards for her undergraduate study in Biological Sciences. After obtaining her Bachelor's degree in 2005, she came to United States and studied Biophysics at Boston University. Graduated with a M.A. in Biophysics in 2007, she decided to pursue her Ph.D. degree in Statistics at North Carolina State University. Her doctoral dissertation research at NCSU is under the direction of Dr. Lexin Li and Dr. Wenbin Lu. During her graduate study, she completed a summer internship at Self-Help Credit Union in 2009, a summer internship at Merck & Co. in 2010, and worked as Graduate Industrial Trainee at GlaxoSmithKline from 2010 to 2011. She will complete her Ph.D. in November 2011.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisors Dr. Lexin Li and Dr. Wenbin Lu for their enormous support. It is a great fortune and honor of mine to have them as my advisors. Their inspiration, encouragement, patience, enthusiasm, and immense knowledge guided me through every step of my research and writing of this dissertation. I could not have imagined having a better advisor for my Ph.D. study.

I owe my sincere gratitude to Dr. Marie Davidian, for being supervisor of my research assistantship. It was not only her generous support that helped me through my study, but also her attitude as a great researcher and her ways of critical thinking that first guided me into statistical research.

I thank Dr. Howard Bondell for being my masters' degree academic advisor during my first two years of study at NCSU, and for serving on my thesis committee. His valuable guidance helped me plan my study well and his insightful comments helped me improve this research.

I thank Dr. Brian Reich for being my committee member, and for his encouragement, inspirational discussions, and valuable suggestions for this research.

I am also grateful for the faculty members in the department for offering great graduate lectures on various topics that led me into statistics and broadened my knowledge in this area, as well as made statistics so fascinating to me.

Special thanks to all my good friends who shared laughs and tears with me in all these years. My life would have been much less colorful without their accompanies.

Last but not least, none of this would have been made possible without my family. My parents have always believed in me and supported me to pursue my dreams. My husband Lei always stands by me and shares the ups and downs in our lives with me. My

daughter Claire has cheered me along the way. I am grateful for the the unconditional love and support from them.

TABLE OF CONTENTS

List of Tables	viii
List of Figures	x
Chapter 1 Introduction	1
Chapter 2 Forward Stagewise Shrinkage and Addition for High Dimensional Censored Regression	5
2.1 Introduction	5
2.2 Model and Method	11
2.2.1 Cox Proportional Hazards Model with Adaptive LASSO	11
2.2.2 A Modified Boosting Algorithm	12
2.2.3 Forward Stagewise Shrinkage and Addition	14
2.3 Simulations	18
2.3.1 Linear Model with $p < n$	18
2.3.2 Linear Model with $p > n$	20
2.3.3 Quadratic Model with More Terms than Sample Size	22
2.4 Applications	26
2.5 Discussion	28
Chapter 3 Groupwise Sufficient Dimension Reduction via Envelope Method 31	
3.1 Introduction	31
3.2 Groupwise Dimension Reduction Subspace and Direct Sum Envelope	35
3.2.1 Groupwise Dimension Reduction Subspace	35
3.2.2 Direct Sum Envelope	39
3.3 Estimation via Envelope Method	40
3.3.1 Envelope for Groupwise Dimension Reduction	41
3.3.2 The Objective Function	43
3.3.3 Estimation	45
3.3.4 Numerical Procedures	46
3.3.5 Dimension Estimation	50
3.4 Simulations	51
3.4.1 A General Model	52
3.4.2 Various Correlations Among Predictors	53
3.4.3 A More Challenging Model	54
3.4.4 No Group Structure	56
3.4.5 Partial Dimension Reduction	57
3.4.6 Dimension Estimation	57
3.5 Application	59

3.6	Discussion	60
Chapter 4 Variable Selection with the Kernel Machine Cox Proportional Hazards Model for Optimal Treatment Strategy		
4.1	Introduction	63
4.2	Model and Method	66
4.2.1	Potential Outcome	66
4.2.2	Cox Proportional Hazards Model for Optimal Treatment Strategy	69
4.2.3	The Kernel Machine Method	71
4.2.4	Connection to the Mixed Effects Cox Model	74
4.2.5	Variable Selection with the Kernel Machine Cox Model	76
4.3	Simulations	77
4.4	Discussion	79
References		87

LIST OF TABLES

Table 2.1	Simulation results for a linear Cox model with $p < n$. Results are averaged over 100 replications. Numbers in parentheses are standard errors.	21
Table 2.2	Simulation results for a linear Cox model with $p > n$. Results are averaged over 100 replications. Numbers in parentheses are standard errors.	23
Table 2.3	Simulation results for a quadratic Cox model with more terms than the sample size. Results are averaged over 100 replications. Numbers in parentheses are standard errors.	25
Table 2.4	Genes selected by FOSSA for the lymphoma data and the breast cancer data. Reported are the order of selection, the gene ID, the estimated coefficient by FOSSA, the estimated coefficient by the Cox model with only the selected genes and the corresponding p -values.	28
Table 3.1	Simulation results for Model (3.5). Median and median absolute deviation (in parentheses) of vector correlation coefficient over 100 replications are shown.	53
Table 3.2	Simulation results for Model (3.6) with $n = 800$. Median and median absolute deviation (in parentheses) of vector correlation coefficient over 100 replications are shown.	54
Table 3.3	Simulation results for Model (3.7). Median and median absolute deviation (in parentheses) of vector correlation coefficient over 100 replications are shown.	55
Table 3.4	Simulation results for Model (3.8). Median and median absolute deviation (in parentheses) of vector correlation coefficient over 100 replications are shown.	56
Table 3.5	Simulation results for Model (3.9). Median and median absolute deviation (in parentheses) of vector correlation coefficient over 100 replications are shown.	58
Table 3.6	Proportions of selecting the correct set of dimensions within the top 1, 2 and 3 choices using the BIC-type dimension estimation criterion with gSIR for Models (3.5, 3.6, 3.7) over 100 replications.	58
Table 3.7	Linear regression p values for the temperature-proxy dataset.	60
Table 4.1	Simulation results for Case 1 with 20% censoring. Results are averaged over 100 replications. Numbers in parentheses are standard errors.	81

Table 4.2	Simulation results for Case 2 with 20% censoring. Results are averaged over 100 replications. Numbers in parentheses are standard errors.	82
Table 4.3	Simulation results for Case 3 with 20% censoring. Results are averaged over 100 replications. Numbers in parentheses are standard errors.	83
Table 4.4	Simulation results for Case 1 with 40% censoring. Results are averaged over 100 replications. Numbers in parentheses are standard errors.	84
Table 4.5	Simulation results for Case 2 with 40% censoring. Results are averaged over 100 replications. Numbers in parentheses are standard errors.	84
Table 4.6	Simulation results for Case 3 with 40% censoring. Results are averaged over 100 replications. Numbers in parentheses are standard errors.	85
Table 4.7	Variable selection results averaged over 100 replications.	86

LIST OF FIGURES

Figure 2.1	Kaplan-Meier estimates of survival curves for high and low risk patients for diffuse large-B-cell lymphoma data. Log-rank tests p values are given.	29
Figure 2.2	Kaplan-Meier estimates of survival curves for high and low risk patients for breast cancer data. Log-rank tests p values are given.	29
Figure 3.1	Cross-validation errors for the temperature dataset.	61

Chapter 1

Introduction

As technology advances, high dimensional data analysis has become ubiquitous in modern scientific research. For instance, in a typical microarray dataset, the number of predictors (genes) p is in thousands or more, whereas the number of observations n is much smaller, often only in hundreds or fewer. For such high dimensional data, usual regression and modeling techniques often encounter difficulties, and certain dimension reduction and/or variable selection methods are highly desirable to either transform high dimensional data into lower dimensional space via linear or nonlinear transformations, or to select important variables that are relevant to the response variable. Moreover, on top of the requirement for dimension reduction and/or variable selection methods, other complications are usually involved in high dimensional data analysis. For example, in many applications, the response variable of interest is censored survival data, e.g., time to death or cancer recurrence in biomedical research, credit default time in finance, and mechanical failure time in engineering. In such situations, special handling of the censoring issue is required in data analysis. Another example is the presence of prior group information in the predictors. For instance, in a global surface temperature reconstruction

study (Mann et al., 2008), 1,209 proxy series are recorded to characterize temperature changes over time, and these proxies come from several distinct groups, such as tree rings, ice cores, cave deposits, lake sediments and historical documentation series. It is usually desirable to incorporate such group information into dimension reduction to obtain more interpretable results. A third example is to apply variable selection method in making optimal treatment decisions in clinical studies. Unlike standard regression problems, the important variables to be selected here is the ones that involved with treatment effects, and therefore the problem becomes more complicated. Variable selection and dimension reduction for high dimensional data with such complications are especially challenging, and are the focuses of our studies.

The rest of this dissertation is divided into three chapters, and in each chapter we focus on a different complication that arises in variable selection and/or dimension reduction. In Chapter 2, we address model estimation and variable selection problems simultaneously for censored survival data with high dimensional predictors, including the case when $p \gg n$. The Cox proportional hazards model (Cox, 1972) is applied in this study. We propose the FOrward Stagewise Shrinkage and Addition (FOSSA) method, which combines the strategies of forward stagewise boosting and shrinkage penalties, and carries out an additive stagewise modeling while applying shrinkage penalties at each iteration. The FOSSA method works naturally for high dimensional problems, and performs simultaneous model estimation and variable selection. In addition, the proposed method is flexible such that it can be easily extended to nonlinear models. Simulation studies and real data analyses are performed to demonstrate the competitive performance of the proposed method.

In Chapter 3, we consider situations where prior group information in the predictors are available for dimension reduction methods in the sufficient dimension reduction (SDR)

framework (Cook, 1998). In many cases like the temperature dataset example in Mann et al. (2008), the predictors naturally fall into different domains or groups, and accounting for these prior group information during dimension reduction steps would provide more interpretable results and lead to more accurate estimates of the underlying directions. We propose a groupwise sufficient dimension reduction method in Chapter 3 to accomplish this goal. The proposed method covers any classical dimension reduction estimators via a direct sum envelope that is constrained by the desired group structure. As such, the prior group information is imposed onto the classical dimension reduction estimator, and leads to simultaneous estimation of groupwise SDR directions. Simulation studies and real data analysis have shown that the proposed method performs well under various situations.

In Chapter 4, we focus on estimating the optimal treatment strategy with variable selection in the context of survival data. In clinical studies, optimal treatment strategies are a set of rules for making personalized effective treatment decisions that are determined based on each individual's various characteristics, such as genetic, environmental and behavioral characteristics, such that the long-term clinical outcome is optimized. When the clinical outcome is censored survival data, the optimal treatment strategy is usually defined as the decision rule that maximizes the expected survival time, or the one that maximizes the survival probability at a given time point. In addition, with the large amount of characteristic and clinical information available, it is important to apply variable selection method during estimation to achieve interpretable and efficient results. We propose to apply shrinkage type variable selection method with the kernel machine Cox proportional hazards model for this problem. The proposed method leads to estimates of the optimal treatment strategy that meet the requirements of both definitions given above, and is flexible by allowing nonparametric specification of the baseline covariate

effects. Moreover, by incorporating shrinkage penalties in the estimation, the noises in the estimated decision rule are greatly reduced. Numerical studies suggest that the proposed method is useful in making correct treatment decisions under various underlying true models.

Chapter 2

Forward Stagewise Shrinkage and Addition for High Dimensional Censored Regression

2.1 Introduction

For high dimensional data analysis, it is often true that only a subset of variables are relevant to the outcome, and as such variable selection becomes a vital ingredient of the data analysis. In many applications, the outcome variable is subject to censoring, e.g., time to death or cancer recurrence in biomedical research, credit default time in finance, and mechanical failure time in engineering. Cox proportional hazards model (Cox, 1972) has been the most popular tool for analyzing censored responses. However, Cox model estimation and variable selection are challenging, especially when the number of covariates (genes) p is greater than the number of observations n , and this is to be the focus of this chapter.

During the past decade, many research efforts have been devoted to the area of variable selection, and a variety of selection methods have been proposed. An outstanding class is the shrinkage method, which minimizes a penalized objective function

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\{L(\boldsymbol{\beta}) + J(\boldsymbol{\beta})\},$$

where $L(\boldsymbol{\beta})$ represents the loss function, such as the residual sum of squares in linear regression and negative log likelihood function in generalized linear models, $J(\boldsymbol{\beta})$ is a penalty term and often involves some tuning parameters. For example, in least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), $J(\boldsymbol{\beta}) = \lambda \sum_{j=1}^p |\beta_j|$; in elastic net (Zou and Hastie, 2005), $J(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2$; and in adaptive LASSO (Zou, 2006), $J(\boldsymbol{\beta}) = \lambda \sum_{j=1}^p |\beta_j|/|\tilde{\beta}_j|$, where $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \dots, \tilde{\beta}_p)^\top$ is a consistent estimator of $\boldsymbol{\beta}$. Other examples include the smoothly clipped absolute deviation (SCAD) method (Fan and Li, 2001), nonnegative garrote (Yuan and Lin, 2007), among many others. Many of those methods have been extended to the Cox model, e.g., Tibshirani (1997), Fan and Li (2002), and Zhang and Lu (2007). In general, this class of solutions can be formulated in terms of a loss function plus a regularization term, and its minimization leads to a sparse estimate of the regression parameter, which in effect achieves simultaneously variable selection and parameter estimation. They have been shown to perform competently in various settings especially when p is small to moderate.

However, when facing high dimensional problems, for instance, microarray data analysis where p is huge, the inherent computational complexity of the aforementioned methods may cause algorithmic instability and yield estimators with large variance. In addition, when $p \gg n$, methods such as the adaptive LASSO become either difficult or infeasible to implement. One solution to address high dimensional problems is variable screening,

which filters large number of predictors via marginal measure between each individual predictor and the response. For linear models, Fan and Lv (2008) proposed sure independence screening (SIS) using marginal correlation, and showed that it would retain all active predictors with a large probability. Another popular and also classical solution for $p \gg n$ is forward stepwise regression (FR) that selects one variable at a time. This strategy avoids computational complexity of dealing with all p predictors simultaneously, and is shown by Wang (2009) to be capable of selecting all active predictors consistently in the linear model setup. Numerical studies have found that both SIS and FR perform competitively compared to other variable selection solutions when $p \gg n$. However, both focus on variable selection only, and model estimation is usually carried out separately after the selection.

Another method that works in a similar fashion as FR is forward stagewise regression, which also builds a model iteratively, and in some of its implementations (Li and Luan, 2005; Lu and Li, 2008), it involves only one variable at each iteration. It differs from FR in that, once a term is added into the model, its coefficient remains unadjusted in all subsequent iterations.

A successful application and generalization of forward stagewise regression is boosting. It repeatedly applies a fitting method, called the base learner, to the reweighted data and updates the estimator by adding a weighted version of the newly fitted base learner at each iteration. The final boosting estimator is constructed by taking a weighted sum of the series of the fitted base learners. Boosting first originated in the machine learning community known as a classification technique called AdaBoost (Schapire, 1990; Freund, 1995; Freund and Schapire, 1997). For two class classification problems with $Y \in \{-1, 1\}$ be a binary outcome, the algorithm of AdaBoost is given as follows (Hastie et al., 2005):

Step 1. Initialization. Set weights $w_i = 1/N, i = 1, \dots, N$. Set $F_0(\mathbf{X}) = 0$.

Step 2. Repeat for $m = 1, \dots, M$

(a) Fit the classifier $f_m(\mathbf{X}) \in \{-1, 1\}$ using weighted training data with weights w_i .

(b) Compute error rate $err_m = \frac{\sum_i w_i I(Y_i \neq f_m(\mathbf{X}_i))}{\sum_i w_i}$, and $c_m = \log\{(1 - err_m)/err_m\}$.

(c) Update $F_m(\mathbf{X}) = F_{m-1}(\mathbf{X}) + c_m f_m(\mathbf{X})$.

(d) Update the weights w_i as $w_i \exp[c_m I\{Y_i \neq f_m(\mathbf{X}_i)\}], i = 1, \dots, N$.

Step 3. Output the final AdaBoost classifier as $\text{sign}\{F_M(\mathbf{X})\}$.

Later Friedman et al. (2000) showed that the above Adaboost algorithm is in fact equivalent to fitting a forward stagewise additive model by minimizing the exponential loss function $E[\exp\{-YF(\mathbf{X})\}]$. Friedman (2001) further proposed a general gradient descent boosting algorithm that can accommodate a variety of loss functions. Suppose we want to find a function estimate $F(\mathbf{X})$ by minimizing the expectation of a particular loss function $\Psi(Y, F)$, and let $f(\mathbf{X}, a)$ be the base learner, then the algorithm for gradient descent boosting is given by (Friedman, 2001):

Step 1. Initialization. Set $F_0(\mathbf{X}) = \text{argmin}_\rho \sum_{i=1}^N \Psi(Y_i, \rho)$.

Step 2. Repeat for $m = 1, \dots, M$

(a) Compute the working response as the negative gradient

$$\tilde{Y}_i = -\frac{\partial \Psi(Y_i, F(\mathbf{X}_i))}{\partial F(\mathbf{X}_i)} \Big|_{F(\mathbf{X}_i) = F_{m-1}(\mathbf{X}_i)}$$

(b) Regress $f(\mathbf{X}_i, a)$ on \tilde{Y}_i . Find $a_m = \text{argmin}_{a, \beta} \sum_{i=1}^N \{\tilde{Y}_i - \beta f(\mathbf{X}_i, a)\}^2$.

(c) Line search to choose the gradient descent step size as

$$\rho_m = \text{argmin}_\rho \sum_{i=1}^N \Psi\{Y_i, F_{m-1}(\mathbf{X}_i) + \rho f(\mathbf{X}_i, a_m)\}$$

(d) Update $F_m(\mathbf{X}) = F_{m-1}(\mathbf{X}) + \rho_m f(\mathbf{X}_i, a_m)$

Step 3. Output the final estimate $F_M(\mathbf{X})$.

As the gradient boosting method allows various loss functions, a number of extensions followed (Bühlmann and Yu, 2003; Bühlmann and Hothorn, 2007; Ridgeway, 1999). Because the base learner used in boosting can be both linear and nonlinear, this method can fit data with a more complicated structure than a linear form. Besides, the base learner often involves one variable at a time, and thus it works for large p regressions. Applications of boosting to censored data were also developed, including Li and Luan (2005) for Cox proportional hazards model and Lu and Li (2008) for transformation models. Despite its flexible nature, its generality to cope with various types of regressions, and its competitive empirical performance, boosting suffers from some noticeable drawbacks. First, boosting does not perform explicit variable selection. A variable is selected if its coefficient is nonzero when the algorithm stops. Relative contributions of individual variables are measured by a heuristic importance measure (Friedman, 2001), while there is no associated inference available to separate the active predictors from the inactive ones. Second, a boosting algorithm often takes many iterations and thus a long time to complete. This is mainly due to the fact that a small learning rate is added to the model update at each iteration to achieve proportional shrinkage (Friedman, 2001). As a consequence, adding an active covariate into the model may take tens of iterations or more to complete. Other examples of high dimensional survival analysis include Ishwaran et al. (2010), Witten and Tibshirani (2010) and van Wieringen et al. (2009).

In this chapter, we aim to address simultaneous model estimation and variable selection in Cox proportional hazards models with high dimensional predictors. We couple the strategies of shrinkage estimation and forward stagewise boosting, and propose a Forward Stagewise Shrinkage and Addition (FOSSA) method, which carries out an additive

stagewise modeling while introducing shrinkage at each iteration. Our contributions are two fold. First, the proposed method works naturally for high dimensional regressions. When facing high dimensional predictors, existing solutions often require a pre-screening as the first step, followed by a refined variable selection and parameter estimation as the second step. By contrast, our method does not require any pre-screening, and in effect combines the two tasks in one stroke. Our intensive simulations show that our method performs competitively compared to the existing methods for both $n > p$ and $n \ll p$ scenarios. Second, our solution inherits the flexibility of the boosting method. We will focus on linear Cox models in this chapter, but extensions to nonlinear Cox models are straightforward and will be briefly discussed at the end. On the other hand, our method also extends the existing boosting method in that, it now performs explicit variable selection, and it greatly reduces the number of iterations required. As we will see later in simulations, the new method often converges in tens of steps, compared to hundreds of iterations or more demanded by the usual boosting.

The rest of the chapter is organized as follows. In Section 2.2, we first briefly review the Cox proportional hazards model, adaptive LASSO, and forward stagewise boosting. We then present our new method in detail. In Section 2.3, we conduct an intensive simulation study to investigate the empirical performance of the proposed method, and to compare with existing variable selection and boosting solutions, In Section 2.4, FOSSA is applied to the analysis of two microarray data examples for further illustration. We conclude the chapter in Section 2.5 with a discussion on future extensions.

2.2 Model and Method

2.2.1 Cox Proportional Hazards Model with Adaptive LASSO

Our discussion hereinafter will assume the context of survival data analysis, while the methodology applies to other censored regressions as well. Considering n random subjects, let T_i be the failure time, C_i be the censoring time, and $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^\top$ be the vector of covariates of the i th subject, $i = 1, \dots, n$. Define the observed event time $\tilde{T}_i = \min(T_i, C_i)$ and the censoring indicator $\delta_i = I(T_i \leq C_i)$. Then the observed data consist of $\{(\tilde{T}_i, \delta_i, \mathbf{X}_i), i = 1, \dots, n\}$. Furthermore, we assume conditional independent censoring, i.e., T_i and C_i are independent conditioning on \mathbf{X}_i , throughout this chapter.

The Cox proportional hazards model assumes that the hazard function $\lambda(t|\mathbf{X}_i)$ of subject i with covariate \mathbf{X}_i is given as:

$$\lambda(t|\mathbf{X}_i) = \lambda_0(t) \exp(\boldsymbol{\beta}^\top \mathbf{X}_i), \text{ for } i = 1, \dots, n, \quad (2.1)$$

where $\lambda_0(t)$ is the unspecified baseline hazard function and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ is the vector of regression coefficients. Assuming there are no tied observations, then the log partial likelihood function (Cox, 1975) is given by:

$$\ell_n(\boldsymbol{\beta}) = \sum_{i=1}^n \delta_i \left[\boldsymbol{\beta}^\top \mathbf{X}_i - \log \left\{ \sum_{j=1}^n \exp(\boldsymbol{\beta}^\top \mathbf{X}_j) I(\tilde{T}_j \geq \tilde{T}_i) \right\} \right]. \quad (2.2)$$

An estimate of $\boldsymbol{\beta}$ is obtained by minimizing $-\ell_n(\boldsymbol{\beta})$ over $\boldsymbol{\beta}$.

To perform variable selection under this model, Zhang and Lu (2007) proposed the adaptive LASSO method, which minimizes the weighted L_1 penalized negative log partial

likelihood,

$$-\frac{1}{n}\ell_n(\boldsymbol{\beta}) + \lambda \sum_{j=1}^p |\beta_j|/|\tilde{\beta}_j|,$$

where $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \dots, \tilde{\beta}_p)^\top$ is the maximizer of (2.2). It is shown that, asymptotically, this adaptive LASSO estimator enjoys the oracle properties, and numerically, the method performs competitively when $p < n$. However, when $p > n$, $\tilde{\boldsymbol{\beta}}$ is not directly available, and thus the adaptive LASSO method cannot be applied to the Cox model with $p > n$.

2.2.2 A Modified Boosting Algorithm

Before we introduce our FOSSA method, we first present a version of boosting algorithm for the Cox model that will help the establishment of FOSSA. It is similar in spirit as the proposal of Li and Luan (2005) though not identical. The basic idea is to model the survival time in the form of $\lambda(t|\mathbf{X}) = \lambda_0(t) \exp\{F(\mathbf{X})\}$, where the function $F(\mathbf{X})$ is updated by iteratively adding new terms. Since the terms added to $F(\mathbf{X})$ in precedent steps remain unchanged in future iterations, they are to be treated as an offset term in subsequent model fitting. We thus denote the log partial likelihood function with an offset term $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^\top$ as

$$\ell_n(\boldsymbol{\beta}, \boldsymbol{\mu}) = \sum_{i=1}^n \delta_i \left[\mu_i + \boldsymbol{\beta}^\top \mathbf{X}_i - \log \left\{ \sum_{j=1}^n \exp(\mu_j + \boldsymbol{\beta}^\top \mathbf{X}_j) I(\tilde{T}_j \geq \tilde{T}_i) \right\} \right].$$

Meanwhile, we use $\ell_n(\beta_g, \boldsymbol{\mu})$ to denote the log partial likelihood function when only the g -th covariate X_g is fitted in the Cox model. That is,

$$\ell_n(\beta_g, \boldsymbol{\mu}) = \sum_{i=1}^n \delta_i \left[\mu_i + \beta_g X_{ig} - \log \left\{ \sum_{j=1}^n \exp(\mu_j + \beta_g X_{jg}) I(\tilde{T}_j \geq \tilde{T}_i) \right\} \right].$$

We next present the boosting algorithm.

Step 1. Set initial values: iteration $k = 0$, the offset $\boldsymbol{\mu}^{[0]} = (\mu_1^{[0]}, \dots, \mu_n^{[0]})^\top = \mathbf{0}$, and the fitted function $F^{[0]}(\mathbf{X}) = 0$.

Step 2. Repeat for $k = 0, \dots, K$.

(a) Obtain the coefficients:

$$\tilde{\beta}_g = \underset{\beta_g}{\operatorname{argmin}} \left\{ -\ell_n(\beta_g, \boldsymbol{\mu}^{[k]}) \right\}, \quad g = 1, \dots, p. \quad (2.3)$$

That is, we fit p univariate Cox models with each covariate X_1, \dots, X_p as the sole predictor plus an offset $\boldsymbol{\mu}^{[k]}$.

(b) Select the covariate with minimum negative log partial likelihood to add in the model, i.e., $g^* = \underset{g=1, \dots, p}{\operatorname{argmin}} \left\{ -\ell_n(\tilde{\beta}_g, \boldsymbol{\mu}^{[k]}) \right\}$.

(c) Update the fitted function $F^{[k+1]}(\mathbf{X}) = F^{[k]}(\mathbf{X}) + \nu \tilde{\beta}_{g^*} X_{g^*}$, where ν is a small learning rate (e.g. 0.05 or 0.1). Also update the offset

$$\boldsymbol{\mu}^{[k+1]} = (F^{[k+1]}(\mathbf{X}_1), \dots, F^{[k+1]}(\mathbf{X}_n))^\top.$$

Step 3. Output the final estimate as $F^{[K]}(\mathbf{X})$.

We first note that this version of the boosting algorithm is slightly different from the gradient descent boosting of Friedman (2001) and Li and Luan (2005) for Cox models. The difference is at Step 2(a), where this algorithm estimates β_g via maximum partial likelihood, whereas in the usual gradient descent boosting, β_g is estimated by fitting a regression model with the negative gradient as the response. As such, the two procedures may end up selecting different variables to add into the model. But if the same variable is selected, the two procedures would yield the same fitted function $F(\mathbf{X})$, because the

gradient descent boosting (Li and Luan, 2005) has a line search step, which essentially fits a Cox model with the selected variable. On the other hand, if there is no response censoring and the loss function is chosen to be the usual squared loss, the two algorithms are equivalent, because both methods repeatedly fit the least squares residuals. We make the following remarks about this algorithm, which also apply to gradient boosting algorithms using component-wise base learners. First, the number of boosting iterations K is usually tuned through cross-validation methods. Second, we note that, at each iteration, univariate regressions are carried out, and variable is added into the model one-at-a-time. This strategy is expected to be helpful for a very large p . Third, a small learning rate ν is imposed to regularize the contribution of the newly selected term through proportional shrinkage. This is because once a coefficient is selected into the model, its coefficient remains unadjusted afterwards, and so adding a small learning rate would help reduce the effect of “incorrectly” estimated parameters to the final model. Finally, no explicit variable selection is performed. Usually it would take many steps to add a truly relevant variable into the model, whereas many irrelevant variables are to be included in the model with a tiny coefficient estimate.

2.2.3 Forward Stagewise Shrinkage and Addition

To capitalize on the advantages whereas to overcome the drawbacks of the adaptive LASSO shrinkage estimation and the boosting method, we propose to couple the two strategies for the purpose of variable selection in high dimensional Cox proportional hazards models. The basic idea is to introduce shrinkage estimation at each iteration of the boosting algorithm, meanwhile dropping the learning rate. As such it in effect replaces the proportional shrinkage dictated by the common learning rate ν with an

adaptive shrinkage at each iteration. We call the resulting algorithm the forward stagewise shrinkage and addition method (FOSSA). Specifically,

Step 1. Set initial values: iteration $k = 0$, the offset $\boldsymbol{\mu}^{[0]} = (\mu_1^{[0]}, \dots, \mu_n^{[0]})^\top = \mathbf{0}$, and the fitted function $F^{[0]}(\mathbf{X}) = 0$.

Step 2. Obtain the coefficients:

$$\hat{\beta}_g = \underset{\beta_g}{\operatorname{argmin}} \left\{ -\frac{1}{n} \ell_n(\beta_g, \boldsymbol{\mu}^{[k]}) + \frac{\lambda}{|\tilde{\beta}_g|} |\beta_g| \right\}, \quad g = 1, \dots, p, \quad (2.4)$$

where $\tilde{\beta}_g$ is the unpenalized estimate obtained in (2.3), and λ is a shrinkage parameter. So in effect, we fit p adaptive LASSO type Cox models with each covariate X_1, \dots, X_p as the sole predictor plus an offset $\boldsymbol{\mu}^{[k]}$.

Step 3. Select the covariate with minimum negative log partial likelihood to add in the model, i.e., $g^* = \underset{g=1, \dots, p}{\operatorname{argmin}} \{-\ell_n(\hat{\beta}_g, \boldsymbol{\mu}^{[k]})\}$.

Step 4. Update the fitted function $F^{[k+1]}(\mathbf{X}) = F^{[k]}(\mathbf{X}) + \hat{\beta}_{g^*} X_{g^*}$. Also update the offset $\boldsymbol{\mu}^{[k+1]} = (F^{[k+1]}(\mathbf{X}_1), \dots, F^{[k+1]}(\mathbf{X}_n))^\top$.

Step 5. Increment k by 1. Go back to Step 2 until convergence. Here we declare the algorithm has converged if the difference of two successive log partial likelihood values is below a preselected threshold (e.g., 10^{-4} or 10^{-6}).

Comparing the FOSSA algorithm with the forward stagewise boosting algorithm, we note that the differences are in Step 2, where an adaptive LASSO penalty is introduced in estimating β_g , and in Step 4, where the learning rate ν is no longer needed. The consequences of those changes are the following. First, variable selection is now carried out explicitly; the truly inactive predictor would usually not enter the model due to the

large penalty, and the truly active predictor could enter the model faster due to the drop of the small learning rate. Second, the new algorithm converges very fast, and thanks to the regularization in coefficient estimation, also achieves a more accurate estimation compared to the usual boosting algorithm. This will later be verified in our simulations.

Implementation of FOSSA requires solving the minimization problem in (2.4) and tuning of the shrinkage parameter λ . For the first task, the optimization is carried out by adopting the idea of Wang and Leng (2007). More specifically, we aim to minimize the objective function

$$-\frac{1}{n}\ell_n(\beta_g, \boldsymbol{\mu}^{[k]}) + \frac{\lambda}{|\tilde{\beta}_g|}|\beta_g|, \quad (2.5)$$

where $\tilde{\beta}_g$ is the maximized partial likelihood estimate obtained in (2.3). Applying the Taylor series expansion of the negative log partial likelihood at $\tilde{\beta}_g$, we obtain

$$-\frac{1}{n}\ell_n(\beta_g, \boldsymbol{\mu}^{[k]}) \approx -\frac{1}{n}\ell_n(\tilde{\beta}_g, \boldsymbol{\mu}^{[k]}) + \frac{1}{n}G(\tilde{\beta}_g)(\beta_g - \tilde{\beta}_g) + \frac{1}{2n}H(\tilde{\beta}_g)(\beta_g - \tilde{\beta}_g)^2,$$

where $G(\tilde{\beta}_g)$ and $H(\tilde{\beta}_g)$ denote the first and second derivative of $-\ell_n(\beta_g, \boldsymbol{\mu}^{[k]})$ with respect to β_g that is evaluated at $\tilde{\beta}_g$. Furthermore, we note that $G(\tilde{\beta}_g) = 0$. Then, by ignoring the constant, we can rewrite the objective function in (2.5) as

$$\frac{1}{2n}H(\tilde{\beta}_g)(\beta_g - \tilde{\beta}_g)^2 + \frac{\lambda}{|\tilde{\beta}_g|}|\beta_g|.$$

Its minimizer is then given by:

$$\hat{\beta}_g = \text{sign}(\tilde{\beta}_g) \left(|\tilde{\beta}_g| - \frac{n\lambda}{H(\tilde{\beta}_g)|\tilde{\beta}_g|} \right)_+,$$

which can be equivalently written as

$$\hat{\beta}_g = \begin{cases} \tilde{\beta}_g - n\lambda\{H(\tilde{\beta}_g)\tilde{\beta}_g\}^{-1} & \text{if } \tilde{\beta}_g^2 > n\lambda H(\tilde{\beta}_g)^{-1} \\ 0 & \text{otherwise} \end{cases}$$

For the task of tuning λ , we employ a Bayesian information criterion (BIC)

$$-2\hat{\ell}_n + \log(n) d_\lambda, \tag{2.6}$$

where $\hat{\ell}_n$ denotes the final log partial likelihood function after the algorithm converges, and d_λ is the number of nonzero covariates in the final model. We then search over a grid values of λ and choose the one that minimizes (2.6). In this algorithm, we have chosen a common shrinkage parameter λ for all covariates across all iterations. We also note that, despite this choice of common λ , the amount of regularization is adaptively dependent upon the magnitude of the individual covariate $|\tilde{\beta}_g|$.

Alternatively, we may also specify a different tuning parameter $\lambda_g^{[k]}$ in (2.4) for each covariate g at each iteration k . We can continue to use BIC to tune this parameter, except that $\hat{\ell}_n$ in (2.6) is replaced with the log partial likelihood function at the current state, and d_λ only takes two values 0 or 1. Consequently, there are only two choices of λ , either $(\tilde{\beta}_g)^2 H(\tilde{\beta}_g)/n$ or 0, which yields either $\hat{\beta}_g = 0$ or $\hat{\beta}_g = \tilde{\beta}_g$, respectively. This way of tuning is computationally simpler. We have compared the two tuning strategies, and found that (results not reported here), the common λ with a grid search achieves a comparable performance as the varying λ when p is small, and shows an edge when p is large. For this reason, we will adopt the common λ strategy in the rest of the chapter.

2.3 Simulations

We conduct an intensive simulation study to evaluate the empirical performance of FOSSA. We consider three scenarios: a usual setup of a linear Cox model with $p < n$, a linear Cox model with $p > n$, and a full quadratic Cox model where $p < n$ but the total number of terms in the model, including the linear terms, the quadratic terms and all the two-way interaction terms, far exceeds the sample size. We also compare FOSSA with existing solutions, including LASSO, adaptive LASSO (aLASSO), forward stepwise selection (FR), and gradient descent boosting using component-wise univariate linear base learners. As a benchmark, we include the oracle estimator in comparison as well.

2.3.1 Linear Model with $p < n$

We first generate the covariates $\mathbf{X}_i, i = 1, \dots, n$, according to a multivariate normal distribution with mean zero, variance one, and an order one autoregressive correlation structure, where $\text{corr}(X_{ij}, X_{ik}) = 0.5^{|j-k|}, j, k = 1, \dots, p$. The failure time T_i 's are then generated following the Cox proportional hazards model given in (2.1) with $\lambda_0(t) = 1$ and $\beta = (1, 0.8, 0, \dots, 0, 0.6)^\top$. In other words, only the first two and the last predictors are active. The censoring time C_i 's are generated independently from a uniform $(0, C_0)$ distribution, where C_0 controls the censoring proportion. We have examined two censoring proportions in our study, 20% and 40%. Since the two sets of results show very similar qualitative patterns, we only report the case with 40% censoring. We vary the sample size $n = 100$ and $n = 150$, and the number of predictors $p = 20$ and $p = 30$. This is a typical setup in the Cox model variable selection with a small p . A total of 100 data replications are performed. We have noted in our simulations that the FR algorithm sometimes fails to converge. For that reason, we limit the maximum number of iterations in FR to be

$\min(\lceil n/\log(n) \rceil, p)$, and the final model is chosen based on BIC. For boosting, we use the `mboost` package in R (Bühlmann and Hothorn, 2007) and choose the learning rate $\nu = 0.1$. The number of boosting iterations is tuned via 5 fold cross-validation.

We evaluate and compare methods by three categories of criteria. The first is the average mean squared error (MSE), $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{\Sigma} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, where $\boldsymbol{\Sigma}$ is taken to be the covariance of the covariates \mathbf{X} in this setup. This criterion measures the estimation accuracy of the model parameter $\boldsymbol{\beta}$. The second category examines the performance in terms of variable selection accuracy. Criteria in this category include the average size of the selected models (Size), the frequency of selecting all truly active predictors (Cover) out of 100 data replications, the frequency of selecting the exact model (Exact), the percentage of correct zeros being identified (Corr0), and the percentage of incorrect zeros being identified (Incorr0). The third category compares the boosting algorithm and the new FOSSA algorithm in terms of the average iterations (Iter) required.

Table 2.1 reports the results. In terms of both parameter estimation accuracy and variable selection accuracy, the three methods, adaptive LASSO, FR and FOSSA perform similarly. LASSO is less accurate in estimating $\boldsymbol{\beta}$, as reflected by a larger MSE, since its estimation is biased. The usual boosting estimator always selects more variables than necessary, as indicated by a much larger model size than the truth. Comparing the number of iterations, FOSSA requires far fewer steps than boosting. Overall, FOSSA achieves a comparable performance as adaptive LASSO and FR, and a superior performance than LASSO and boosting, in this typical $p < n$ setup.

2.3.2 Linear Model with $p > n$

In the second example, the data are generated in the same way as before, except that $\beta = (1, 1, 0, \dots, 0, 1)^\top$, with $p = 100, 500$, and 1000 . The censoring proportions are chosen as 20% and 40%, and the sample size is $n = 100$. So now we have a linear Cox model with $p > n$. In this case, adaptive LASSO does not work, and thus is not included in the comparison. As a result, we compare the performance of LASSO, FR, boosting and FOSSA. We employ the `glmnet` function in R (Friedman et al., 2010) for the implementation of LASSO in this $p > n$ scenario.

Table 2.2 reports the results. Among the four methods, we see that FOSSA achieves the smallest MSE, while FR yields a substantially large MSE, indicating a very inaccurate parameter estimation. In terms of variable selection, FOSSA constantly achieves an average model size that is closest to the true value, which is three in our example, whereas LASSO, FR and boosting all result in a much larger model. FOSSA also performs much better than LASSO, FR and boosting in selecting the exact model, which is rarely picked by the latter three methods. Moreover, the FOSSA algorithm converges quickly even with a very large p , and requires much less iteration steps compared to boosting. In summary, for the $p > n$ scenario, FOSSA performs noticeably better than its competitors LASSO, FR and boosting.

Table 2.1: Simulation results for a linear Cox model with $p < n$. Results are averaged over 100 replications. Numbers in parentheses are standard errors.

p	n	Method	MSE	Size	Cover	Exact	Corr0	Incorr0	Iter
$p = 20$	$n = 100$	Oracle	0.095 (0.011)						
		LASSO	0.514 (0.026)	4.28 (0.17)	0.93	0.34	0.921	0.023	
		aLASSO	0.259 (0.019)	3.65 (0.10)	0.91	0.49	0.956	0.033	
		FR	0.280 (0.030)	3.68 (0.10)	0.94	0.46	0.956	0.023	
		Boosting	0.265 (0.015)	7.53 (0.26)	0.98	0.03	0.732	0.007	171.1
		FOSSA	0.210 (0.015)	3.63 (0.10)	0.93	0.48	0.958	0.027	5.0
	$n = 150$	Oracle	0.062 (0.006)						
		LASSO	0.296 (0.017)	4.04 (0.13)	0.98	0.43	0.938	0.007	
		aLASSO	0.119 (0.010)	3.49 (0.09)	0.98	0.63	0.970	0.007	
		FR	0.103 (0.010)	3.36 (0.06)	0.99	0.71	0.978	0.003	
		Boosting	0.154 (0.009)	7.48 (0.25)	1.00	0.02	0.736	0.000	173.1
		FOSSA	0.120 (0.008)	3.52 (0.09)	0.99	0.64	0.969	0.003	4.9
$p = 30$	$n=100$	Oracle	0.083 (0.008)						
		LASSO	0.938 (0.046)	3.50 (0.15)	0.74	0.30	0.970	0.107	
		aLASSO	0.470 (0.033)	3.40 (0.12)	0.82	0.47	0.977	0.073	
		FR	0.389 (0.040)	4.48 (0.16)	0.99	0.33	0.945	0.003	
		Boosting	0.286 (0.016)	9.47 (0.35)	1.00	0.02	0.760	0.000	178.3
		FOSSA	0.231 (0.015)	4.17 (0.16)	0.97	0.45	0.956	0.010	5.5
	$n=150$	Oracle	0.057 (0.006)						
		LASSO	0.409 (0.021)	4.17 (0.12)	0.98	0.31	0.956	0.007	
		aLASSO	0.148 (0.010)	3.51 (0.09)	0.98	0.60	0.980	0.007	
		FR	0.151 (0.015)	3.76 (0.10)	1.00	0.50	0.972	0.000	
		Boosting	0.166 (0.009)	8.89 (0.37)	1.00	0.01	0.782	0.000	182.6
		FOSSA	0.113 (0.008)	3.68 (0.10)	0.99	0.54	0.974	0.003	5.0

2.3.3 Quadratic Model with More Terms than Sample Size

In reality, the true association between the survival time and the covariates is often likely to be more complicated than the linear Cox model depicted in (2.1). One option to incorporate such a complex association is to include higher order polynomials of the covariates, e.g., the quadratic and interaction terms. In principle, the selection methods that work for the linear model work for the polynomial model too, by treating all those higher order terms as additional covariates. However, practically, the effective number of predictors in a polynomial model grows rapidly with the number p of the original predictors, and based upon the observations in Section 2.3.2, the existing solutions are expected to suffer from a deteriorating accuracy in both model estimation and variable selection. In this section, we consider a simulation example of this type. The original covariates \mathbf{X}_i 's are generated in the same way as in Section 2.3.1. Next the failure time T_i 's are generated based on a quadratic Cox model, where $\lambda(t|\mathbf{X}) = \exp(X_1 - X_2 + X_2^2 - 0.8X_1X_5 + 0.8X_6X_9)$. The censoring time C_i 's are generated independently from a uniform distribution with 20% censoring proportion for this more challenging scenario. The sample size is fixed at $n = 100$, and p takes value of 20 and 30. In addition to the original covariates, all their quadratic and two-way interaction terms are included as the candidate variables, resulting in an effective number \tilde{p} of covariates equal to $p + p(p + 1)/2 = 230$ and 495, respectively. Since \tilde{p} is much larger than n , we again compare our FOSSA solution with LASSO, FR and boosting.

Table 2.3 reports the results. Here, to simplify the computation of MSE, we take Σ equal to an identity matrix in MSE. Boosting and FOSSA achieve smallest MSEs in this scenario, but the former comes with the price of a much larger than necessary model size and a large iteration number. LASSO achieve slightly higher MSE, and selects much

Table 2.2: Simulation results for a linear Cox model with $p > n$. Results are averaged over 100 replications. Numbers in parentheses are standard errors.

Censoring Rate=20 %								
p	Method	MSE	Size	Cover	Exact	Corr0	Incorr0	Iter
100	Oracle	0.083 (0.010)						
	LASSO	0.493 (0.021)	10.96 (0.45)	1.00	0.00	0.918	0.000	
	FR	2.462 (0.293)	9.45 (0.48)	1.00	0.04	0.934	0.000	
	Boosting	0.408 (0.019)	14.40 (0.54)	1.00	0.00	0.882	0.000	157.9
	FOSSA	0.334 (0.017)	4.42 (0.18)	1.00	0.42	0.985	0.000	6.2
500	Oracle	0.071 (0.006)						
	LASSO	0.770 (0.028)	14.00 (0.51)	1.00	0.00	0.978	0.000	
	FR	30.722 (1.300)	20.96 (0.03)	1.00	0.00	0.964	0.000	
	Boosting	0.619 (0.026)	21.49 (0.96)	1.00	0.00	0.963	0.000	154.5
	FOSSA	0.487 (0.021)	5.20 (0.26)	1.00	0.35	0.996	0.000	6.8
1000	Oracle	0.102 (0.014)						
	LASSO	0.794 (0.031)	16.15 (0.68)	1.00	0.00	0.987	0.000	
	FR	52.839 (2.861)	21.00 (0.00)	1.00	0.00	0.982	0.000	
	Boosting	0.626 (0.027)	25.39 (1.07)	1.00	0.00	0.978	0.000	162.3
	FOSSA	0.469 (0.019)	6.44 (0.38)	1.00	0.27	0.997	0.000	8.1
Censoring Rate=40 %								
p	Method	MSE	Size	Cover	Exact	Corr0	Incorr0	Iter
100	Oracle	0.101 (0.012)						
	LASSO	0.594 (0.028)	10.60 (0.42)	1.00	0.01	0.922	0.000	
	FR	5.649 (0.838)	10.53 (0.54)	1.00	0.00	0.922	0.000	
	Boosting	0.520 (0.022)	13.41 (0.55)	1.00	0.00	0.893	0.000	185.8
	FOSSA	0.383 (0.019)	4.52 (0.20)	1.00	0.37	0.984	0.000	6.0
500	Oracle	0.105 (0.010)						
	LASSO	0.923 (0.034)	13.37 (0.55)	1.00	0.01	0.979	0.000	
	FR	89.465 (6.957)	21.00 (0.00)	0.99	0.00	0.964	0.003	
	Boosting	0.807 (0.034)	18.29 (0.85)	1.00	0.00	0.969	0.000	150.5
	FOSSA	0.606 (0.028)	4.96 (0.25)	1.00	0.37	0.996	0.000	6.2
1000	Oracle	0.134 (0.018)						
	LASSO	0.942 (0.036)	14.77 (0.65)	0.99	0.00	0.988	0.003	
	FR	644.778 (352.016)	21.00 (0.00)	0.99	0.00	0.982	0.003	
	Boosting	0.758 (0.030)	23.92 (0.99)	1.00	0.00	0.979	0.000	177.5
	FOSSA	0.549 (0.025)	5.39 (0.25)	1.00	0.22	0.998	0.000	6.7

more variables than necessary. FR appears to have an unstable parameter estimation with a large MSE, and also a relatively large model size.

Table 2.3: Simulation results for a quadratic Cox model with more terms than the sample size. Results are averaged over 100 replications. Numbers in parentheses are standard errors.

\tilde{p}	Method	MSE	Size	Cover	Exact	Corr 0	Incorr 0	Iter
230	ORACLE	0.168 (0.016)						
	LASSO	1.860 (0.082)	17.56 (0.57)	0.77	0.00	0.943	0.073	
	FR	14.473 (0.799)	19.12 (0.33)	0.90	0.00	0.937	0.024	
	Boosting	1.647 (0.082)	21.62 (0.73)	0.81	0.00	0.925	0.056	222.3
	FOSSA	1.742 (0.079)	9.71 (0.39)	0.70	0.01	0.977	0.110	13.1
495	ORACLE	0.138 (0.014)						
	LASSO	2.422 (0.083)	18.06 (0.74)	0.60	0.00	0.972	0.132	
	FR	27.891 (1.650)	20.93 (0.06)	0.90	0.00	0.967	0.034	
	Boosting	2.046 (0.078)	25.27 (1.01)	0.74	0.00	0.958	0.070	190.7
	FOSSA	2.007 (0.073)	11.45 (0.49)	0.64	0.02	0.986	0.116	14.3

2.4 Applications

We illustrate the application of our FOSSA method in the analysis of two microarray data. The first is a study of diffuse large-B-cell lymphoma (Rosenwald et al., 2002), which consists of $n = 240$ lymphoma patients and $p = 7399$ candidate genes. The response is the patient’s survival time after the chemotherapy, among which 102 are censored, yielding a 42.5% censoring proportion. The second is a breast cancer study (van Houwelingen et al., 2006), where there are $n = 295$ female patients with breast cancer and $p = 4919$ candidate genes. The response is again the survival time, and among 295 patients, 216 have censored response, which results in a relatively high censoring rate of 73%. In both studies, the goal is to identify important genes that affect the survival phenotype. Given the large value of p , we consider a linear Cox model for both data.

For the lymphoma data, the FOSSA algorithm converges in 12 iterations, yielding a selection of 11 genes out of 7399 candidates. Among them, six were also selected by Li and Luan (2005) in their analysis of the same dataset using a boosting method, and four belongs to the four gene expression signature groups identified by Rosenwald et al. (2002). We also note that, to apply the boosting algorithm, Li and Luan (2005) conducted a preliminary screening using a univariate Cox model to first bring the number of candidates from 7399 to 50. By contrast, our method does not require any pre-screening step. Those 11 genes are listed in Table 2.4, by the order that they enter the model in FOSSA, along with their gene ID’s and the estimated coefficients by FOSSA. We also fit a linear Cox model with only those 11 selected genes, and report the coefficient estimates and the corresponding p -values. It is seen that the estimates obtained from the final Cox model with high significance and those from FOSSA are quite compatible. In order to evaluate the predictive performance of the FOSSA method, we randomly split the data

into a training dataset of sample size 160 and a testing dataset of sample size 80. We first build the model with the training dataset and then define the high-risk and low-risk group of patients for each dataset based on the predictive model. The cutoff value is determined by the median of $\hat{\beta}^\top \mathbf{X}$ from the training set. Figure 2.1 shows the Kaplan-Meier estimates of survival curves together with the log-rank test results. It is seen that FOSSA achieves a good separation between these two risk groups for both the training and testing sets.

For the breast cancer data, FOSSA converges in 8 iterations, and selects 7 genes out of 4919 candidates. Table 2.4 lists those selected genes and their FOSSA coefficients. We again fit a Cox model using only those 7 genes, with all found significant by their p -values, and the two models yield compatible estimates. Predictive performance of FOSSA is also evaluated by randomly splitting the data into a training set of size 196 and a testing set of size 99. The Kaplan-Meier estimates and log-rank test results are reported in Figure 2.2. Again different risk groups determined by the predictive model built using FOSSA are highly significant for both the training and testing sets. These results indicate that FOSSA can be useful in building predictive models for censored survival data in high dimensional problems.

Table 2.4: Genes selected by FOSSA for the lymphoma data and the breast cancer data. Reported are the order of selection, the gene ID, the estimated coefficient by FOSSA, the estimated coefficient by the Cox model with only the selected genes and the corresponding p -values.

Diffuse large-B-cell lymphoma				
Order	Gene ID	$\hat{\beta}$ (FOSSA)	$\hat{\beta}$ (Cox)	p -value (Cox)
1	AA805575	-0.134	-0.105	2.31e-02
2	AA262133	0.764	1.497	4.01e-08
3	W46566	-0.276	-0.320	2.78e-02
4	AA830781	0.259	0.323	4.47e-02
5	AA243583	-0.274	-0.644	5.42e-06
6	AA293559	-0.106	-0.258	5.66e-03
7	LC_32424	0.176	0.544	1.36e-02
8	AA721746	0.102	0.148	3.27e-01
9	N48691	-0.078	-0.383	1.04e-01
10	AI219836	0.027	0.387	1.37e-02
11	AI391470	0.009	0.222	1.16e-01
Breast cancer				
Order	Gene ID	$\hat{\beta}$ (FOSSA)	$\hat{\beta}$ (Cox)	p -value (Cox)
1	NM_006607	2.502	2.741	1.69e-05
2	AL110226	1.516	2.562	2.43e-05
3	Contig58368_RC	0.788	1.136	2.73e-02
4	NM_002811	1.153	2.760	8.37e-04
5	NM_006399	-0.366	-1.410	1.11e-02
6	NM_006054	0.366	2.263	4.01e-03
7	NM_013290	0.159	1.304	2.35e-02

2.5 Discussion

In this chapter, we have proposed a forward stagewise shrinkage and addition method for model estimation and variable selection in Cox proportional hazards models with high dimensional covariates. It carries out an additive stagewise modeling while introducing shrinkage estimation at each iteration. Compared with the existing variable selection methods, our method performs comparably as the best players in the typical $p < n$ setup,

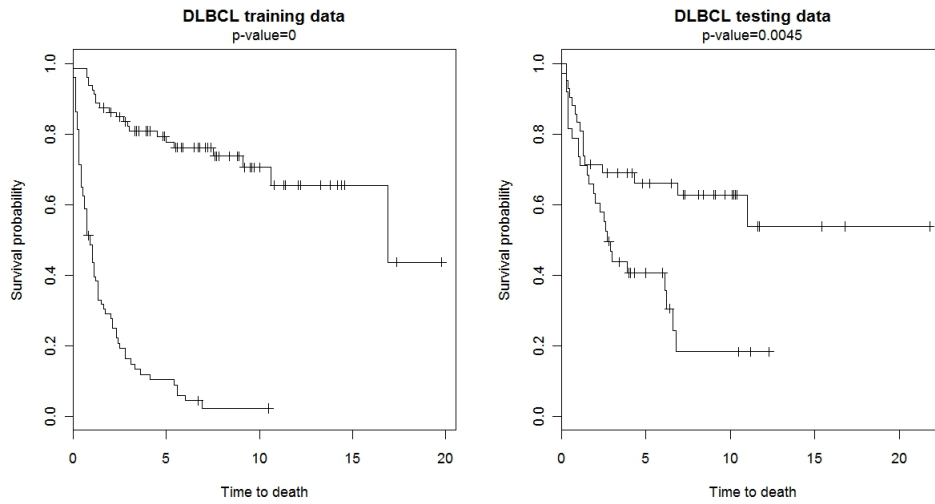


Figure 2.1: Kaplan-Meier estimates of survival curves for high and low risk patients for diffuse large-B-cell lymphoma data. Log-rank tests p values are given.

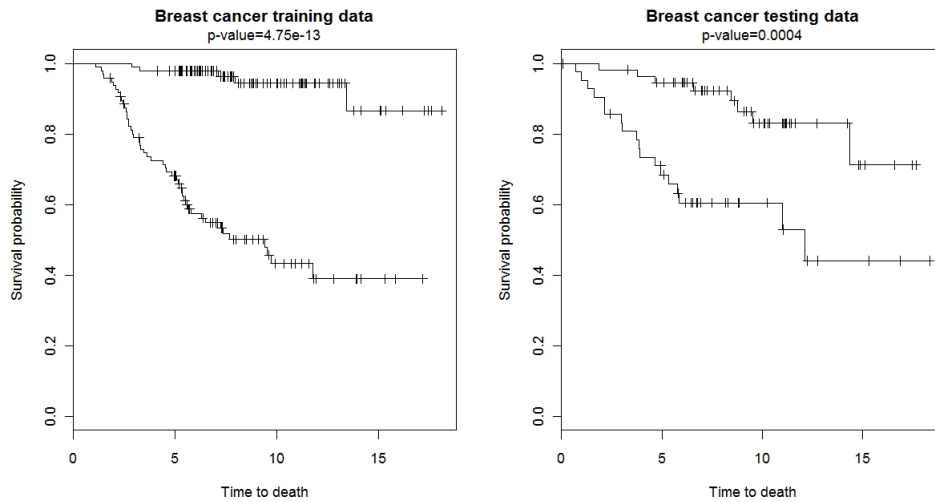


Figure 2.2: Kaplan-Meier estimates of survival curves for high and low risk patients for breast cancer data. Log-rank tests p values are given.

whereas it clearly outperforms the existing solutions when p far exceeds n . Compared with the existing boosting method, our solution explicitly conducts variable selection, and substantially reduces the number of iterations required and thus computing time by dropping the small learning rate in the usual boosting algorithm. Therefore, it provides a useful addition to the statistical toolbox for the analysis of high dimensional survival data.

When the association between the response and the covariates is more complicated than a linear relation, a common remedy is to introduce higher order polynomial terms, e.g., the quadratic and interaction terms, into the model. Although the same methodology for linear model selection can be applied here, our results in Section 2.3.3 indicate that the problem becomes harder than merely having a larger number of covariates. This can be partly attributed to the high correlations among the covariates and their polynomial terms. Our solution exhibits a competitive performance under this scenario. Moreover, our method can be straightforwardly extended to nonlinear Cox models by considering more flexible base learner, e.g., the component-wise cubic smoothing splines (Bühlmann and Yu, 2003; Li and Luan, 2005; Lu and Li, 2008), and employing the group LASSO type penalty (Yuan and Lin, 2006). This extension is certainly of interest for further investigation.

Chapter 3

Groupwise Sufficient Dimension Reduction via Envelope Method

3.1 Introduction

In many high dimensional data applications, the predictors originate from different domains or groups. For example, in a global surface temperature reconstruction study (Mann et al., 2008), 1,209 proxy series are recorded to characterize temperature changes over time. Since the instrumental climate record data are only available from the mid 19th century, in order to better understand the variability in climate and to address the question that whether the recent temperature increases are anomalous, these proxy information are used to reconstruct global surface temperatures over the past 2,000 years. One important feature of these 1,209 proxies is that they come from several distinct groups, such as tree rings, ice cores, cave deposits, lake sediments and historical documentation series. For this dataset, with such high dimensionality ($p=1,209$) in the predictors compared to the sample size ($n < 200$), it is highly desirable to summarize the data

via certain dimension reduction method before performing further analysis. Moreover, with the presence of the group information, it is usually more helpful to incorporate such group information into dimension reduction procedures than treating all predictors equally, as there are several advantages. Firstly, the results are more interpretable, and as a result also more informative. For example, it would be more helpful to get reduced data separated in different groups and know that certain groups of proxies have significant associations with temperature records, rather than obtain linear combinations of all 1,209 proxies as the reduced-dimensional predictors. Secondly, intuitively incorporating group information would lead to more accurate estimates as the total number of unknown parameters to be estimated in dimension reduction have been greatly reduced.

One of the traditional dimension reduction methods is the principal component analysis (PCA), which has been widely applied to climate research studies such as the above temperature data example. PCA reduces the dimension of predictors by constructing orthogonal linear transformations of predictors while preserving the maximum variation. A significant drawback of PCA is that the analysis is performed solely based on the covariance structure of the predictors, and no response information is involved. Cook (1998) developed the sufficient dimension reduction (SDR) concept, by which a reduction of the p dimensional predictor \mathbf{X} , usually through a projection of \mathbf{X} onto a lower dimensional space, is deemed sufficient if there is no information loss in the conditional distribution of the response Y given \mathbf{X} , or in the conditional mean of Y given \mathbf{X} after the reduction. More specifically, in the SDR framework, the dimension reduction subspace is defined as a subspace \mathcal{S} of \mathbb{R}^p , such that $F(Y|\mathbf{X}) = F(Y|P_{\mathcal{S}}\mathbf{X})$, where $P_{\mathcal{S}}$ is the projection matrix onto \mathcal{S} . Therefore, we have $Y \perp\!\!\!\perp \mathbf{X}|P_{\mathcal{S}}\mathbf{X}$, and the subspace \mathcal{S} characterizes all the information in \mathbf{X} such that the regression information in $F(Y|\mathbf{X})$ is fully preserved. Under mild conditions, the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ can be defined as the intersection of all

dimension reduction subspaces, and therefore has the minimum dimension (Cook, 1998). Similarly, \mathcal{S} is a mean dimension reduction subspace if $E(Y|\mathbf{X}) = E(Y|P_{\mathcal{S}}\mathbf{X})$, and the central mean subspace $\mathcal{S}_{E(Y|\mathbf{X})}$ is the intersection of all such mean dimension reduction subspaces (Cook and Li, 2002). Compared to PCA, SDR methods fully incorporate the information in both the predictors and response, and therefore are more response specific.

Many popular and well studied dimension reduction methods are SDR methods. One popular example is the sliced inverse regression (SIR; Li 1991), which recovers central subspace based on the conditional first moment. SIR estimators are obtained by the relation that $\Sigma^{-1}E(\mathbf{X}|Y)$ lies in $\mathcal{S}_{Y|\mathbf{X}}$ almost surely under the linearity condition (Li, 1991), where $\Sigma = \text{cov}(\mathbf{X})$. More specifically, the direction estimators given by SIR are $\{\Sigma^{-1/2}\boldsymbol{\eta}_i, i = 1, \dots, d\}$, where $\boldsymbol{\eta}_i$ are eigenvectors of $\text{cov}\{E(\mathbf{Z}|Y)\}$ with $\mathbf{Z} = \Sigma^{-1/2}\{\mathbf{X} - E(\mathbf{X})\}$. Sliced average variance estimator (SAVE; Cook and Weisberg 1991) utilizes the conditional second moment, and estimates the central subspace via $\Sigma^{-1}\{\Sigma - \text{var}(\mathbf{X}|Y)\}$. Directional regression (DR; Li and Wang 2007) combines the information in the first two conditional moments and provides direction estimates of the central subspace via $\Sigma^{-1}[2\Sigma - E\{(\mathbf{X} - \tilde{\mathbf{X}})(\mathbf{X} - \tilde{\mathbf{X}})^\top | Y, \tilde{Y}\}]$, where $(\tilde{\mathbf{X}}, \tilde{Y})$ is an independent copy of (\mathbf{X}, Y) . SDR methods that focus on the conditional mean $E(Y|\mathbf{X})$ include minimum average variance estimator (MAVE; Xia et al. 2002) and principal Hessian directions (PHD; Li 1992), among others. However, neither PCA nor the SDR methods account for prior group information during dimension reduction.

Recently, several approaches have been proposed to incorporate group information into dimension reduction procedures (Naik and Tsai, 2005; Li, 2009; Li et al., 2010b). Naik and Tsai (2005) proposed the constrained inverse regression (CIR) method, in which the kernel matrix of a classical dimension reduction method (e.g. $\text{cov}\{E(\mathbf{Z}|Y)\}$ in SIR) is projected onto a constrained subspace based on group information. Li (2009) incorporated

the group information by first performing dimension reduction on each group separately, then building the final estimate as the individual estimates assembled together. However, Li et al. (2010b) showed that CIR with SIR as the initial estimate and assembled SIR (aSIR) lead to equivalent population estimates, and neither is unbiased unless certain conditions hold, among which one requires that the predictors in different groups are conditionally independent given the response Y . It is easily seen that this is a fairly restrictive assumption which can be rarely met. The bias of the aSIR estimates come from the fact that directions in different groups are coupled with each other, and should be estimated simultaneously rather than separately. Li et al. (2010b) proposed a groupwise MAVE (gMAVE) method, in which a direct sum group structure is imposed into the differential operator of $E(Y|\mathbf{X})$ during the ordinary MAVE estimation. This method does not require any restrictive independence assumptions, and recovers the conditional mean information $E(Y|\mathbf{X})$ while preserving the group structure. Asymptotic consistency of the gMAVE estimator is also obtained.

In this chapter, we propose a groupwise sufficient dimension reduction procedure via the envelop method. The objective is to preserve the full regression information in the conditional distribution $F(Y|\mathbf{X})$ while incorporating prior group structure information in the predictors during dimension reduction. The underlying idea of our method is to cover any classical dimension reduction estimators via a direct sum envelope that is constrained by the desired group structure. More specifically, we estimate the direct sum envelope of any SDR estimator, which is defined as the smallest subspace that has the following two properties: a) covers the subspace spanned by the SDR estimator, and b) contains the group structure. As such, the prior group information is imposed onto the classical dimension reduction estimator, and thus one can preserve the group structure in the predictors while summarizing the true underlying directions. The key contributions

of this chapter are as follows. First of all, the proposed method can be applied to any classical estimator of the central subspace such as SIR, SAVE and DR. Secondly, the envelope idea is fairly general and flexible, and can be applied to impose other structures onto dimension reduction subspaces. For instance, in Li et al. (2010a), the dimension folding Kronecker envelope is imposed onto dimension reduction subspaces to preserve the matrix structure in the predictors. Moreover, compared to the gSAVE method which preserves the regression information in $E(Y|\mathbf{X})$, the proposed method targets at recovering $F(Y|\mathbf{X})$, and therefore is more general and can be applied to many more situations. The rest of this chapter is organized as follows. In Section 3.2, we provide a theoretical formulation of groupwise dimension reduction subspace, groupwise central subspace and direct sum envelope; in Section 3.3, the population and finite sample estimation procedures are discussed; in Section 3.4, the empirical performance of the proposed method is illustrated through simulation studies; in Section 3.5 the usefulness of the proposed method is demonstrated via real data analysis on the temperature-proxy dataset; in the end, a brief discussion is presented.

3.2 Groupwise Dimension Reduction Subspace and Direct Sum Envelope

3.2.1 Groupwise Dimension Reduction Subspace

In this section, we provide a mathematical formulation of groupwise dimension reduction subspace. Most of the definitions in this section are defined analogous to the ones in Li et al. (2010b), which are in the context of groupwise mean dimension reduction. We first define a *groupwise dimension reduction subspace*, which captures the full regression

information in $F(Y|\mathbf{X})$ while preserving the group structure in \mathbf{X} .

Let $\mathcal{S}_1, \dots, \mathcal{S}_g \subset \mathbb{R}^p$ be an orthogonal decomposition of \mathbb{R}^p , such that

$$\mathbb{R}^p = \mathcal{S}_1 \oplus \dots \oplus \mathcal{S}_g, \quad (3.1)$$

where \oplus denotes the direct sum operator. With the presence of group information, $\mathcal{S}_1, \dots, \mathcal{S}_g$ can be defined to represent the group structure in the predictors. For example, consider two groups of predictors (X_1, X_2, X_3) and (X_4, X_5) , then $\mathcal{S}_1 = \text{span}(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ and $\mathcal{S}_2 = \text{span}(\mathbf{e}_4, \mathbf{e}_5)$, where \mathbf{e}_i is a 5-dimensional vector with the i th element equals to 1 and others equal to 0, and $\text{span}(\mathbf{A})$ represents the subspace spanned by the columns of matrix \mathbf{A} . The *groupwise dimension reduction subspace* is then defined as follows.

Definition 3.1 For a given orthogonal decomposition $\{\mathcal{S}_1, \dots, \mathcal{S}_g\}$ of \mathbb{R}^p , if there are subspaces $\mathcal{T}_\ell \subseteq \mathcal{S}_\ell$ for $\ell = 1, \dots, g$ such that

$$F(Y|\mathbf{X}) = F(Y|P_{\mathcal{T}_1}\mathbf{X}, \dots, P_{\mathcal{T}_g}\mathbf{X}),$$

where $F(\cdot)$ denotes the cumulative distribution function, then $\mathcal{T}_1 \oplus \dots \oplus \mathcal{T}_g$ is a *groupwise dimension reduction subspace with respect to $\{\mathcal{S}_1, \dots, \mathcal{S}_g\}$* .

By definition, groupwise dimension reduction subspace is a subspace that recovers full regression information in the conditional distribution $F(Y|\mathbf{X})$ while preserving the group structure of the predictors. It is noted that a groupwise dimension reduction subspace itself is a dimension reduction subspace. In other words, a groupwise dimension reduction subspace must hold all the properties of a dimension reduction subspace, in addition to containing the group structure. In fact, it can be considered as a dimension reduction subspace constrained by the orthogonal decomposition in (3.1).

Similarly as the central subspace, a *groupwise central subspace* can be defined as the groupwise dimension reduction subspace that is contained in all groupwise dimension reduction subspaces. Such a groupwise central subspace with minimum dimension is of particular interest to us since it provides maximum dimension reduction without loss of information. One intuitive way of constructing the groupwise central subspace is to take the intersection of all groupwise dimension reduction subspaces. This first requires Lemma 1 of Li et al. (2010b), as follows:

Lemma 1 of Li et al. (2010b): “Suppose that $\mathcal{T}' = \mathcal{T}'_1 \oplus \dots \oplus \mathcal{T}'_g$ and $\mathcal{T}'' = \mathcal{T}''_1 \oplus \dots \oplus \mathcal{T}''_g$ with $\mathcal{T}'_\ell \subseteq \mathcal{S}_\ell$ and $\mathcal{T}''_\ell \subseteq \mathcal{S}_\ell$ for $\ell = 1, \dots, g$. Then $\mathcal{T}' \cap \mathcal{T}'' = (\mathcal{T}'_1 \cap \mathcal{T}''_1) \oplus \dots \oplus (\mathcal{T}'_g \cap \mathcal{T}''_g)$.”

This lemma states that the intersection of direct sums is equivalent to the direct sum of intersections. In addition, we assume the following implication holds throughout this chapter, which is true under mild conditions (Cook, 1998; Yin et al., 2008):

$$F(Y|\mathbf{X}) = F(Y|P_{\mathcal{T}'}\mathbf{X}) \text{ and } F(Y|\mathbf{X}) = F(Y|P_{\mathcal{T}''}\mathbf{X}) \Rightarrow F(Y|\mathbf{X}) = F(Y|P_{\mathcal{T}' \cap \mathcal{T}''}\mathbf{X}) \quad (3.2)$$

We then have the following lemma.

Lemma 3.1 *Suppose (3.2) holds, then for a given orthogonal decomposition $\{\mathcal{S}_1, \dots, \mathcal{S}_g\}$, the intersection of two groupwise dimension reduction subspaces with respect to $\{\mathcal{S}_1, \dots, \mathcal{S}_g\}$ is also a groupwise dimension reduction subspace with respect to the same orthogonal decomposition.*

The proof is easy and thus omitted here. This lemma states the property of closure under intersection of groupwise dimension reduction subspaces, and justifies the following definition of the *groupwise central subspace*.

Definition 3.2 *The intersection of all groupwise dimension reduction subspaces with*

respect to a given orthogonal decomposition $\{\mathcal{S}_1, \dots, \mathcal{S}_g\}$ is the groupwise central subspace with respect to $\{\mathcal{S}_1, \dots, \mathcal{S}_g\}$, which is denoted as $\mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g)$.

We note here that $\mathcal{S}_{Y|\mathbf{X}}$ is contained in any groupwise dimension reduction subspace. Consequently, we have $\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g)$. In other words, given the groupwise central subspace, it is possible to perform further dimension reduction on \mathbf{X} if the group structure of \mathbf{X} is disregarded, and imposing the group structure limits the maximum reduction one can perform. On the other hand, if the group structure is to be preserved, the groupwise central subspace is the smallest subspace that contains the full regression information in $F(Y|\mathbf{X})$.

Next we examine the groupwise central subspace from a basis spanning point of view. By construction, we have

$$\mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g) = \mathcal{T}_1^* \oplus \dots \oplus \mathcal{T}_g^*$$

for some subspaces $\mathcal{T}_1^* \subseteq \mathcal{S}_1, \dots, \mathcal{T}_g^* \subseteq \mathcal{S}_g$. Let the dimension of \mathcal{S}_ℓ and \mathcal{T}_ℓ^* for $\ell = 1, \dots, g$ be p_ℓ and d_ℓ respectively, and the dimensions of $\mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g)$ be d . Here $p_1 + \dots + p_g = p$ and $d_1 + \dots + d_g = d$. Let $\boldsymbol{\gamma}_\ell \in \mathbb{R}^{p \times p_\ell}$ denote a basis matrix of \mathcal{S}_ℓ , i.e., $\text{span}(\boldsymbol{\gamma}_\ell) = \mathcal{S}_\ell$. Let $\boldsymbol{\beta}_\ell \in \mathbb{R}^{p_\ell \times d_\ell}$ be a matrix satisfying $\text{span}(\boldsymbol{\gamma}_\ell \boldsymbol{\beta}_\ell) = \mathcal{T}_\ell^*$. In other words, the bases of each subspace \mathcal{T}_ℓ^* can be separated into two parts: the $\boldsymbol{\gamma}_\ell$'s which represent group information and the $\boldsymbol{\beta}_\ell$'s which are the true underlying directions within each group. In the example where there are two groups of predictors (X_1, X_2, X_3) and (X_4, X_5) , we have $p_1 = 3, p_2 = 2$, and $\boldsymbol{\gamma}_1 = (\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$, $\boldsymbol{\gamma}_2 = (\mathbf{e}_4, \mathbf{e}_5)$. Let us assume the true regression model is $Y = \exp(X_1 + X_2 - X_3) + (X_4 + X_5)^2$, that is, each group of predictors contribute to the model through a single linear combination of the members in the group. Therefore, we have $d_1 = d_2 = 1$, and $\boldsymbol{\beta}_1 = (1, 1, -1)^\top, \boldsymbol{\beta}_2 = (1, 1)^\top$. It follows that

$\mathcal{T}_1^* = \text{span}(\boldsymbol{\gamma}_1\boldsymbol{\beta}_1) = \text{span}\{(1, 1, -1, 0, 0)^\top\}$ and $\mathcal{T}_2^* = \text{span}(\boldsymbol{\gamma}_2\boldsymbol{\beta}_2) = \text{span}\{(0, 0, 0, 1, 1)^\top\}$.

It should be noted that the group information $\boldsymbol{\gamma}_\ell$'s are known as prior knowledge, and therefore are not the target of our study. Consequently, the primary interest in our study lies in the $\boldsymbol{\beta}_\ell$'s and they are the main targets of our groupwise dimension reduction estimation.

In addition, since the $\boldsymbol{\gamma}_\ell$'s are orthogonal to each other, we note the following equality (Conway, 1990):

$$\text{span}(\boldsymbol{\gamma}_1\boldsymbol{\beta}_1) \oplus \cdots \oplus \text{span}(\boldsymbol{\gamma}_g\boldsymbol{\beta}_g) = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_g) \text{span}(\boldsymbol{\beta}_1 \oplus \dots \oplus \boldsymbol{\beta}_g), \quad (3.3)$$

where for matrices, the operation \oplus denotes building a block diagonal matrix with the elements, i.e., $\boldsymbol{\beta}_1 \oplus \cdots \oplus \boldsymbol{\beta}_g$ is a block diagonal matrix with matrices $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_g$ on the diagonals. (3.3) suggests that the groupwise central subspace can be decomposed into the $\boldsymbol{\gamma}_\ell$'s and the $\boldsymbol{\beta}_\ell$'s, and as a result the estimation can be focused on the $\boldsymbol{\beta}_\ell$'s easily.

3.2.2 Direct Sum Envelope

We next propose the direct sum envelope method for groupwise dimension reduction. We starts with an existing SDR estimator, and with the direct sum envelope we aim to find the smallest subspace that not only covers the SDR estimator, but also preserves the group structure in the predictors. This idea is motivated by the Kronecker envelope method in Li et al. (2010a). They propose to use the Kronecker envelope to cover classical estimates of the central subspace to achieve dimension reduction while preserving the matrix structure of predictors. The main purpose of applying the Kronecker envelope in Li et al. (2010a) is to impose the matrix structure to existing estimates. Similarly, we propose to apply a direct sum envelope to classical estimates to impose the group

structure in the predictors. In order to do so, we first introduce the definition of the *direct sum envelope* of a given random matrix \mathbf{U} .

Definition 3.3 *Let \mathbf{U} be a $p \times r$ random vector or random matrix. For a given orthogonal decomposition $\{\mathcal{S}_1, \dots, \mathcal{S}_g\}$, define $\mathcal{F} = \{\mathcal{T}_1 \oplus \dots \oplus \mathcal{T}_g : \mathcal{T}_1 \subseteq \mathcal{S}_1, \dots, \mathcal{T}_g \subseteq \mathcal{S}_g, \text{span}(\mathbf{U}) \subseteq \mathcal{T}_1 \oplus \dots \oplus \mathcal{T}_g \text{ almost surely}\}$. Then the intersection of all members of \mathcal{F} is called the *direct sum envelope of \mathbf{U} with respect to $\{\mathcal{S}_1, \dots, \mathcal{S}_g\}$* . We denote such an envelope as $\mathcal{E}_{\{\mathcal{S}_1, \dots, \mathcal{S}_g\}}^\oplus(\mathbf{U})$.*

By definition, $\mathcal{E}_{\{\mathcal{S}_1, \dots, \mathcal{S}_g\}}^\oplus(\mathbf{U})$ is the smallest subspace that both covers $\text{span}(\mathbf{U})$ and conforms to the orthogonal decomposition in (3.1). As such, the group information can be effectively incorporated into SDR methods. The next proposition justifies the above definition of $\mathcal{E}_{\{\mathcal{S}_1, \dots, \mathcal{S}_g\}}^\oplus(\mathbf{U})$, guarantees its existence and assures it is uniquely defined.

Proposition 3.1 *The direct sum envelope $\mathcal{E}_{\{\mathcal{S}_1, \dots, \mathcal{S}_g\}}^\oplus(\mathbf{U})$ uniquely exists.*

PROOF. (a) By taking $\mathcal{T}_\ell = \mathcal{S}_\ell$ and (3.1), we see that \mathcal{F} is non-empty. (b) If $\text{span}(\mathbf{U}) \subseteq \mathcal{T}'_1 \oplus \dots \oplus \mathcal{T}'_g$ and $\text{span}(\mathbf{U}) \subseteq \mathcal{T}''_1 \oplus \dots \oplus \mathcal{T}''_g$, then $\text{span}(\mathbf{U}) \subseteq (\mathcal{T}'_1 \oplus \dots \oplus \mathcal{T}'_g) \cap (\mathcal{T}''_1 \oplus \dots \oplus \mathcal{T}''_g) = (\mathcal{T}'_1 \cap \mathcal{T}''_1) \oplus \dots \oplus (\mathcal{T}'_g \cap \mathcal{T}''_g)$. So \mathcal{F} is a π -system (Billingsley, 1986, p.36), i.e., the intersection of members of \mathcal{F} is still a member of \mathcal{F} . (c) The intersection, denoted as $\mathcal{T}^*_1 \oplus \dots \oplus \mathcal{T}^*_g$, of all members of \mathcal{F} is unique since it has the smallest dimension in the sense that $d_\ell \leq \tilde{d}_\ell$ for all $\ell = 1, \dots, g$, where d_ℓ is the dimension of \mathcal{T}^*_ℓ , and \tilde{d}_ℓ is the dimension of \mathcal{T}_ℓ . \square

3.3 Estimation via Envelope Method

In this section, we connect direct sum envelope with groupwise central subspace, and provide both population and sample estimation procedures of the groupwise central sub-

space.

3.3.1 Envelope for Groupwise Dimension Reduction

According to the definition of direct sum envelope, if the column space of the random matrix \mathbf{U} lies almost surely in a dimension reduction central subspace, then intuitively its direct sum envelope should be closely linked to the groupwise central subspace. If such a connection exists, the groupwise central subspace could be estimated through the corresponding direct sum envelope. Next we establish this connection and propose the estimator of groupwise central subspace. The next theorem is to serve as the foundation of our estimation of the groupwise central subspace.

Theorem 3.1 *If there is a random vector or matrix $\mathbf{U} \in \mathbb{R}^{p \times r}$ such that $\text{span}(\mathbf{U}) \subseteq \mathcal{S}_{Y|\mathbf{X}}$ almost surely, then*

$$\mathcal{E}_{\{\mathcal{S}_1, \dots, \mathcal{S}_g\}}^{\oplus}(\mathbf{U}) \subseteq \mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g).$$

PROOF. $\mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g)$ is a dimension reduction subspace. By the definition of $\mathcal{S}_{Y|\mathbf{X}}$, we have $\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g)$. It follows that $\text{span}(\mathbf{U}) \subseteq \mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g)$ almost surely. Also by definition, $\mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g) = \mathcal{T}_1^* \oplus \dots \oplus \mathcal{T}_g^*$ for some subspaces $\mathcal{T}_1^* \subseteq \mathcal{S}_1, \dots, \mathcal{T}_g^* \subseteq \mathcal{S}_g$. Therefore, $\mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g) \in \mathcal{F}$, where \mathcal{F} is defined in Definition 3.3. By the definition of direct sum envelope, $\mathcal{E}_{\{\mathcal{S}_1, \dots, \mathcal{S}_g\}}^{\oplus}(\mathbf{U})$ is the intersection of all members of \mathcal{F} . Hence, $\mathcal{E}_{\{\mathcal{S}_1, \dots, \mathcal{S}_g\}}^{\oplus}(\mathbf{U}) \subseteq \mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g)$. \square

This theorem states that if there exists a random vector or matrix \mathbf{U} such that its column space is a subspace in the classical dimensional reduction central subspace $\mathcal{S}_{Y|\mathbf{X}}$ almost

surely, then its direct sum envelope $\mathcal{E}_{\{\mathcal{S}_1, \dots, \mathcal{S}_g\}}^\oplus(\mathbf{U})$ lies in the groupwise central subspace. It implies that we can build up an estimate for $\mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g)$ from *any* classical estimate of $\mathcal{S}_{Y|\mathbf{X}}$ that is based on such random vectors or matrices. For example, for SIR, $\mathbf{U} = \Sigma^{-1}E(\mathbf{X}|Y)$ with $r = 1$; for SAVE, $\mathbf{U} = \Sigma^{-1}[\Sigma - \text{cov}(\mathbf{X}|Y)]$ with $r = p$; and for DR, $\mathbf{U} = \Sigma^{-1}[2\Sigma - E\{(\mathbf{X} - \tilde{\mathbf{X}})(\mathbf{X} - \tilde{\mathbf{X}})^\top | Y, \tilde{Y}\}]$ with $r = p$. We will denote the envelope method based on SIR, SAVE and DR as groupwise SIR (gSIR), groupwise SAVE (gSAVE) and groupwise DR (gDR) respectively in the sequel. Similar estimation method can also be developed by including those estimators of the central mean subspace as well, e.g., PHD, MAVe, and their variants.

It should be noted that if \mathbf{U} does not capture all the information in $\mathcal{S}_{Y|\mathbf{X}}$, then one cannot conclude that by Theorem 3.1 $\mathcal{E}_{\{\mathcal{S}_1, \dots, \mathcal{S}_g\}}^\oplus(\mathbf{U})$ would capture all the information in $\mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g)$ either. However, if \mathbf{U} captures the full information in $\mathcal{S}_{Y|\mathbf{X}}$, then it is desirable to see the corresponding property in $\mathcal{E}_{\{\mathcal{S}_1, \dots, \mathcal{S}_g\}}^\oplus(\mathbf{U})$, such that $\mathcal{E}_{\{\mathcal{S}_1, \dots, \mathcal{S}_g\}}^\oplus(\mathbf{U})$ covers the entire space of $\mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g)$, and this is called exhaustiveness. More specifically, let F_0 be the true distribution of (\mathbf{X}, Y) and F_n be the empirical distribution of the observed data $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$. Let $\beta(F_n)$ denote an estimator of a subspace \mathcal{S} , then it is said to be exhaustive if $\text{span}\{\beta(F_0)\} = \mathcal{S}$. With such a property the target subspace can be estimated explicitly. Examples of exhaustive estimators for the central subspace include SAVE and DR (Li and Wang, 2007). We next show that the proposed groupwise dimension reduction method inherits the exhaustiveness property, if the initial random matrix to be covered by the direct sum envelope is also an exhaustive estimator of $\mathcal{S}_{Y|\mathbf{X}}$, as stated in the following proposition.

Proposition 3.2 *If the estimator $\beta(F_n)$ of the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ is exhaustive, then the direct sum envelope of $\beta(F_n)$ is an exhaustive estimator of the groupwise central*

subspace $\mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g)$.

PROOF. By the exhaustiveness of $\beta(F_n)$, $\text{span}\{\beta(F_0)\} = \mathcal{S}_{Y|\mathbf{X}}$. From Theorem 3.1, $\mathcal{E} = \mathcal{E}_{\{\mathcal{S}_1, \dots, \mathcal{S}_g\}}^\oplus\{\beta(F_0)\} \subseteq \mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g)$. On the other hand, by $\text{span}\{\beta(F_0)\} \subseteq \mathcal{E}$, $\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{E}$. Thus \mathcal{E} satisfies that $F(Y|\mathbf{X}) = F(Y|P_{\mathcal{E}}\mathbf{X})$, and can be written as $\mathcal{T}_1 \oplus \dots \oplus \mathcal{T}_g$ for some subspaces $\mathcal{T}_1 \subseteq \mathcal{S}_1, \dots, \mathcal{T}_g \subseteq \mathcal{S}_g$. Hence, \mathcal{E} is a groupwise dimension reduction subspace. It follows that $\mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g) \subseteq \mathcal{E}$. Therefore, $\mathcal{E} = \mathcal{E}_{\{\mathcal{S}_1, \dots, \mathcal{S}_g\}}^\oplus\{\beta(F_0)\} = \mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g)$. \square

By this proposition, we can explicitly estimate $\mathcal{S}_{Y|\mathbf{X}}(\mathcal{S}_1, \dots, \mathcal{S}_g)$ by estimating the direct sum envelope of an exhaustive estimator of $\mathcal{S}_{Y|\mathbf{X}}$ (e.g. the SAVE and DR estimator).

3.3.2 The Objective Function

Next we discuss the estimation procedure of the direct sum envelope. We first consider a population level objective function and show that the solution to the minimization of this objective function would give the direct sum envelope that we desire. Letting $\|\cdot\|$ denote the Frobenius matrix norm. We have the following theorem.

Theorem 3.2 *Suppose that the elements of $\mathbf{U} \in \mathbb{R}^{p \times r}$ are measurable with respect to a random vector W and have finite variance. Consider the objective function*

$$L(\mathbf{b}_1, \dots, \mathbf{b}_g, \mathbf{f}) = E \left\| \Sigma^{1/2} \mathbf{U} - \Sigma^{1/2} \mathbf{\Gamma}(\mathbf{b}_1 \oplus \dots \oplus \mathbf{b}_g) \mathbf{f}(W) \right\|^2,$$

where $\mathbf{b}_\ell \in \mathbb{R}^{p_\ell \times d_\ell}$, f denotes a function mapping from the support of W to $\mathbb{R}^{d \times r}$ with finite second moment, $d = d_1 + \dots + d_g$, $\mathbf{\Gamma} = (\gamma_1, \dots, \gamma_g) \in \mathbb{R}^{p \times p}$. Let $(\mathbf{b}_1^*, \dots, \mathbf{b}_g^*, \mathbf{f}^*)$

denote the minimizer of $L(\mathbf{b}_1, \dots, \mathbf{b}_g, \mathbf{f})$. Then

$$\Gamma \text{span}(\mathbf{b}_1^* \oplus \dots \oplus \mathbf{b}_g^*) = \mathcal{E}_{\{\mathcal{S}_1, \dots, \mathcal{S}_g\}}^\oplus(\mathbf{U}).$$

PROOF. Write $\mathcal{E}_{\{\mathcal{S}_1, \dots, \mathcal{S}_g\}}^\oplus(\mathbf{U}) = \text{span}(\gamma_1 \mathbf{b}_1^0) \oplus \dots \oplus \text{span}(\gamma_g \mathbf{b}_g^0)$, which also equals $\Gamma \text{span}(\mathbf{b}_1^0 \oplus \dots \oplus \mathbf{b}_g^0)$ by (3.3). Since elements of \mathbf{U} are measurable with respect to W , there exists a random matrix $\phi(W) \in \mathbb{R}^{d \times r}$ such that $\mathbf{U} = \Gamma(\mathbf{b}_1^0 \oplus \dots \oplus \mathbf{b}_g^0)\phi(W)$. Then,

$$\Sigma^{1/2} \mathbf{U} = \Sigma^{1/2} \Gamma(\mathbf{b}_1^0 \oplus \dots \oplus \mathbf{b}_g^0)\phi(W).$$

So $L(\mathbf{b}_1, \dots, \mathbf{b}_g, \mathbf{f})$ reaches its minimum at 0 within the range of $(\mathbf{b}_1, \dots, \mathbf{b}_g, \mathbf{f})$. Therefore any minimizer $(\mathbf{b}_1^*, \dots, \mathbf{b}_g^*, \mathbf{f}^*)$ must satisfy $\Sigma^{1/2} \mathbf{U} = \Sigma^{1/2} \Gamma(\mathbf{b}_1^* \oplus \dots \oplus \mathbf{b}_g^*) \mathbf{f}^*(W)$ almost surely. Consequently,

$$\Gamma(\mathbf{b}_1^0 \oplus \dots \oplus \mathbf{b}_g^0)\phi(W) = \Gamma(\mathbf{b}_1^* \oplus \dots \oplus \mathbf{b}_g^*) \mathbf{f}^*(W) \text{ almost surely.}$$

As such $\Gamma \text{span}(\mathbf{b}_1^* \oplus \dots \oplus \mathbf{b}_g^*)$ contains \mathbf{U} almost surely, so it belongs to \mathcal{F} . But it has the same dimension as $\mathcal{E}_{\{\mathcal{S}_1, \dots, \mathcal{S}_g\}}^\oplus(\mathbf{U})$, and $\mathcal{E}_{\{\mathcal{S}_1, \dots, \mathcal{S}_g\}}^\oplus(\mathbf{U})$ is unique. Then the conclusion follows. \square

For gSIR, $W = Y$, and for gDR, $W = (Y, \tilde{Y})^\top$. Here it is important to note that again, as prior group information, $\Gamma = (\gamma_1, \dots, \gamma_g)$ are not the objective of our estimation. Following Li et al. (2010b), in this general objective function, we have $\mathbf{U} = \Sigma^{-1} E(\mathbf{X}|Y)$ for gSIR, $\mathbf{U} = \Sigma^{-1} \{\Sigma - \text{cov}(\mathbf{X}|Y)\} \Sigma^{-1/2}$ for gSAVE, and $\mathbf{U} = \Sigma^{-1} [2\Sigma - E\{(\mathbf{X} - \tilde{\mathbf{X}})(\mathbf{X} - \tilde{\mathbf{X}})^\top | Y, \tilde{Y}\}] \Sigma^{-1/2}$ for gDR.

3.3.3 Estimation

We propose to minimize the above objective function by alternatively estimating $(\mathbf{b}_1, \dots, \mathbf{b}_g)$ and \mathbf{f} . In this way, the minimization is achieved through iterations between two steps, each being a least squares problem with a closed-form solution. As such the computation is easy and fast.

Theorem 3.3 *The minimization of $L(\mathbf{b}_1, \dots, \mathbf{b}_g, \mathbf{f})$ can be broken into two steps.*

1. Given fixed $\mathbf{f}(W)$, the minimizer $(\mathbf{b}_1, \dots, \mathbf{b}_g)$ of $L(\mathbf{b}_1, \dots, \mathbf{b}_g, \mathbf{f})$ is

$$\begin{pmatrix} \text{vec}(\mathbf{b}_1) \\ \vdots \\ \text{vec}(\mathbf{b}_g) \end{pmatrix} = \{E(\mathbf{V}_2^\top \mathbf{V}_2)\}^{-1} E(\mathbf{V}_2^\top \mathbf{V}_1), \text{ where } \mathbf{f}(W) = \begin{pmatrix} \mathbf{f}_1(W) \\ \vdots \\ \mathbf{f}_g(W) \end{pmatrix},$$

with $\mathbf{f}_\ell(W) \in \mathbb{R}^{d_\ell \times r}$, $\ell = 1, \dots, g$, $\mathbf{V}_1 = \text{vec}(\boldsymbol{\Sigma}^{1/2} \mathbf{U}) \in \mathbb{R}^{pr \times 1}$, and $\mathbf{V}_2 = \{\mathbf{f}_1(W)^\top \otimes (\boldsymbol{\Sigma}^{1/2} \boldsymbol{\gamma}_1), \dots, \mathbf{f}_g(W)^\top \otimes (\boldsymbol{\Sigma}^{1/2} \boldsymbol{\gamma}_g)\} \in \mathbb{R}^{pr \times (p_1 d_1 + \dots + p_g d_g)}$.

2. Given fixed $(\mathbf{b}_1, \dots, \mathbf{b}_g)$, the minimizer \mathbf{f} of $L(\mathbf{b}_1, \dots, \mathbf{b}_g, \mathbf{f})$ is

$$\text{vec}\{\mathbf{f}(w)\} = (\mathbf{V}_3^\top \mathbf{V}_3)^{-1} \mathbf{V}_3^\top \mathbf{V}_1(w),$$

where $\mathbf{V}_1(w) = \text{vec}\{\boldsymbol{\Sigma}^{1/2} \mathbf{U}(w)\} \in \mathbb{R}^{pr \times 1}$, and $\mathbf{V}_3 = I_r \otimes \{\boldsymbol{\Sigma}^{1/2} \boldsymbol{\Gamma}(\mathbf{b}_1 \oplus \dots \oplus \mathbf{b}_g)\} \in \mathbb{R}^{pr \times dr}$.

PROOF. We first rewrite the objective function as

$$L(\mathbf{b}_1, \dots, \mathbf{b}_g, \mathbf{f}) = E \left\| \text{vec}(\boldsymbol{\Sigma}^{1/2} \mathbf{U}) - \text{vec}\{\boldsymbol{\Sigma}^{1/2} \boldsymbol{\Gamma}(\mathbf{b}_1 \oplus \dots \oplus \mathbf{b}_g) \mathbf{f}(W)\} \right\|^2.$$

The first conclusion is obtained by noting that

$$\begin{aligned}
& \text{vec}\{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Gamma}(\mathbf{b}_1 \oplus \dots \oplus \mathbf{b}_g)\mathbf{f}(W)\} \\
&= \sum_{\ell=1}^g \text{vec}\{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\gamma}_\ell \mathbf{b}_\ell \mathbf{f}_\ell(W)\} \\
&= \sum_{\ell=1}^g \{\mathbf{f}_\ell(W)^\top \otimes (\boldsymbol{\Sigma}^{1/2}\boldsymbol{\gamma}_\ell)\} \text{vec}(\mathbf{b}_\ell) \\
&= \{\mathbf{f}_1(W)^\top \otimes (\boldsymbol{\Sigma}^{1/2}\boldsymbol{\gamma}_1), \dots, \mathbf{f}_g(W)^\top \otimes (\boldsymbol{\Sigma}^{1/2}\boldsymbol{\gamma}_g)\} \begin{pmatrix} \text{vec}(\mathbf{b}_1) \\ \vdots \\ \text{vec}(\mathbf{b}_g) \end{pmatrix}.
\end{aligned}$$

Then the result follows because the minimizer of $E\|\mathbf{V}_1 - \mathbf{V}_2\mathbf{c}\|^2$ over $\mathbf{c} \in \mathbb{R}^{p_1d_1+\dots+p_gd_g}$ is $\{E(\mathbf{V}_2^\top\mathbf{V}_2)\}^{-1}E(\mathbf{V}_2^\top\mathbf{V}_1)$.

The second conclusion is obtained by noting that, for each fixed w in the support of W , $\mathbf{f}(w)$ is the minimizer of

$$\begin{aligned}
& E \left\{ \left\| \text{vec}(\boldsymbol{\Sigma}^{1/2}\mathbf{U}) - \text{vec}\{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Gamma}(\mathbf{b}_1 \oplus \dots \oplus \mathbf{b}_g)\}\mathbf{f} \right\|^2 \mid W = w \right\} \\
&= \left\| \text{vec}\{\boldsymbol{\Sigma}^{1/2}\mathbf{U}(w)\} - \text{vec}\{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Gamma}(\mathbf{b}_1 \oplus \dots \oplus \mathbf{b}_g)\}\mathbf{f}(w) \right\|^2 \\
&= \left\| \text{vec}\{\boldsymbol{\Sigma}^{1/2}\mathbf{U}(w)\} - (I_r \otimes \{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Gamma}(\mathbf{b}_1 \oplus \dots \oplus \mathbf{b}_g)\}) \text{vec}\{\mathbf{f}(w)\} \right\|^2.
\end{aligned}$$

Then the result follows because the minimizer of $\|\mathbf{V}_1(w) - \mathbf{V}_3\mathbf{c}\|^2$ over $\mathbf{c} \in \mathbb{R}^{dr}$ is $(\mathbf{V}_3^\top\mathbf{V}_3)^{-1}\mathbf{V}_3^\top\mathbf{V}_1(w)$. \square

3.3.4 Numerical Procedures

We next provide the numerical procedure that minimizes the sample counterpart of $L(\mathbf{b}_1, \dots, \mathbf{b}_g, \mathbf{f})$. Consider an independent and identically distributed (i.i.d.) sample

$(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$, the sample covariance matrix Σ of the predictors \mathbf{X} is estimated by

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top.$$

Without loss of generality, we assume \mathbf{X} is centered at 0 in the sequel.

Following many classical dimension reduction methods (e.g. SIR, SAVE and DR), we first slice the response Y by finding a partition J_1, \dots, J_s on the support of Y . We first illustrate the numerical procedures by describing the algorithm for gSIR as follows.

Step 1. Obtain initial estimates of $(\mathbf{b}_1, \dots, \mathbf{b}_g)$. We use assembled SIR (aSIR) to obtain the initial estimates. Regress Y on $\gamma_1^\top \mathbf{X}, \dots, \gamma_g^\top \mathbf{X}$ separately, let $\hat{\mathbf{b}}_l$ be the first d_l dimensions of the SIR estimates for $\widehat{S}_{Y|\gamma_l^\top \mathbf{X}}$ for $l = 1, \dots, g$.

Step 2. Given $(\mathbf{b}_1, \dots, \mathbf{b}_g)$, update the estimate of $\mathbf{f}(W)$ (for gSIR, $W = Y$ and $r = 1$). For $h = 1, \dots, s$, compute

$$\widehat{\mathbf{U}}(h) = \widehat{\Sigma}^{-1} E_n(\mathbf{X} | Y \in J_h) = \widehat{\Sigma}^{-1} \frac{1}{n_h} \sum_{i=1}^n \mathbf{X}_i I(Y_i \in J_h),$$

$$\widehat{\mathbf{V}}_1(h) = \text{vec}\{\widehat{\Sigma}^{1/2} \widehat{\mathbf{U}}(h)\} = \text{vec}\{\widehat{\Sigma}^{-1/2} \frac{1}{n_h} \sum_{i=1}^n \mathbf{X}_i I(Y_i \in J_h)\},$$

$$\widehat{\mathbf{V}}_3 = I_r \otimes \{\widehat{\Sigma}^{1/2} \Gamma(\hat{\mathbf{b}}_1 \oplus \dots \oplus \hat{\mathbf{b}}_g)\} = \widehat{\Sigma}^{1/2} \Gamma(\hat{\mathbf{b}}_1 \oplus \dots \oplus \hat{\mathbf{b}}_g),$$

where n_h is the number of observations in partition J_h . Then compute $\text{vec}\{\hat{\mathbf{f}}(h)\}$ by

$$\text{vec}\{\hat{\mathbf{f}}(h)\} = (\widehat{\mathbf{V}}_3^\top \widehat{\mathbf{V}}_3)^{-1} \widehat{\mathbf{V}}_3^\top \widehat{\mathbf{V}}_1(h).$$

Step 3. Given $\mathbf{f}(W)$, update the estimate of $(\mathbf{b}_1, \dots, \mathbf{b}_g)$. For $h = 1, \dots, s$, compute

$$\widehat{\mathbf{V}}_2(h) = \{\hat{\mathbf{f}}_1(h)^\top \otimes (\widehat{\Sigma}^{1/2}\gamma_1), \dots, \hat{\mathbf{f}}_g(h)^\top \otimes (\widehat{\Sigma}^{1/2}\gamma_g)\}.$$

Then compute

$$\begin{pmatrix} \text{vec}(\hat{\mathbf{b}}_1) \\ \vdots \\ \text{vec}(\hat{\mathbf{b}}_g) \end{pmatrix} = \left\{ \sum_{h=1}^s \frac{n_h}{n} \widehat{\mathbf{V}}_2(h)^\top \widehat{\mathbf{V}}_2(h) \right\}^{-1} \left\{ \sum_{h=1}^s \frac{n_h}{n} \widehat{\mathbf{V}}_2(h)^\top \widehat{\mathbf{V}}_1(h) \right\}.$$

Step 4. Return to Step 2 and iterate, until

$$\sum_{h=1}^s \frac{n_h}{n} \left\| \{ \widehat{\Sigma}^{1/2} \widehat{\mathbf{U}}(h) \} - \{ \widehat{\Sigma}^{1/2} \Gamma(\hat{\mathbf{b}}_1 \oplus \dots \oplus \hat{\mathbf{b}}_g) \hat{\mathbf{f}}(h) \} \right\|^2$$

converges.

The algorithm for gSAVE is similar to gSIR and therefore omitted here. For the gDR method, the algorithm is also similar, except that because of the presence of $(\tilde{\mathbf{X}}, \tilde{Y})$ the terms are more complicated to compute. Here we have $\mathbf{U} = \Sigma^{-1} [2\Sigma - E\{(\mathbf{X} - \tilde{\mathbf{X}})(\mathbf{X} - \tilde{\mathbf{X}})^\top | Y, \tilde{Y}\}] \Sigma^{-1/2}$ with $r = p$ and $W = (Y, \tilde{Y})^\top$. As given in Li et al. (2010b), for $h, k = 1, \dots, s$, we have the following conditional expectation

$$\begin{aligned} & E\{(\mathbf{X} - \tilde{\mathbf{X}})(\mathbf{X} - \tilde{\mathbf{X}})^\top | Y \in J_h, \tilde{Y} \in J_k\} \\ &= E(\mathbf{X}\mathbf{X}^\top | Y \in J_h) - E(\mathbf{X} | Y \in J_h)E(\tilde{\mathbf{X}}^\top | \tilde{Y} \in J_k) \\ &\quad - E(\tilde{\mathbf{X}} | \tilde{Y} \in J_k)E(\mathbf{X}^\top | Y \in J_h) + E(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top | \tilde{Y} \in J_k). \end{aligned}$$

Its sample estimate is given by

$$\begin{aligned}
E_n(h, k) &= E_n\{(\mathbf{X} - \tilde{\mathbf{X}})(\mathbf{X} - \tilde{\mathbf{X}})^\top | Y \in J_h, \tilde{Y} \in J_k\} \\
&= \frac{1}{n_h} \sum_{Y_i \in J_h} \mathbf{X}_i \mathbf{X}_i^\top - \frac{1}{n_h n_k} \sum_{Y_i \in J_h} \mathbf{X}_i \sum_{\tilde{Y}_j \in J_k} \tilde{\mathbf{X}}_j^\top \\
&\quad - \frac{1}{n_h n_k} \sum_{\tilde{Y}_j \in J_k} \tilde{\mathbf{X}}_j \sum_{Y_i \in J_h} \mathbf{X}_i^\top + \frac{1}{n_k} \sum_{\tilde{Y}_j \in J_k} \tilde{\mathbf{X}}_j \tilde{\mathbf{X}}_j^\top.
\end{aligned}$$

The numerical procedure for gDR can be summarized as follows:

Step 1. Obtain initial estimates of $(\mathbf{b}_1, \dots, \mathbf{b}_g)$. Same as in gSIR, we use aSIR estimates to obtain the initial estimates of $(\mathbf{b}_1, \dots, \mathbf{b}_g)$, i.e., let $\hat{\mathbf{b}}_l$ be the first d_l dimensions of the SIR estimates for $\hat{S}_{Y|\gamma_l^\top \mathbf{X}}$ for $l = 1, \dots, g$.

Step 2. Given $(\mathbf{b}_1, \dots, \mathbf{b}_g)$, update the estimate of $\mathbf{f}(W)$ (for gDR, $W = (Y, \tilde{Y})^\top$ and $r = p$). For $h, k = 1, \dots, s$, compute

$$\hat{\mathbf{U}}(h, k) = \hat{\Sigma}^{-1} \{2\hat{\Sigma} - E_n(h, k)\} \hat{\Sigma}^{-1/2},$$

$$\hat{\mathbf{V}}_1(h, k) = \text{vec}\{\hat{\Sigma}^{1/2} \hat{\mathbf{U}}(h, k)\} = \text{vec}\{2I_p - \hat{\Sigma}^{-1/2} E_n(h, k) \hat{\Sigma}^{-1/2}\},$$

$$\hat{\mathbf{V}}_3 = I_r \otimes \{\hat{\Sigma}^{1/2} \Gamma(\hat{\mathbf{b}}_1 \oplus \dots \oplus \hat{\mathbf{b}}_g)\}.$$

Then compute $\text{vec}\{\hat{\mathbf{f}}(h, k)\}$ by

$$\text{vec}\{\hat{\mathbf{f}}(h, k)\} = (\hat{\mathbf{V}}_3^\top \hat{\mathbf{V}}_3)^{-1} \hat{\mathbf{V}}_3^\top \hat{\mathbf{V}}_1(h, k).$$

Step 3. Given $\mathbf{f}(W)$, update the estimate of $(\mathbf{b}_1, \dots, \mathbf{b}_g)$. For $h, k = 1, \dots, s$, compute

$$\widehat{\mathbf{V}}_2(h, k) = \{\hat{\mathbf{f}}_1(h, k)^\top \otimes (\widehat{\Sigma}^{1/2} \boldsymbol{\gamma}_1), \dots, \hat{\mathbf{f}}_g(h, k)^\top \otimes (\widehat{\Sigma}^{1/2} \boldsymbol{\gamma}_g)\}.$$

Then compute

$$\begin{pmatrix} \text{vec}(\hat{\mathbf{b}}_1) \\ \vdots \\ \text{vec}(\hat{\mathbf{b}}_g) \end{pmatrix} = \left\{ \sum_{h=1}^s \sum_{k=1}^s \frac{n_h n_k}{n^2} \widehat{\mathbf{V}}_2(h, k)^\top \widehat{\mathbf{V}}_2(h, k) \right\}^{-1} \left\{ \sum_{h=1}^s \sum_{k=1}^s \frac{n_h n_k}{n^2} \widehat{\mathbf{V}}_2(h, k)^\top \widehat{\mathbf{V}}_1(h, k) \right\}.$$

Step 4. Return to Step 2 and iterate, until

$$\sum_{h=1}^s \sum_{k=1}^s \frac{n_h n_k}{n^2} \left\| \{ \widehat{\Sigma}^{1/2} \widehat{\mathbf{U}}(h, k) \} - \{ \widehat{\Sigma}^{1/2} \boldsymbol{\Gamma}(\hat{\mathbf{b}}_1 \oplus \dots \oplus \hat{\mathbf{b}}_g) \hat{\mathbf{f}}(h, k) \} \right\|^2$$

converges.

3.3.5 Dimension Estimation

In the above estimation procedures, we assume the true dimension of each group $\{d_1, d_2, \dots, d_g\}$ is known. However, such information is usually not available in practical applications. Therefore, certain method for estimating $\{d_1, d_2, \dots, d_g\}$ is desirable. We suggest using a Bayesian information criterion (BIC) type method to determine the dimensions for the proposed method. Let $L(d_{w1}, d_{w2}, \dots, d_{wg})$ be the minimum value of the objective function obtained from the minimization procedure described in Section 3.3.4 when the working dimensions are chosen as $\{d_{w1}, d_{w2}, \dots, d_{wg}\}$, then the BIC-type

criterion is in the following form:

$$BIC(d_{w1}, d_{w2}, \dots, d_{wg}) = L(d_{w1}, d_{w2}, \dots, d_{wg}) + C_n \log(n) p_w, \quad (3.4)$$

where C_n is a scaling factor and p_w is the number of unknown parameters corresponding to $\{d_{w1}, d_{w2}, \dots, d_{wg}\}$ (i.e. $p_w = \sum_{l=1}^g p_l \times d_{wl} + r \times \sum_{l=1}^g d_{wl}$). The estimates $\{\hat{d}_1, \hat{d}_2, \dots, \hat{d}_g\}$ are obtained by minimizing the BIC-type criterion given in (3.4). We have numerically tested the performance of this criterion with $C_n = 0.05\sqrt{L(1, 1, \dots, 1)}n^{-1/2}$, and the results are presented in Section 3.4.6. Optimal choice of C_n and theoretical properties of this dimension estimation method require further investigation.

3.4 Simulations

We perform simulation studies to illustrate the empirical performance of the proposed method under various situations. SIR and DR estimates are chosen as the initial random matrix \mathbf{U} to demonstrate the performance of the proposed envelope method. It should be noted that the proposed method can also be applied to SAVE, as well as other SDR estimators. However, because the unstable performance of SAVE for monotone trends with small to medium sample sizes (Li and Wang, 2007), the results are omitted here. The performance of aSIR and aDR (Li, 2009) are compared with gSIR and gDR. For methods that require slicing of the response variable, the number of slices are chosen as 10. To measure the estimation accuracy, median and median absolute deviation (adjusted by the constant 1.4826) of vector correlation coefficients (Hotelling, 1936) over 100 replications are reported.

3.4.1 A General Model

We first consider a general regression model in which the relationships between the response variable and the predictors are nonlinear, and the predictors are involved in both the mean and the variance components of the regression function. The model is in the following form:

$$Y = \exp(0.2\boldsymbol{\beta}_1^\top \mathbf{V}_1) + 0.5 \sin(0.2\pi\boldsymbol{\beta}_2^\top \mathbf{V}_2) + 0.01(4 + 0.1\boldsymbol{\beta}_1^\top \mathbf{V}_1 + \boldsymbol{\beta}_3^\top \mathbf{V}_3)^2\epsilon, \quad (3.5)$$

where $\mathbf{X} = (\mathbf{V}_1^\top, \mathbf{V}_2^\top, \mathbf{V}_3^\top)^\top$ is a $p = 20$ dimensional predictor and multivariate normal distributed with mean zero and compound symmetry covariance matrix of variance one and covariance 0.5. The dimensions for the three groups \mathbf{V}_1 , \mathbf{V}_2 and \mathbf{V}_3 are $p_1 = 10$, $p_2 = 5$ and $p_3 = 5$, respectively. ϵ follows a standard normal distribution and is independent of \mathbf{X} . The true regression coefficients are $\boldsymbol{\beta}_1 = (1, -1, 0, \dots, 0)^\top$, $\boldsymbol{\beta}_2 = (1, 0, \dots, 0)^\top$, and $\boldsymbol{\beta}_3 = (0, \dots, 0, 1, -1)^\top$. We vary the sample size n as 100, 200, 400, 800 and 1200. The set of working dimensions are fixed at true values (1,1,1) for all four methods.

Table 3.1 reports the simulation results for Model (3.5). Median and median absolute deviation (in parentheses, adjusted by the constant 1.4826) of vector correlation coefficients over 100 replications are shown. gSIR and gDR achieve better performance than aSIR and aDR for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_3$. Note that the estimation of $\boldsymbol{\beta}_3$ becomes particularly challenging since the information regarding $\boldsymbol{\beta}_3$ is only contained in the variance component. In fact aSIR and aDR do not achieve better performance in estimating $\boldsymbol{\beta}_3$ as n increases. On the contrary, gSIR and gDR obtain better estimates as n increases and overall clearly outperform aSIR and aDR for $\boldsymbol{\beta}_3$. For $\boldsymbol{\beta}_2$, all four methods provide similar results, especially for large n . This is probably due to the fact that the regression model related to $\boldsymbol{\beta}_2$ is relatively simple and consequently the estimation is easy.

Table 3.1: Simulation results for Model (3.5). Median and median absolute deviation (in parentheses) of vector correlation coefficient over 100 replications are shown.

	Method	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 1200$
β_1	gSIR	0.890 (0.071)	0.962 (0.023)	0.987 (0.006)	0.995 (0.003)	0.996 (0.002)
	aSIR	0.885 (0.048)	0.928 (0.023)	0.950 (0.012)	0.957 (0.007)	0.961 (0.005)
	gDR	0.854 (0.086)	0.942 (0.026)	0.979 (0.010)	0.991 (0.005)	0.995 (0.003)
	aDR	0.837 (0.090)	0.913 (0.035)	0.943 (0.018)	0.956 (0.009)	0.962 (0.005)
β_2	gSIR	0.944 (0.044)	0.986 (0.011)	0.994 (0.005)	0.997 (0.002)	0.998 (0.001)
	aSIR	0.955 (0.032)	0.982 (0.014)	0.991 (0.006)	0.996 (0.003)	0.998 (0.002)
	gDR	0.933 (0.059)	0.974 (0.023)	0.989 (0.008)	0.996 (0.003)	0.998 (0.002)
	aDR	0.920 (0.062)	0.972 (0.024)	0.988 (0.010)	0.995 (0.005)	0.997 (0.003)
β_3	gSIR	0.441 (0.354)	0.422 (0.299)	0.620 (0.279)	0.821 (0.170)	0.870 (0.100)
	aSIR	0.357 (0.276)	0.199 (0.187)	0.204 (0.196)	0.139 (0.115)	0.118 (0.094)
	gDR	0.338 (0.276)	0.399 (0.290)	0.510 (0.354)	0.625 (0.321)	0.683 (0.260)
	aDR	0.370 (0.286)	0.342 (0.279)	0.330 (0.278)	0.239 (0.184)	0.186 (0.152)

3.4.2 Various Correlations Among Predictors

In the second simulated example, we examine the change in performance of these four methods when the correlation among predictors increases. The regression model for this example is given as follows:

$$Y = \text{sign}(\beta_1^\top \mathbf{V}_1 + \epsilon_1) \log(|\beta_2^\top \mathbf{V}_2 + 5 + 2\epsilon_2|). \quad (3.6)$$

Here $\mathbf{X} = (\mathbf{V}_1^\top, \mathbf{V}_2^\top)^\top$ is a $p = 20$ dimensional predictor, multivariate normal distributed with mean zero and compound symmetry covariance matrix of variance one and covariance ρ , where $\rho = 0.2, 0.5, 0.8$. The dimensions for the two groups $\mathbf{V}_1, \mathbf{V}_2$ are $p_1 = 10$ and $p_2 = 10$, respectively. ϵ_1, ϵ_2 are independent standard normal errors. We set the values of β 's as $\beta_1 = (1, 1, 1, 0, \dots, 0)^\top$ and $\beta_2 = (1, -1, -1, 0, \dots, 0)^\top$. We also vary the sample size n as $n = 100, 200, 400, 800$ and 1200 . The results for all n show similar patterns and only the results for $n = 800$ are reported.

Table 3.2: Simulation results for Model (3.6) with $n = 800$. Median and median absolute deviation (in parentheses) of vector correlation coefficient over 100 replications are shown.

	Method	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$
β_1	gSIR	0.988 (0.005)	0.977 (0.014)	0.931 (0.029)
	aSIR	0.988 (0.006)	0.976 (0.014)	0.927 (0.034)
	gDR	0.983 (0.010)	0.966 (0.014)	0.874 (0.076)
	aDR	0.981 (0.010)	0.967 (0.019)	0.896 (0.051)
β_2	gSIR	0.988 (0.006)	0.980 (0.010)	0.955 (0.021)
	aSIR	0.986 (0.008)	0.881 (0.055)	0.701 (0.124)
	gDR	0.984 (0.007)	0.972 (0.016)	0.897 (0.051)
	aDR	0.977 (0.009)	0.931 (0.034)	0.831 (0.072)

Table 3.2 reports the results for Model (3.6) with $n = 800$. The performance of all methods deteriorate as ρ increases. However, the impact of the increase of ρ is much smaller on gSIR and gDR, especially for β_2 . It has been noted in Li et al. (2010b) that the assembled methods can only obtain unbiased estimates under certain conditions, one of which is that predictors from different groups are conditionally independent given the response Y (i.e., $\mathbf{V}_1 \perp\!\!\!\perp \mathbf{V}_2|Y$). Therefore, we expect the performance of aSIR and aDR to deteriorate as ρ increases. In addition, as the correlation increases, the relationship between the response variable and predictors becomes more complicated, and therefore the increase of ρ might also show an impact on gSIR and gDR, but a relatively small one. The simulation results indeed support these two points.

3.4.3 A More Challenging Model

In the models we have examined so far, the number of true underlying directions in each group is one. In this section we consider a more challenging example, in which the true

Table 3.3: Simulation results for Model (3.7). Median and median absolute deviation (in parentheses) of vector correlation coefficient over 100 replications are shown.

	Method	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 1200$
β_1	gSIR	0.502 (0.300)	0.561 (0.277)	0.757 (0.232)	0.879 (0.090)	0.923 (0.054)
	aSIR	0.435 (0.328)	0.499 (0.245)	0.622 (0.223)	0.649 (0.190)	0.679 (0.135)
	gDR	0.526 (0.320)	0.558 (0.296)	0.776 (0.243)	0.883 (0.103)	0.946 (0.042)
	aDR	0.441 (0.272)	0.528 (0.283)	0.662 (0.237)	0.715 (0.165)	0.734 (0.132)
β_2	gSIR	0.422 (0.319)	0.571 (0.299)	0.771 (0.206)	0.868 (0.107)	0.895 (0.085)
	aSIR	0.430 (0.298)	0.580 (0.331)	0.712 (0.228)	0.766 (0.142)	0.806 (0.138)
	gDR	0.326 (0.284)	0.536 (0.410)	0.728 (0.205)	0.883 (0.109)	0.928 (0.053)
	aDR	0.493 (0.306)	0.635 (0.299)	0.745 (0.230)	0.825 (0.119)	0.847 (0.106)

dimensions for each group are higher than one. The regression model is as follows:

$$Y = (2 + \beta_{11}^\top \mathbf{V}_1)^2 + \exp(0.8\beta_{12}^\top \mathbf{V}_1) + 2\beta_{21}^\top \mathbf{V}_2 + 2 \sin(0.3\pi\beta_{22}^\top \mathbf{V}_2) + 0.1\epsilon. \quad (3.7)$$

Here $\mathbf{X} = (\mathbf{V}_1^\top, \mathbf{V}_2^\top)^\top$ is a $p = 10$ dimensional predictor, multivariate normal distributed with mean zero and compound symmetry covariance matrix of variance one and covariance 0.5. \mathbf{V}_1 and \mathbf{V}_2 each has dimension 5. ϵ is again a standard normal error. The true regression coefficients are set as $\beta_1 = (\beta_{11}, \beta_{12})$, with $\beta_{11} = (1, 1, 0, 0, 0)^\top$, $\beta_{12} = (0, 0, 0, 1, -1)^\top$, and $\beta_2 = (\beta_{21}, \beta_{22})$, with $\beta_{21} = (1, 1, 0, 0, 0)^\top$ and $\beta_{22} = (0, 1, 1, 0, 0)^\top$. In this model, the true dimension for both groups is two. We again vary the sample size n as $n = 100, 200, 400, 800$ and 1200 .

Table 3.3 summarizes the results for Model (3.7). The results show that gSIR and gDR again outperform aSIR and aDR, for both β_1 and β_2 , especially for n large. These results suggest that the proposed method work well for relatively complicated models.

Table 3.4: Simulation results for Model (3.8). Median and median absolute deviation (in parentheses) of vector correlation coefficient over 100 replications are shown.

Method	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 1200$
gSIR	0.907 (0.055)	0.980 (0.011)	0.995 (0.002)	0.998 (0.001)	0.999 (0.001)
SIR	0.980 (0.008)	0.991 (0.003)	0.996 (0.001)	0.998 (0.001)	0.999 (0.000)
gDR	0.882 (0.059)	0.965 (0.017)	0.988 (0.005)	0.996 (0.002)	0.997 (0.001)
DR	0.960 (0.015)	0.984 (0.006)	0.992 (0.002)	0.996 (0.001)	0.998 (0.001)

3.4.4 No Group Structure

In some situations, there is no underlying group structure in the predictors. In this example, we examine the performance of gSIR and gDR while imposing an “incorrect” group structure under such a situation. We consider the following model:

$$Y = \exp(0.2\boldsymbol{\beta}^\top \mathbf{X}) + 0.1\epsilon. \quad (3.8)$$

Here $\mathbf{X} = (\mathbf{V}_1^\top, \mathbf{V}_2^\top, \mathbf{V}_3^\top)^\top$ is a $p = 20$ dimensional predictor, multivariate normal distributed with mean zero and compound symmetry covariance matrix of variance one and covariance 0.5. The pseudo-grouped predictors $\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3$ have dimensions $p_1 = 10, p_2 = 5$ and $p_3 = 5$ respectively. ϵ follows a standard normal distribution. We have $\boldsymbol{\beta}_1 = (1, -1, 0, \dots, 0)^\top, \boldsymbol{\beta}_2 = (1, 0, \dots, 0)^\top, \boldsymbol{\beta}_3 = (0, \dots, 0, 1, -1)^\top$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\beta}_3^\top)^\top$. gSIR and gDR are performed as if there were 3 groups of predictors. To evaluate the groupwise estimates, we first obtain $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1 \oplus \hat{\boldsymbol{\beta}}_2 \oplus \hat{\boldsymbol{\beta}}_3)$, then report the vector correlation coefficient between $\boldsymbol{\beta}$ and its projection onto $\text{span}(\hat{\boldsymbol{\beta}})$, which is given by $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}^\top \hat{\boldsymbol{\beta}})^{-1} \hat{\boldsymbol{\beta}}^\top \boldsymbol{\beta}$. The performance of gSIR and gDR are compared with ordinary SIR and DR estimates.

Table 3.4 summarizes the simulation results for Model (3.8). As expected, the performance of gSIR and gDR are slightly worse than those of SIR and DR for small n . However, as n increases, the difference diminishes and becomes negligible for $n \geq 800$.

These results suggest that the proposed method can still achieve satisfactory performance when group structure is incorrectly imposed in the situation where no real group structure is present.

3.4.5 Partial Dimension Reduction

At last we consider the situation of partial dimension reduction, in which only parts of the data need to be reduced and the others would like to be kept intact. The proposed envelope method can be easily modified to accommodate partial dimension reduction by fixing \mathbf{b}_g to an identity matrix for the group to be protected from reduction in the estimation procedure. In this example, we consider the following model:

$$Y = V_1 + \boldsymbol{\beta}^T \mathbf{V}_2 + V_1 \times (\boldsymbol{\beta}^T \mathbf{V}_2) + 0.5\epsilon, \quad (3.9)$$

where $\mathbf{X} = (V_1, \mathbf{V}_2^T)^T$ is multivariate normal distributed with mean zero and compound symmetry covariance matrix of variance one and covariance 0.5. V_1 is a scalar and \mathbf{V}_2 is a 10 dimensional random vector. ϵ is a standard normal error. We set $\boldsymbol{\beta}$ as $(1, -1, 0, \dots, 0)^T$, and vary n as $n = 100, 200, 400, 800, 1200$.

Table 3.5 reports the results. gSIR and gDR clearly outperform aSIR and aDR. This result probably comes from the fact that incorporating the information in V_1 helps with estimating $\boldsymbol{\beta}$. This example demonstrates the usefulness of the proposed method in partial dimension reduction.

3.4.6 Dimension Estimation

In this section, we evaluate the BIC-type criterion given in (3.4) for dimension estimation using the gSIR method as an illustration for Models (3.5, 3.6, 3.7). For this evaluation, C_n

Table 3.5: Simulation results for Model (3.9). Median and median absolute deviation (in parentheses) of vector correlation coefficient over 100 replications are shown.

Method	$n = 100$	$n = 200$	$n = 400$	$n = 800$	$n = 1200$
gSIR	0.924 (0.051)	0.973 (0.014)	0.990 (0.005)	0.994 (0.003)	0.997 (0.002)
aSIR	0.911 (0.044)	0.951 (0.022)	0.964 (0.012)	0.972 (0.005)	0.975 (0.003)
gDR	0.913 (0.060)	0.961 (0.020)	0.985 (0.006)	0.993 (0.005)	0.995 (0.002)
aDR	0.877 (0.080)	0.947 (0.027)	0.974 (0.011)	0.985 (0.006)	0.987 (0.003)

Table 3.6: Proportions of selecting the correct set of dimensions within the top 1, 2 and 3 choices using the BIC-type dimension estimation criterion with gSIR for Models (3.5, 3.6, 3.7) over 100 replications.

		$n=100$	$n=200$	$n=400$	$n=800$	$n=1200$
Model (3.5)	Top 1	0.29	0.38	0.57	0.59	0.59
	Top 2	0.49	0.62	0.81	0.85	0.92
	Top 3	0.62	0.72	0.90	0.97	0.97
Model (3.6)	Top 1	0.67	0.93	0.98	0.99	1.00
	Top 2	0.90	0.99	1.00	1.00	1.00
	Top 3	0.99	1.00	1.00	1.00	1.00
Model (3.7)	Top 1	0.56	0.60	0.65	0.63	0.69
	Top 2	0.76	0.84	0.88	0.89	0.95
	Top 3	0.87	0.91	0.92	0.97	0.98

is chosen as $0.05\sqrt{L(1, 1, \dots, 1)}n^{-1/2}$ based on its numerical performance. The proportions of selecting the correct set of dimensions within the top one, two and three choices over 100 replications are reported in Table 3.6. These results show that the performance of this BIC-type method depend on the complexity of the model. In general, the true dimensions can be selected within the top three choices for large n . However, when the coefficients are relatively difficult to be estimated accurately (e.g. β_3 in Model (3.5)), the true set of dimensions are also difficult to be selected as the top choice. This BIC-type criterion for the gDR method was not evaluated due to the long computational time. In addition, further work is required in order to determine the optimal choice of C_n .

3.5 Application

Next we illustrate the application of the proposed method in the analysis of the temperature and proxy dataset (Mann et al., 2008) discussed in Section 3.1. This comprehensive dataset contains 1,209 yearly resolved climate proxies with some dated back to B.C. and instrumental temperature record anomalies for the Northern Hemisphere for years 1850-1998 A.D. Usually one important objective of analyzing climate data is to build reconstructions of past temperatures for a long time, in order to study the pattern of climate change. To simplify the problem, in this analysis we only include proxies that have less than 10% missing data from year 999-1998 A.D., and two proxies are further excluded due to numerical issues. The resulting 116 proxies belong to 6 distinct groups, including tree composites/reconstructions, lake sediments, various composites or reconstructions or historical records, cave deposits, ice cores and tree rings.

For illustration purposes, we apply the gSIR method to this dataset and perform a naïve analysis. We first reduce the dimensionality of the proxies using the gSIR method for data from year 1850 to 1998 A.D. The working dimension is chosen as one for each group, in other words, we extract information from each group with a single linear combination and consider them as constructed covariates. Then a linear regression is fit using these six constructed covariates. Table 3.7 presents the corresponding p -values for all six groups. Compared to traditional dimension reduction method such as PCA, the proposed groupwise dimension reduction method enables one to draw conclusions regarding each individual group. The results show that proxies from lake sediments, cave deposits, ice cores and tree rings have significant effects in predicting temperatures. The adjusted R-squared for this linear regression is 0.6188, while the adjusted R-squared for the linear regression using all 116 proxies is 0.6865. Therefore, in terms of adjusted R-squared, lit-

Table 3.7: Linear regression p values for the temperature-proxy dataset.

Group	p -value
Tree composites/reconstructions	0.1577
Lake sediments	0.0144
Various composites or reconstructions or historical records	0.2465
Cave deposits	8.73e-13
Ice cores	6.15e-08
Tree rings	0.0230

the information is lost by reducing the proxy dimension from 116 to 6 using the gSIR method.

In order to assess the predictive performance of the proposed method, we perform 100 replications of 5 fold cross-validation using the above described dimension reduction - linear regression procedure. For the dimension reduction step, gSIR, aSIR, PCA with six principle components, and assembled PCA (aPCA) are compared, where gSIR, aSIR and aPCA all use one as the working dimension for each group. Therefore, all four dimension reduction methods provide six constructed covariates for the linear regression step. Average residual mean squared errors (RMSE) on the testing dataset are plotted in Figure 3.1. It is seen that gSIR clearly outperforms the other three methods. These results suggest that the proposed groupwise dimension reduction method could be useful in terms of improving both estimation accuracy and interpretability in high dimensional data analysis.

3.6 Discussion

In this chapter, we have proposed a groupwise sufficient dimension reduction method which preserves the group structure in the predictors and recovers full regression informa-

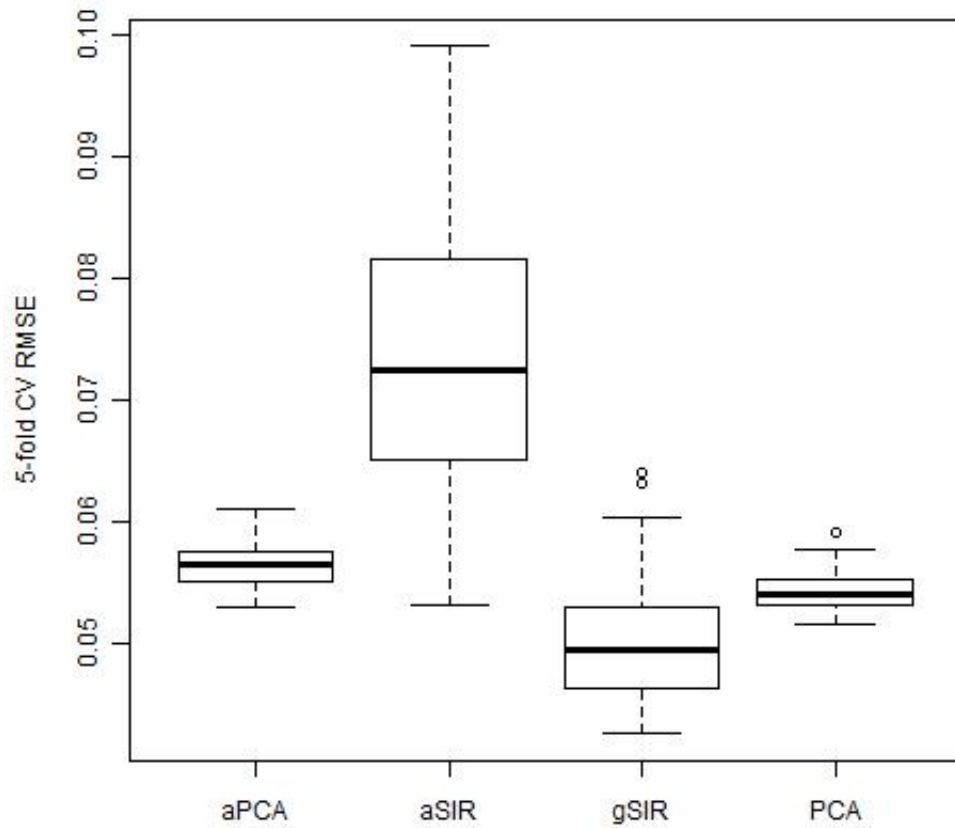


Figure 3.1: Cross-validation errors for the temperature dataset.

tion in the conditional distribution of $Y|\mathbf{X}$. Incorporating such prior group information in dimension reduction procedures provides more interpretable results, and more accurate estimates. The proposed method imposes the group structure to any classical estimator of $\mathcal{S}_{Y|\mathbf{X}}$ that based on certain random vectors or matrices via a direct sum envelope. As such, the smallest subspace that both covers the column space spanned by the random matrix and carries the group structure is estimated, which leads to the estimate of the groupwise central subspace. Compared to the assembled method, our proposal estimates the directions of different groups simultaneously. In addition, compared to the gMAVE method, the proposed method leads to the estimates of the more general groupwise central subspace, instead of groupwise central mean subspace, and thus is applicable to a wider range of situations. Simulation studies and real data analysis have shown that the proposed method achieves satisfactory performance in finite samples under various situations, including cases where there is no true group structure in the predictors, and where only partial dimension reduction is desired. For dimension estimation, we have suggested a BIC-type method and tested its empirical performance. However, its theoretical properties and the optimal choice of the scaling factor C_n still requires further research.

Chapter 4

Variable Selection with the Kernel Machine Cox Proportional Hazards Model for Optimal Treatment Strategy

4.1 Introduction

In clinical studies, optimal treatment strategies are sets of rules for making subject-specific treatment decisions that are determined based on each individual's various characteristics, such as genetic, environmental and behavioral characteristics and so on, such that the long-term clinical outcome is optimized. During the past few decades, many studies have been focused on obtaining such optimal treatment strategies. Rubin (1974, 1978) studied the causal effects of single point treatment in both randomized and observational studies using the potential outcome model. Murphy (2003) studied methods

for estimating dynamic treatment regimes, which are a series of treatment decision rules based on time-varying covariates and are adaptive to the status change across time for an individual. Some other work on optimal treatment strategies include Robins (2004), Murphy (2005), Zhao et al. (2009) and Brinkley et al. (2010). Many of the studies are focused on simple clinical outcomes such as CD4 count in HIV studies, and the objective of optimal treatment strategy is usually to maximize the potential outcomes. However, in many applications, instead of simple clinical outcomes, censored survival time is of primary interest. Few studies are targeted to optimal treatment decisions in the survival framework. One example is Zhao et al. (2009). In this chapter, we aim to estimate the optimal treatment strategy in the context of survival data.

To estimate the optimal treatment strategy, one usually needs to adjust the treatment and treatment-covariate interaction effect on survival times with the baseline covariate effect. However, the true function form of the baseline covariate effect is often complicated and nonlinear. The kernel machine method is a powerful and convenient tool for function approximation and incorporating complex feature spaces, especially for high-dimensional data (Vapnik, 1998; Schölkopf and Smola, 2002; Suykens, 2002). Support vector machine (SVM) (Vapnik, 1998) is a popular example of the kernel machine method. Recently, linear, logistic and Cox proportional hazards kernel machine models have been proposed to model genetic pathway effects, and corresponding testing methods are developed (Liu et al., 2007, 2008; Cai et al., 2010). However, little research has been done to develop method that apply kernel machine to survival data for optimal treatment strategies.

Another challenge for estimating optimal treatment strategies arises from the high dimensionality of the covariates. As technology advances, there are many clinical, genetic and environmental information available for making treatment decisions. Nonetheless, it is often true that only a subset of variables are relevant to the treatment assignment,

and including redundant information in the decision rules usually yields unstable and high variant estimator. As such variable selection becomes an important component of the analysis. In the context of optimal treatment decisions, Qian and Murphy (2011) applied L_1 penalized least squares in estimating optimal treatment rules and performing variable selection in the Q-learning framework. Lu et al. (2011) proposed a loss based method in the form of A-learning and applied shrinkage penalties to estimate optimal treatment strategy. Incorporating variable selection in the estimation has been shown to often provide better estimated treatment rules.

In this chapter, we focus on estimating the optimal treatment strategy in the context of survival data. Let $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^p$ denote the p -dimensional baseline covariates, $A \in \mathcal{A}$ denote the treatment decision for a patient, then a treatment strategy $g(\mathbf{X}) \in \mathcal{G}$ is a mapping from \mathcal{X} to \mathcal{A} , where \mathcal{G} denote the decision rule space that contains all possible treatment rules. For survival data, the optimal treatment strategy g^{opt} can be defined in two ways: the treatment strategy that maximizes the expected survival time, and the one that maximizes the survival probability at a given time point t_0 . We propose a two-step procedure which aims to estimate the optimal treatment strategy for survival data. The optimal treatment strategy obtained from the proposed model satisfies both definitions given above. In the first step, a kernel machine Cox proportional hazards model is applied to obtain an initial estimate of the optimal treatment rule; in the second step, shrinkage penalties are applied to incorporate variable selection in the estimation. By using the kernel machine method, the proposed procedure does not require explicit specification of the baseline covariate effect function and therefore is very flexible. In addition, incorporating shrinkage penalties in the estimation leads to a more parsimonious model and effectively reduces noises in the estimated decision rule. The rest of this chapter is organized as follows. In Section 4.2 we first briefly review the potential outcome model,

then describe the kernel machine Cox proportional hazards model, its connection to the mixed effects Cox model, and the variable selection step. In Section 4.3, simulation studies are performed to examine the empirical performance of the proposed method. Section 4.4 concludes the chapter with some discussions.

4.2 Model and Method

4.2.1 Potential Outcome

For simplicity, in this chapter we focus our discussion on two treatment groups such that the treatment $a \in \mathcal{A} = \{-1, 1\}$, where -1 represents the control and 1 represents the treatment. The discussion can however be easily extended to multiple treatment groups, as discussed in Section 4.4. Following Rubin (1974) and Rubin (1978), we define the potential survival time (potential outcome) $T^*(a)$ as the survival time that a patient would have if assigned treatment $a \in \mathcal{A}$, and assume the following two standard assumptions:

- A1: The stable unit treatment value assumption,

$$T = T^*(1)I(A = 1) + T^*(-1)I(A = -1).$$

That is, the actual survival time of an individual who received treatment $A = a$ is the same as the potential survival time $T^*(a)$, in all conditions.

- A2: The strong ignorability assumption, $A \perp\!\!\!\perp \{T^*(-1), T^*(1)\} | \mathbf{X}$. That is, The treatment assignment A is independent of the potential survival times given covariates \mathbf{X} . This assumption requires that we have full information of confounding variables. Note that for randomized studies, this assumption holds naturally. In fact,

it is reasonable to assume $A \perp\!\!\!\perp \{T^*(-1), T^*(1)\}$ in randomized studies. However, for observational studies, one should be cautious when making this assumption.

Denote $g(\mathbf{X}) \in \mathcal{G}$ as the treatment rule given the covariates \mathbf{X} , where \mathcal{G} is the set of all possible treatment regimes. Based on the above two assumptions, it follows that (Lu et al., 2011)

$$E[T^*\{g(\mathbf{X})\}] = E_{\mathbf{X}}[E(T|A = 1, \mathbf{X})I\{g(\mathbf{X}) = 1\} + E(T|A = -1, \mathbf{X})I\{g(\mathbf{X}) = -1\}]. \quad (4.1)$$

According to (4.1), the optimal treatment rule defined as the rule that maximizes expected survival time, i.e., $g^{opt}(\mathbf{X}) = \operatorname{argmax}_{g \in \mathcal{G}} E[T^*\{g(\mathbf{X})\}]$, is then given by

$$g^{opt}(\mathbf{X}) = \operatorname{sign}\{E(T|A = 1, \mathbf{X}) - E(T|A = -1, \mathbf{X})\}. \quad (4.2)$$

We note that when $E(T|A = 1, \mathbf{X}) = E(T|A = -1, \mathbf{X})$, the treatment rule in (4.2) is not well defined. Since $E(T|A = 1, \mathbf{X}) = E(T|A = -1, \mathbf{X})$ indicates that these two treatments are equally good for the subject, we randomly assign the subject to one of these two treatments in this situation, and adopt this rule throughout this chapter.

Similarly, for any given time point t_0 we have

$$\begin{aligned} P[T^*\{g(\mathbf{X})\} > t_0] &= E_{\mathbf{X}}[P(T > t_0|A = 1, \mathbf{X})I\{g(\mathbf{X}) = 1\} \\ &\quad + P(T > t_0|A = -1, \mathbf{X})I\{g(\mathbf{X}) = -1\}]. \end{aligned} \quad (4.3)$$

PROOF. We first note that for any treatment regime $g \in \mathcal{G}$, the potential survival time is given by

$$T^*\{g(\mathbf{X})\} = T^*(1)I\{g(\mathbf{X}) = 1\} + T^*(-1)I\{g(\mathbf{X}) = -1\}. \quad (4.4)$$

In addition, for $a \in \mathcal{A} = \{-1, 1\}$ we have

$$\begin{aligned} P(T > t_0 | A = a, \mathbf{X}) &= E\{I(T > t_0) | A = a, \mathbf{X}\} \\ &= E[I\{T^*(1)I(A = 1) + T^*(-1)I(A = -1) > t_0\} | A = a, \mathbf{X}] \\ &= E[I\{T^*(1) > t_0\}I(A = 1) \\ &\quad + I\{T^*(-1) > t_0\}I(A = -1) | A = a, \mathbf{X}] \\ &= E[I\{T^*(a) > t_0\} | A = a, \mathbf{X}] \\ &= E[I\{T^*(a) > t_0\} | \mathbf{X}] \\ &= P[T^*(a) > t_0 | \mathbf{X}], \end{aligned} \quad (4.5)$$

where the second equation holds by A1, and the second last equation holds by A2. Then we have

$$\begin{aligned}
& P\left[T^*\{g(\mathbf{X})\} > t_0\right] \\
&= E\left(I[T^*\{g(\mathbf{X})\} > t_0]\right) \\
&= E_{\mathbf{X}}\left\{E\left(I[T^*\{g(\mathbf{X})\} > t_0]|\mathbf{X}\right)\right\} \\
&= E_{\mathbf{X}}\left\{E\left(I[T^*(1)I\{g(\mathbf{X}) = 1\} + T^*(-1)I\{g(\mathbf{X}) = -1\}] > t_0|\mathbf{X}\right)\right\} \\
&= E_{\mathbf{X}}\left\{E\left(I\{T^*(1) > t_0\}I\{g(\mathbf{X}) = 1\} + I\{T^*(-1) > t_0\}I\{g(\mathbf{X}) = -1\}|\mathbf{X}\right)\right\} \\
&= E_{\mathbf{X}}\left\{E\left[I\{T^*(1) > t_0\}|\mathbf{X}\right]I\{g(\mathbf{X}) = 1\} + E\left[I\{T^*(-1) > t_0\}|\mathbf{X}\right]I\{g(\mathbf{X}) = -1\}\right\} \\
&= E_{\mathbf{X}}\left\{P[T^*(1) > t_0|\mathbf{X}]I\{g(\mathbf{X}) = 1\} + P[T^*(-1) > t_0|\mathbf{X}]I\{g(\mathbf{X}) = -1\}\right\} \\
&= E_{\mathbf{X}}\left\{P(T > t_0|A = 1, \mathbf{X})I\{g(\mathbf{X}) = 1\} + P(T > t_0|A = -1, \mathbf{X})I\{g(\mathbf{X}) = -1\}\right\},
\end{aligned}$$

where the third equation follows (4.4) and the last equation follows (4.5). \square

Based on (4.3), the optimal treatment strategy defined as the rule that maximizes the survival probability at a given time point t_0 , i.e., $g^{opt}(\mathbf{X}) = \operatorname{argmax}_{g \in \mathcal{G}} P[T^*\{g(\mathbf{X})\} > t_0]$, is given by

$$g^{opt}(\mathbf{X}) = \operatorname{sign}\{P(T > t_0|A = 1, \mathbf{X}) - P(T > t_0|A = -1, \mathbf{X})\}. \quad (4.6)$$

4.2.2 Cox Proportional Hazards Model for Optimal Treatment Strategy

We next discuss the Cox proportional hazards model that leads us to the solution of $g^{opt}(\mathbf{X})$. Considering n random subjects, let T_i be the failure time, C_i be the censor-

ing time, $A_i \in \mathcal{A} = \{-1, 1\}$ be the treatment, and $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$ be the p -dimensional covariates of the i -th subject, for $i = 1, \dots, n$. Define the observed event time $\tilde{T}_i = \min(T_i, C_i)$ and the censoring indicator $\delta_i = I(T_i \leq C_i)$. Then the observed data consist of $\{(\tilde{T}_i, \delta_i, A_i, \mathbf{X}_i), i = 1, \dots, n\}$. Furthermore, we assume conditional independent censoring, i.e., $T \perp\!\!\!\perp C | \mathbf{X}$.

We propose the use of a semiparametric Cox proportional hazards model that relates the survival time T to treatment A and covariates \mathbf{X} through a hazard function $\lambda(t|A, \mathbf{X})$:

$$\lambda(t|A, \mathbf{X}) = \lambda_0(t) \exp\{h(\mathbf{X}) - Af(\mathbf{X})\}, \quad (4.7)$$

where $\lambda_0(t)$ is the unspecified baseline hazard function, $h(\mathbf{X})$ is an unknown centered smooth function of \mathbf{X} corresponding to the baseline effect of \mathbf{X} , and $f(\mathbf{X})$ is a parametric function of \mathbf{X} . $Af(\mathbf{X})$ corresponds to the combination of treatment effect and the treatment-covariate interaction effects. Here the interest of our study is to make the correct treatment decision A for an individual with covariates \mathbf{X} , such that both the expected survival time and the survival probability at a given time t_0 are maximized. According to Model (4.7), it is easy to see that

$$\frac{\lambda(t|A = 1, \mathbf{X})}{\lambda(t|A = -1, \mathbf{X})} = \exp\{-2f(\mathbf{X})\}.$$

If $f(\mathbf{X}) > 0$, we have $\lambda(t|A = 1, \mathbf{X}) < \lambda(t|A = -1, \mathbf{X})$ for any $t > 0$. Consequently, the survival functions $S(t) = Pr(T > t)$ have the following relationship:

$$\begin{aligned} S(t|A = 1, \mathbf{X}) &= \exp\left\{-\int_0^t \lambda(u|A = 1, \mathbf{X}) du\right\} > \\ \exp\left\{-\int_0^t \lambda(u|A = -1, \mathbf{X}) du\right\} &= S(t|A = -1, \mathbf{X}) \end{aligned}$$

for any $t > 0$. It is easily seen then $E(T|A = 1, \mathbf{X}) > E(T|A = -1, \mathbf{X})$. Similarly, one can show that if $f(\mathbf{X}) < 0$, $S(t|A = 1, \mathbf{X}) < S(t|A = -1, \mathbf{X})$ and $E(T|A = 1, \mathbf{X}) < E(T|A = -1, \mathbf{X})$. Therefore, we have

$$\begin{aligned} \text{sign}\{f(\mathbf{X})\} &= \text{sign}\{S(t_0|A = 1, \mathbf{X}) - S(t_0|A = -1, \mathbf{X})\} \\ &= \text{sign}\{E(T|A = 1, \mathbf{X}) - E(T|A = -1, \mathbf{X})\} \end{aligned} \quad (4.8)$$

Combining (4.8) with (4.2) and (4.6), the optimal treatment strategy according to Model (4.7) is given by $g^{opt}(\mathbf{X}) = \text{sign}\{f(\mathbf{X})\}$, for both definitions. To simplify our problem, we fix $f(\mathbf{X})$ in its linear form throughout this chapter, i.e. $f(\mathbf{X}) = \boldsymbol{\beta}^\top \tilde{\mathbf{X}}$, where $\tilde{\mathbf{X}} = (1, \mathbf{X}^\top)^\top$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{p+1})^\top$ is the vector of regression coefficients for the treatment and treatment-covariate interaction term $A\tilde{\mathbf{X}}$. The model then becomes

$$\lambda(t|A, \mathbf{X}) = \lambda_0(t) \exp\{h(\mathbf{X}) - A(\boldsymbol{\beta}^\top \tilde{\mathbf{X}})\}. \quad (4.9)$$

Based on this model, the optimal treatment strategy is $g^{opt}(\mathbf{X}) = \text{sign}(\boldsymbol{\beta}^\top \tilde{\mathbf{X}})$.

4.2.3 The Kernel Machine Method

Model (4.9) allows nonparametric adjustment of the baseline covariates effect that not related to the treatment through the unknown function $h(\cdot)$, and therefore greatly improves the flexibility of the model. Note that if $h(\mathbf{X}) = \boldsymbol{\eta}^\top \mathbf{X}$, the model is reduced to the standard linear Cox model with interaction terms. We follow the general nonparametric modeling setting and assume that $h(\cdot)$ lies in a certain function space \mathcal{H}_K , which is generated by a given positive definite kernel function $K(\cdot, \cdot; \rho)$, where ρ are possible kernel parameters and usually unknown. One popular kernel function is the d th polynomial kernel

$K(\mathbf{X}_1, \mathbf{X}_2) = (\mathbf{X}_1^\top \mathbf{X}_2 + \rho)^d$, where ρ is the intercept and d is the polynomial degree. Another popular kernel function is the Gaussian kernel $K(\mathbf{X}_1, \mathbf{X}_2) = \exp\{-\|\mathbf{X}_1 - \mathbf{X}_2\|^2/\rho\}$, where $\|\cdot\|$ represents the Euclidean norm and $\rho > 0$ is a scale parameter. The function space generated by the Gaussian kernel is spanned by radial basis functions. More mathematical properties about this kernel function can be found in Bühmann (2003).

Under certain regularity conditions, from the Mercer's Theorem (Cristianini and Shawe-Taylor, 2000), a given kernel function corresponds to a specific function space spanned by a set of orthogonal basis functions $\{\phi_j(\mathbf{X})\}_{j=1}^J$. This leads to the primal representation of any $h(\mathbf{X}) \in \mathcal{H}_K$ as $h(\mathbf{X}) = \sum_{j=1}^J w_j \phi_j(\mathbf{X})$. However, it is generally difficult to explicitly specify the form of the basis functions, and an equivalent dual representation is usually employed. The dual representation allows any unknown function $h(\mathbf{X})$ in \mathcal{H}_K be written as a linear combination of the kernel function evaluated at each observed data point, as we describe in the following.

Estimation of $\{\boldsymbol{\beta}, h(\cdot)\}$ is achieved by maximizing the penalized log partial likelihood (PPL) function given as

$$PPL = \sum_{i=1}^n \delta_i \left[h(\mathbf{X}_i) - A_i(\boldsymbol{\beta}^\top \tilde{\mathbf{X}}_i) - \log \left\{ \sum_{j=1}^n \exp\{h(\mathbf{X}_j) - A_j(\boldsymbol{\beta}^\top \tilde{\mathbf{X}}_j)\} I(\tilde{T}_j \geq \tilde{T}_i) \right\} \right] - \frac{\alpha}{2} \|h\|_{\mathcal{H}_K}^2 \quad (4.10)$$

By the Representer Theorem (Kimeldorf and Wahba, 1971), the solution of $h(\cdot) \in \mathcal{H}_K$ for (4.10) can be represented as

$$h(\mathbf{X}_i) = \sum_{j=1}^n \gamma_j K(\mathbf{X}_i, \mathbf{X}_j; \rho) = \boldsymbol{\gamma}^\top K_i(\rho), \quad (4.11)$$

where $K_i(\rho) = (K(\mathbf{X}_i, \mathbf{X}_1; \rho), \dots, K(\mathbf{X}_i, \mathbf{X}_n; \rho))^\top$, and $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_n)^\top$ is an n dimen-

sional unknown parameter. This corresponds to the dual representation. Denote $\mathbb{K}(\rho)$ as the $n \times n$ matrix with the (i, j) th element being $K_{ij}(\rho) = K(\mathbf{X}_i, \mathbf{X}_j, \rho)$, then $\mathbf{h} = (h(\mathbf{X}_1), \dots, h(\mathbf{X}_n))^\top = \mathbb{K}(\rho)\boldsymbol{\gamma}$. Substituting (4.11) into (4.10) gives

$$\begin{aligned} PPL &= \sum_{i=1}^n \delta_i [\boldsymbol{\gamma}^\top K_i(\rho) - A_i(\boldsymbol{\beta}^\top \tilde{\mathbf{X}}_i)] \\ &\quad - \log \left\{ \sum_{j=1}^n \exp(\boldsymbol{\gamma}^\top K_j(\rho) - A_j(\boldsymbol{\beta}^\top \tilde{\mathbf{X}}_j)) I(\tilde{T}_j \geq \tilde{T}_i) \right\} - \frac{\alpha}{2} \boldsymbol{\gamma}^\top \mathbb{K}(\rho) \boldsymbol{\gamma} \end{aligned} \quad (4.12)$$

Note that there are two tuning parameters α and ρ in the above PPL function. The penalty parameter α governs the tradeoff between the complexity and the fit of the model. Specifically, $\alpha = 0$ leads to a saturated model, and $\alpha = \infty$ reduces the model to a standard linear Cox model with only the treatment effect and the treatment-covariate interactions $A\tilde{\mathbf{X}}$. ρ is a kernel parameter and often controls the smoothness property of the kernel function. For example, for the Gaussian kernel, $\rho \rightarrow 0$ corresponds to no similarities among individuals, while $\rho = \infty$ assumes all individuals are the same.

Given α and ρ , $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be estimated alternatively by maximizing (4.12) via the Newton-Raphson iterative algorithm. The detailed estimation procedure is given as follows:

Step 1. Set $k = 0$. Obtain initial estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

$$\boldsymbol{\gamma}^{[0]} = (0, \dots, 0)^\top.$$

$$\boldsymbol{\beta}^{[0]} = \text{maximum likelihood estimate of the partial likelihood function with } \boldsymbol{\gamma} = \mathbf{0}.$$

Step 2. Given $\boldsymbol{\beta}^{[k]}$, update the estimate of $\boldsymbol{\gamma}$ as

$$\boldsymbol{\gamma}^{[k+1]} = \boldsymbol{\gamma}^{[k]} - H^{-1}(\boldsymbol{\gamma}^{[k]})U(\boldsymbol{\gamma}^{[k]}),$$

where $U(\boldsymbol{\gamma}^{[k]}) = \frac{\partial PPL(\boldsymbol{\beta}^{[k]}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{[k]}}$ and $H(\boldsymbol{\gamma}^{[k]}) = \frac{\partial^2 PPL(\boldsymbol{\beta}^{[k]}, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^\top}|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{[k]}}$.

$\boldsymbol{\gamma}^{[k+1]}$ is then further adjusted by centering $\mathbb{K}(\rho)\boldsymbol{\gamma}^{[k+1]}$.

Step 3. Given $\boldsymbol{\gamma}^{[k+1]}$, estimate $\boldsymbol{\beta}$. In R, $\boldsymbol{\beta}^{[k+1]}$ can be obtained by treating $\mathbb{K}(\rho)\boldsymbol{\gamma}^{[k+1]}$ as an offset term, and then maximizing the partial likelihood function with respect to $\boldsymbol{\beta}$ through the `coxph` function.

Step 4. Set $k = k + 1$. Go back to Step 2 until convergence.

We propose to tune the tuning parameters $\{\alpha, \rho\}$ via grid search either through cross-validation or through an independent testing dataset, i.e., choose the combination of $\{\alpha, \rho\}$ that maximize the partial likelihood function of the testing dataset. Alternatively, one can estimate $\{\boldsymbol{\beta}, \mathbf{h}, \alpha, \rho\}$ simultaneously through the connection of the proposed kernel machine Cox model to the mixed effects Cox model as described in the next section.

4.2.4 Connection to the Mixed Effects Cox Model

Estimators of $\{\boldsymbol{\beta}, h(\cdot)\}$ in (4.12) correspond to the estimator of a mixed effects Cox model using the Laplace approximation (Liu et al., 2007, 2008; Cai et al., 2010). Similar derivations can be found in Ripatti and Palmgren (2000), Therneau (2003) and Pankratz et al. (2005). Consider the following mixed effects Cox model:

$$\lambda_i(t) = \lambda_0(t) \exp\{h_i - A_i(\boldsymbol{\beta}^\top \tilde{\mathbf{X}}_i)\}, \quad (4.13)$$

where $\mathbf{h} = (h_1, \dots, h_n)^\top$ is a $n \times 1$ vector of subject-specific random effects following $\mathbf{h} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \tau^2 \mathbb{K}(\rho) = \frac{1}{\alpha} \mathbb{K}(\rho)$. Let $\ell_n(\boldsymbol{\beta}, \mathbf{h})$ denote the log partial likelihood

function conditional on \mathbf{h} . The mixed effects model leads to the observed data partial likelihood function:

$$\exp(L) = (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \int \exp\{\ell_n(\boldsymbol{\beta}, \mathbf{h}) - \frac{1}{2} \mathbf{h}^\top \boldsymbol{\Sigma}^{-1} \mathbf{h}\} d\mathbf{h} \quad (4.14)$$

Setting $\mathbf{h} = \mathbb{K}(\rho)\boldsymbol{\gamma}$, one can easily see that

$$PPL(\boldsymbol{\beta}, \mathbf{h}) = \ell_n(\boldsymbol{\beta}, \mathbf{h}) - \frac{1}{2} \mathbf{h}^\top \boldsymbol{\Sigma}^{-1} \mathbf{h}. \quad (4.15)$$

Combining (4.15) with (4.14) gives:

$$\begin{aligned} \exp(L) &= (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \int \exp\{PPL(\boldsymbol{\beta}, \mathbf{h})\} d\mathbf{h} \\ &\approx (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \int \exp\{PPL(\boldsymbol{\beta}, \hat{\mathbf{h}}) - \frac{1}{2} (\mathbf{h} - \hat{\mathbf{h}})^\top (-H_{\hat{\mathbf{h}}\hat{\mathbf{h}}}) (\mathbf{h} - \hat{\mathbf{h}})\} d\mathbf{h} \\ &= |\boldsymbol{\Sigma}|^{-1/2} | -H_{\hat{\mathbf{h}}\hat{\mathbf{h}}} |^{-1/2} \exp\{PPL(\boldsymbol{\beta}, \hat{\mathbf{h}})\} \\ &\quad \int (2\pi)^{-n/2} | -H_{\hat{\mathbf{h}}\hat{\mathbf{h}}}^{-1} |^{-1/2} \exp\{-\frac{1}{2} (\mathbf{h} - \hat{\mathbf{h}})^\top (-H_{\hat{\mathbf{h}}\hat{\mathbf{h}}}) (\mathbf{h} - \hat{\mathbf{h}})\} d\mathbf{h} \\ &= |\boldsymbol{\Sigma}|^{-1/2} | -H_{\hat{\mathbf{h}}\hat{\mathbf{h}}} |^{-1/2} \exp\{PPL(\boldsymbol{\beta}, \hat{\mathbf{h}})\} \end{aligned}$$

where $\hat{\mathbf{h}}$ maximizes PPL and $H_{\hat{\mathbf{h}}\hat{\mathbf{h}}}$ is the second derivative of the PPL with respect to \mathbf{h} evaluated at $\hat{\mathbf{h}}$. Therefore, the integrated log partial likelihood function becomes

$$L = PPL(\boldsymbol{\beta}, \hat{\mathbf{h}}) - \frac{1}{2} (\log |\boldsymbol{\Sigma}| + \log | -H_{\hat{\mathbf{h}}\hat{\mathbf{h}}}|). \quad (4.16)$$

If we ignore the last two terms in (4.16), and maximize L with respect to $\boldsymbol{\beta}$, then $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{h}})$ jointly maximize $PPL(\boldsymbol{\beta}, \mathbf{h})$. Ripatti and Palmgren (2000) have suggested that the loss

of information by ignoring these two terms is slight. It follows that one can obtain the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{h}}$ of the kernel machine Cox model by fitting the mixed effects Cox model (4.13) using Laplace approximation, and the tuning parameters τ and ρ can be considered as variance parameters in a mixed effects Cox model.

4.2.5 Variable Selection with the Kernel Machine Cox Model

We propose to apply shrinkage penalties for variable selection in the kernel machine Cox model. In practice, there are many choices of the penalty terms, such as adaptive LASSO (Zou, 2006) and the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001). We employ the adaptive LASSO penalty here. Denote the solution that maximizes (4.12) as $\{\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}\}$, then the adaptive LASSO solution solves

$$\min_{\boldsymbol{\beta}} \left\{ -\frac{1}{n} \ell_n(\boldsymbol{\beta}, \tilde{\boldsymbol{\gamma}}) + \lambda \sum_{j=1}^{p+1} |\beta_j| / |\tilde{\beta}_j| \right\}, \quad (4.17)$$

where $\ell_n(\boldsymbol{\beta}, \tilde{\boldsymbol{\gamma}})$ is the log partial likelihood function given by:

$$\begin{aligned} \ell_n(\boldsymbol{\beta}, \tilde{\boldsymbol{\gamma}}) &= \sum_{i=1}^n \delta_i [\tilde{\boldsymbol{\gamma}}^\top K_i(\rho) - A_i(\boldsymbol{\beta}^\top \tilde{\mathbf{X}}_i)] \\ &\quad - \log \left\{ \sum_{j=1}^n \exp\{\tilde{\boldsymbol{\gamma}}^\top K_j(\rho) - A_j(\boldsymbol{\beta}^\top \tilde{\mathbf{X}}_j)\} I(\tilde{T}_j \geq \tilde{T}_i) \right\}. \end{aligned}$$

Denote the solution to (4.17) as $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_{p+1})^\top$. In this variable selection step, the tuning parameter λ can be tuned via regular methods such as cross-validation and BIC. In this chapter, we employ the BIC-type criterion and choose λ such that the objective function

$$-2\ell_n(\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}) + \log(n) d_\lambda$$

is minimized, where d_λ is the number of nonzero estimates in $\hat{\boldsymbol{\beta}}$ given λ .

4.3 Simulations

We perform simulation studies to investigate the empirical performance of the proposed method under various situations. The estimation accuracy, variable selection performance, and the ability to make correct treatment decisions are evaluated.

In all settings, \mathbf{X}_i for $i = 1, \dots, n$ are generated from a multivariate normal distribution with mean zero, variance one, and an order one autoregressive correlation structure, where $\text{corr}(X_{ij}, X_{ik}) = 0.5^{|j-k|}$, $j, k = 1, \dots, p$. A_i are generated independently with $Pr(A = 1) = 0.5$. The censoring time C_i 's are generated from a uniform $(0, C_0)$ distribution, where C_0 controls the censoring proportion. The survival time T_i are generated according to Model (4.9) with $\lambda_0(t) = 1$ and various forms of h . We consider three different function forms of h , including a linear form, a function with interaction terms and a complex nonlinear function, as follows:

- Case 1: $h(\mathbf{X}) = \boldsymbol{\eta}_1^\top \mathbf{X}$, where $\boldsymbol{\eta}_1 = (0.6, -0.9, 0.8, 0.9, 0, \dots, 0, 1)^\top$.
- Case 2: $h(\mathbf{X}) = 0.6(\boldsymbol{\eta}_2^\top \mathbf{X})(\boldsymbol{\eta}_3^\top \mathbf{X})$, where $\boldsymbol{\eta}_2 = (1, -1, 1, 0, \dots, 0)^\top$ and $\boldsymbol{\eta}_3 = (1, 0, \dots, 0, -1, 0, 1)^\top$.
- Case 3: $h(\mathbf{X}) = 2 \sin(\boldsymbol{\eta}_2^\top \mathbf{X}) + 0.3(1 + \boldsymbol{\eta}_3^\top \mathbf{X})^2$, where $\boldsymbol{\eta}_2$ and $\boldsymbol{\eta}_3$ are same as in case 2.

After generated according to the above functions, \mathbf{h} is centered in each replication. We set $\boldsymbol{\beta} = (0.8, 1, 0, \dots, 0, -0.9, 0.8)^\top$ in all three cases. As such, covariates in the baseline function and those in the interaction term are allowed to be different. We consider two censoring proportions: 20% and 40%, and vary the sample size as $n = 100, 200$. The

dimension for \mathbf{X} is set as $p = 10, 20, 30$ for 20% censoring and $p = 10, 20$ for 40% censoring. 100 replications are performed for each scenario. We apply the Gaussian kernel in estimating h in this simulation study.

We compare the proposed kernel machine Cox model (Kernel), the kernel machine Cox model with adaptive LASSO (aLASSO), and a linear model which assumes a linear form $\boldsymbol{\eta}^\top \mathbf{X}$ as the baseline function h (Linear). To complete the comparison, two oracle models are also included as benchmarks: Oracle1 applies true values of h in estimating $\boldsymbol{\beta}$; Oracle2 not only applies true values of h , but also assumes the true model for $\boldsymbol{\beta}$, i.e., only covariates with nonzero coefficients are included during estimation. For the kernel machine Cox model estimation, in each scenario a common testing dataset of sample size 500 is generated for selection of the tuning parameters, i.e., the optimal $\{\alpha, \rho\}$ are chosen such that the partial likelihood function of the testing dataset is maximized. The grids are chosen as $\alpha = (2^{-6}, 2^{-5}, \dots, 2^7)$, and $\rho = (\sigma_m^2/4, \sigma_m^2/2, \sigma_m^2, 2\sigma_m^2, 4\sigma_m^2)$, where $\sigma_m = \text{median}(\|\mathbf{X}_i - \mathbf{X}_j\|, i \neq j)$.

We evaluate and compare these five methods by two criteria. The first criterion is mean squared error (MSE), defined as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$, which measures the estimation accuracy. The second criterion is percentage of correct decisions (PCD), defined as $\sum_{i=1}^n I\{\text{sign}(\boldsymbol{\beta}^\top \tilde{\mathbf{X}}_i) = \text{sign}(\hat{\boldsymbol{\beta}}^\top \tilde{\mathbf{X}}_i)\}/n$, which evaluates the ability to make correct treatment decisions. Average MSE and PCD together with their corresponding standard errors are reported. For the kernel machine Cox model with adaptive LASSO method, we also report the following measures that summarize its variable selection performance: the average size of the selected models (Size), the frequency of selecting the exact model (Exact), the frequency of selecting all important predictors (Cover) out of 100 data replications, the average percentage of correct zeros being identified (Corr0), and the average percentage of incorrect zeros (Incorr0).

Table 4.1 to Table 4.6 summarize the MSE and PCD results. The performance of all methods improve as n increases, p decreases and as censoring rate decreases. In all cases, the kernel machine Cox model achieves better performance than the linear model, in terms of both MSE and PCD. In general, the improvements are further enhanced with the variable selection step. It is noted that although when p is small, in certain cases the penalized estimator gives a larger MSE than the unpenalized one (e.g. $n = 100, p = 10$ for Case 3), the PCD for the penalized estimator still improves. Table 4.7 summarizes the variable selection performance of the kernel machine Cox model with adaptive LASSO method. Overall, the procedure performs well in terms of variable selection. It gives a reasonable model size estimation, achieves a moderate to high frequency of selecting the exact model, and a high coverage percentage. These simulation results suggest that the proposed method is useful in making correct treatment decisions in various situations.

4.4 Discussion

In this chapter, we have proposed a two-step procedure that combines the kernel machine Cox proportional hazards model with shrinkage penalty based variable selection method for estimating optimal treatment strategy in the survival framework. The proposed procedure does not require explicit specification of the baseline covariate effect function and thus very flexible. Incorporating variable selection in the estimation also leads to a more parsimonious model and usually gives better decision rules. Our simulation studies have shown that the propose method works well in various situations, and is a promising method in estimating optimal treatment strategies. In this chapter, we have limited our discussion to two treatment groups for simplicity. However, the proposed method can be easily extended to multiple treatment groups. For example, assume the possible

treatment set is given by $\mathcal{A} = \{1, 2, \dots, K\}$, Model (4.9) can be modified accordingly as

$$\begin{aligned} \lambda(t|A, \mathbf{X}) = \lambda_0(t) \exp\{h(\mathbf{X}) - I(A = 1)(\boldsymbol{\beta}_1^\top \tilde{\mathbf{X}}) - I(A = 2)(\boldsymbol{\beta}_2^\top \tilde{\mathbf{X}}) - \\ \dots - I(A = K - 1)(\boldsymbol{\beta}_{K-1}^\top \tilde{\mathbf{X}})\}. \end{aligned} \quad (4.18)$$

The optimal treatment strategy is then given by

$$g^{opt}(\mathbf{X}) = \operatorname{argmax}_{k \in \mathcal{A}} \{\boldsymbol{\beta}_k^\top \tilde{\mathbf{X}}\},$$

where $\boldsymbol{\beta}_K = \mathbf{0}$. In other words, in this multiple treatment setting, the optimal treatment is selected as the one that corresponds to the smallest hazard function. The variable selection step can then be applied accordingly by applying shrinkage penalties on $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \dots, \boldsymbol{\beta}_{K-1}^\top)^\top$. Alternatively, if the goal for variable selection is to determine the treatment rule based on a common subset of covariates for all possible treatments, one can apply the group LASSO type penalty (Yuan and Lin, 2006), and define $p + 1$ groups such that $\{\beta_{1j}, \beta_{2j}, \dots, \beta_{(K-1)j}\}, j = 1, \dots, p + 1$ are in one group. Another interesting extension of the proposed method is to apply the method for situations where multiple decision rules at different time points are required throughout the study, which corresponds to the studies for dynamic treatment regimes. In addition, it would be useful to study the selection of the kernel function for the proposed method. These problems are certainly of interest for future studies.

Table 4.1: Simulation results for Case 1 with 20% censoring. Results are averaged over 100 replications. Numbers in parentheses are standard errors.

		$n=100$			$n=200$		
		$p=10$	$p=20$	$p=30$	$p=10$	$p=20$	$p=30$
MSE	Oracle1	0.324 (0.016)	0.911 (0.043)	1.870 (0.082)	0.129 (0.007)	0.313 (0.013)	0.538 (0.022)
	Oracle2	0.083 (0.006)	0.100 (0.008)	0.098 (0.008)	0.038 (0.003)	0.038 (0.003)	0.035 (0.003)
	Linear	0.667 (0.039)	4.071 (0.235)	29.898 (6.112)	0.184 (0.012)	0.639 (0.026)	1.394 (0.054)
	Kernel	0.377 (0.018)	1.118 (0.044)	2.199 (0.108)	0.137 (0.007)	0.346 (0.013)	0.643 (0.024)
	aLASSO	0.253 (0.018)	0.554 (0.044)	0.864 (0.061)	0.084 (0.006)	0.143 (0.009)	0.206 (0.014)
PCD	Oracle1	0.924 (0.003)	0.889 (0.004)	0.861 (0.004)	0.949 (0.002)	0.922 (0.002)	0.906 (0.002)
	Oracle2	0.954 (0.003)	0.959 (0.002)	0.952 (0.003)	0.971 (0.002)	0.969 (0.002)	0.971 (0.002)
	Linear	0.910 (0.003)	0.857 (0.004)	0.794 (0.005)	0.948 (0.002)	0.911 (0.002)	0.891 (0.003)
	Kernel	0.912 (0.003)	0.861 (0.004)	0.817 (0.005)	0.946 (0.002)	0.911 (0.002)	0.892 (0.003)
	aLASSO	0.929 (0.004)	0.908 (0.004)	0.878 (0.006)	0.961 (0.002)	0.953 (0.002)	0.952 (0.003)

Table 4.2: Simulation results for Case 2 with 20% censoring. Results are averaged over 100 replications. Numbers in parentheses are standard errors.

		$n=100$			$n=200$		
		$p=10$	$p=20$	$p=30$	$p=10$	$p=20$	$p=30$
MSE	Oracle1	0.337 (0.017)	0.991 (0.053)	2.232 (0.111)	0.135 (0.008)	0.319 (0.012)	0.548 (0.022)
	Oracle2	0.088 (0.007)	0.085 (0.007)	0.097 (0.009)	0.041 (0.003)	0.036 (0.003)	0.041 (0.003)
	Linear	0.677 (0.039)	3.307 (0.199)	44.819 (19.975)	0.314 (0.015)	0.550 (0.022)	1.214 (0.050)
	Kernel	0.508 (0.026)	1.163 (0.048)	2.493 (0.115)	0.238 (0.011)	0.447 (0.017)	0.716 (0.025)
	aLASSO	0.562 (0.041)	0.785 (0.049)	0.877 (0.058)	0.227 (0.015)	0.425 (0.020)	0.529 (0.025)
PCD	Oracle1	0.925 (0.003)	0.890 (0.004)	0.862 (0.004)	0.949 (0.002)	0.922 (0.002)	0.911 (0.002)
	Oracle2	0.958 (0.003)	0.960 (0.002)	0.956 (0.003)	0.970 (0.002)	0.971 (0.002)	0.973 (0.002)
	Linear	0.886 (0.005)	0.820 (0.005)	0.765 (0.006)	0.922 (0.003)	0.889 (0.003)	0.858 (0.004)
	Kernel	0.896 (0.005)	0.853 (0.004)	0.823 (0.005)	0.931 (0.003)	0.901 (0.003)	0.880 (0.003)
	aLASSO	0.910 (0.006)	0.890 (0.005)	0.886 (0.006)	0.946 (0.003)	0.938 (0.003)	0.939 (0.003)

Table 4.3: Simulation results for Case 3 with 20% censoring. Results are averaged over 100 replications. Numbers in parentheses are standard errors.

		$n=100$			$n=200$		
		$p=10$	$p=20$	$p=30$	$p=10$	$p=20$	$p=30$
MSE	Oracle1	0.327 (0.017)	0.855 (0.048)	1.779 (0.082)	0.124 (0.008)	0.302 (0.010)	0.503 (0.019)
	Oracle2	0.078 (0.006)	0.081 (0.008)	0.085 (0.007)	0.034 (0.002)	0.037 (0.003)	0.037 (0.003)
	Linear	0.738 (0.040)	3.063 (0.187)	22.304 (3.060)	0.377 (0.018)	0.626 (0.026)	1.179 (0.038)
	Kernel	0.598 (0.030)	1.333 (0.055)	2.578 (0.098)	0.315 (0.013)	0.597 (0.020)	0.930 (0.026)
	aLASSO	0.661 (0.040)	1.162 (0.063)	1.820 (0.080)	0.334 (0.017)	0.567 (0.025)	0.738 (0.033)
PCD	Oracle1	0.921 (0.003)	0.887 (0.004)	0.856 (0.003)	0.948 (0.002)	0.919 (0.002)	0.907 (0.003)
	Oracle2	0.955 (0.003)	0.959 (0.002)	0.952 (0.003)	0.969 (0.002)	0.968 (0.002)	0.970 (0.002)
	Linear	0.870 (0.005)	0.816 (0.005)	0.740 (0.007)	0.915 (0.003)	0.876 (0.003)	0.853 (0.003)
	Kernel	0.884 (0.004)	0.831 (0.004)	0.774 (0.005)	0.922 (0.003)	0.882 (0.003)	0.858 (0.003)
	aLASSO	0.895 (0.005)	0.849 (0.007)	0.794 (0.012)	0.940 (0.003)	0.917 (0.004)	0.912 (0.004)

Table 4.4: Simulation results for Case 1 with 40% censoring. Results are averaged over 100 replications. Numbers in parentheses are standard errors.

		$n=100$		$n=200$	
		$p=10$	$p=20$	$p=10$	$p=20$
MSE	Oracle1	0.463 (0.024)	1.405 (0.071)	0.177 (0.008)	0.439 (0.018)
	Oracle2	0.115 (0.010)	0.129 (0.012)	0.047 (0.003)	0.050 (0.004)
	Linear	1.233 (0.097)	13.357 (2.316)	0.272 (0.017)	1.046 (0.051)
	Kernel	0.533 (0.027)	1.507 (0.071)	0.188 (0.009)	0.483 (0.018)
	aLASSO	0.350 (0.025)	0.706 (0.057)	0.109 (0.007)	0.188 (0.012)
PCD	Oracle1	0.907 (0.004)	0.869 (0.004)	0.940 (0.002)	0.910 (0.002)
	Oracle2	0.948 (0.003)	0.954 (0.003)	0.965 (0.002)	0.966 (0.002)
	Linear	0.891 (0.004)	0.821 (0.005)	0.935 (0.002)	0.895 (0.003)
	Kernel	0.896 (0.004)	0.846 (0.004)	0.935 (0.003)	0.900 (0.002)
	aLASSO	0.916 (0.004)	0.888 (0.006)	0.955 (0.003)	0.947 (0.003)

Table 4.5: Simulation results for Case 2 with 40% censoring. Results are averaged over 100 replications. Numbers in parentheses are standard errors.

		$n=100$		$n=200$	
		$p=10$	$p=20$	$p=10$	$p=20$
MSE	Oracle1	0.486 (0.026)	1.457 (0.075)	0.182 (0.011)	0.437 (0.017)
	Oracle2	0.117 (0.011)	0.113 (0.011)	0.047 (0.004)	0.046 (0.004)
	Linear	1.216 (0.109)	10.980 (3.630)	0.377 (0.023)	0.764 (0.031)
	Kernel	0.645 (0.032)	1.651 (0.075)	0.302 (0.017)	0.519 (0.019)
	aLASSO	0.633 (0.052)	0.932 (0.059)	0.258 (0.018)	0.417 (0.023)
PCD	Oracle1	0.909 (0.003)	0.869 (0.004)	0.942 (0.002)	0.913 (0.003)
	Oracle2	0.950 (0.003)	0.955 (0.003)	0.966 (0.002)	0.967 (0.002)
	Linear	0.873 (0.005)	0.794 (0.006)	0.913 (0.003)	0.877 (0.003)
	Kernel	0.885 (0.004)	0.834 (0.005)	0.919 (0.003)	0.891 (0.003)
	aLASSO	0.898 (0.006)	0.869 (0.005)	0.939 (0.003)	0.930 (0.003)

Table 4.6: Simulation results for Case 3 with 40% censoring. Results are averaged over 100 replications. Numbers in parentheses are standard errors.

		$n=100$		$n=200$	
		$p=10$	$p=20$	$p=10$	$p=20$
MSE	Oracle1	0.457 (0.024)	1.291 (0.065)	0.171 (0.009)	0.414 (0.014)
	Oracle2	0.103 (0.008)	0.105 (0.009)	0.043 (0.002)	0.052 (0.004)
	Linear	1.113 (0.065)	7.111 (0.646)	0.421 (0.022)	0.846 (0.034)
	Kernel	0.670 (0.032)	1.555 (0.069)	0.340 (0.015)	0.668 (0.023)
	aLASSO	0.746 (0.047)	1.322 (0.075)	0.329 (0.017)	0.585 (0.031)
PCD	Oracle1	0.906 (0.004)	0.865 (0.004)	0.940 (0.002)	0.907 (0.003)
	Oracle2	0.947 (0.003)	0.949 (0.003)	0.965 (0.002)	0.962 (0.002)
	Linear	0.853 (0.005)	0.790 (0.006)	0.904 (0.003)	0.860 (0.004)
	Kernel	0.874 (0.004)	0.818 (0.005)	0.916 (0.003)	0.870 (0.003)
	aLASSO	0.879 (0.006)	0.831 (0.008)	0.936 (0.003)	0.905 (0.004)

Table 4.7: Variable selection results averaged over 100 replications.

	Censor	n	p	Size	Exact	Cover	Corr0	Incorr0			
Case 1	20%	100	10	4.73	0.60	1.00	0.896	0.000			
			20	5.55	0.22	0.89	0.900	0.038			
			30	5.73	0.17	0.81	0.925	0.075			
		200	10	4.44	0.67	1.00	0.937	0.000			
			20	4.60	0.68	1.00	0.965	0.000			
			30	5.17	0.49	1.00	0.957	0.000			
	40%	100	10	4.74	0.55	0.98	0.891	0.005			
			20	5.44	0.23	0.83	0.898	0.073			
			200	10	4.36	0.71	1.00	0.949	0.000		
		200	20	4.81	0.61	1.00	0.952	0.000			
			Case 2	20%	100	10	4.79	0.40	0.92	0.873	0.025
						20	5.32	0.27	0.82	0.909	0.055
30	5.90	0.23				0.82	0.920	0.065			
200	10	4.70		0.52	1.00	0.900	0.000				
	20	4.93		0.47	1.00	0.945	0.000				
	30	4.85		0.53	0.99	0.968	0.003				
40%	100	10	4.80	0.29	0.83	0.853	0.058				
		20	5.41	0.15	0.74	0.895	0.093				
		200	10	4.65	0.58	1.00	0.907	0.000			
	200	20	4.83	0.52	0.99	0.951	0.003				
		Case 3	20%	100	10	5.01	0.35	0.94	0.844	0.020	
					20	5.32	0.14	0.68	0.894	0.123	
30	4.57				0.04	0.40	0.932	0.321			
200	10		4.60	0.53	1.00	0.914	0.000				
	20		5.12	0.45	0.99	0.934	0.003				
	30		5.51	0.23	0.93	0.941	0.018				
40%	100	10	4.70	0.34	0.83	0.869	0.055				
		20	4.68	0.08	0.55	0.910	0.211				
		200	10	4.66	0.56	1.00	0.906	0.000			
	200	20	5.04	0.40	0.97	0.936	0.010				

REFERENCES

- Billingsley, P. (1986), *Probability and measure*, Wiley, New York.
- Brinkley, J., Tsiatis, A., and Anstrom, K. (2010), “A generalized estimator of the attributable benefit of an optimal treatment regime,” *Biometrics*, 66, 512–522.
- Bühlmann, P. and Hothorn, T. (2007), “Boosting algorithms: Regularization, prediction and model fitting,” *Statistical Science*, 22, 477–505.
- Bühlmann, P. and Yu, B. (2003), “Boosting with the L2 loss: Regression and classification,” *Journal of the American Statistical Association*, 98, 324–339.
- Bühlmann, M. D. (2003), *Radial basis functions: Theory and implementations*, Cambridge University Press.
- Cai, T., Tonini, G., and Lin, X. (2010), “Kernel machine approach to testing the significance of multiple genetic markers for risk prediction,” *Biometrics*.
- Conway, J. (1990), *A course in functional analysis*, vol. 96, Springer.
- Cook, R. and Li, B. (2002), “Dimension reduction for conditional mean in regression,” *Annals of Statistics*, 455–474.
- Cook, R. and Weisberg, S. (1991), “Discussion of sliced inverse regression for dimension reduction,” *Journal of the American Statistical Association*, 86, 328–332.
- Cook, R. D. (1998), *Regression graphics: Ideas for studying regressions through graphics*, Wiley, New York.
- Cox, D. R. (1972), “Regression models and life-tables,” *Journal of the Royal Statistical Society, Series B*, 34, 187–220.
- (1975), “Partial likelihood,” *Biometrika*, 62, 269–276.
- Cristianini, N. and Shawe-Taylor, J. (2000), *An introduction to support vector machines: and other kernel-based learning methods*, Cambridge University Press.
- 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.
- (2002), “Variable selection for Cox’s proportional hazards model and frailty model,” *Annals of Statistics*, 30, 74–99.

- Fan, J. and Lv, J. (2008), “Sure independence screening for ultrahigh dimensional feature space,” *Journal of The Royal Statistical Society, Series B*, 70, 849–911.
- Freund, Y. (1995), “Boosting a weak learning algorithm by majority,” *Information and Computation*, 121, 256–285.
- Freund, Y. and Schapire, R. E. (1997), “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of Computer and System Sciences*, 55, 119–139.
- Friedman, J., Hastie, T., and Tibshirani, R. (2000), “Additive logistic regression: a statistical view of boosting,” *Annals of Statistics*, 28, 337–374.
- Friedman, J. H. (2001), “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, 29, 1189–1232.
- Friedman, J. H., Hastie, T., and Tibshirani, R. (2010), “Regularization paths for generalized linear models via coordinate descent,” *Journal of Statistical Software*, 33, 1–22.
- Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005), “The elements of statistical learning: data mining, inference and prediction,” *The Mathematical Intelligencer*, 27, 83–85.
- Hotelling, H. (1936), “Relations between two sets of variables,” *Biometrika*, 321–327.
- Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J., and Lauer, M. S. (2010), “High-dimensional variable selection for survival data,” *Journal of the American Statistical Association*, 105, 205–217.
- Kimeldorf, G. and Wahba, G. (1971), “Some results on Tchebycheffian spline functions,” *Journal of Mathematical Analysis and Applications*, 33, 82–95.
- Li, B., Kim, M. K., and Altman, N. (2010a), “On dimension folding of matrix- or array-valued statistical objects,” *Annals of Statistics*, 38, 1094–1121.
- Li, B. and Wang, S. (2007), “On directional regression for dimension reduction,” *Journal of the American Statistical Association*, 102, 997–1008.
- Li, H. and Luan, Y. (2005), “Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data,” *Bioinformatics*, 21, 2403–2409.
- Li, K.-C. (1991), “Sliced inverse regression for dimension reduction,” *Journal of the American Statistical Association*, 86, 316–327.

- (1992), “On principal hessian directions for data visualization and dimension reduction: Another application of Stein’s lemma,” *Journal of the American Statistical Association*, 87, 1025–1039.
- Li, L. (2009), “Exploiting predictor domain information in sufficient dimension reduction,” *Computational Statistics and Data Analysis*, 53, 2665–2672.
- Li, L., Li, B., and Zhu, L.-X. (2010b), “Groupwise dimension reduction,” *Journal of the American Statistical Association*, 105, 1188–1201.
- Liu, D., Ghosh, D., and Lin, X. (2008), “Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models,” *BMC bioinformatics*, 9, 292.
- Liu, D., Lin, X., and Ghosh, D. (2007), “Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models,” *Biometrics*, 63, 1079–1088.
- Lu, W. and Li, L. (2008), “Boosting method for nonlinear transformation models with censored survival data,” *Biostatistics*, 9, 658–667.
- Lu, W., Zhang, H. H., and Zeng, D. (2011), “Variable selection for optimal treatment decision,” *Statistical Methods in Medical Research*, To appear.
- Mann, M. E., Zhang, Z., Hughes, M. K., Bradley, R. S., Miller, S. K., Rutherford, S., and Ni, F. (2008), “Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia,” *Proceedings of the National Academy of Sciences*, 105, 13252–13257.
- Murphy, S. (2003), “Optimal dynamic treatment regimes,” *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 331–366.
- (2005), “An experimental design for the development of adaptive treatment strategies,” *Statistics in Medicine*, 24, 1455–1481.
- Naik, P. A. and Tsai, C.-L. (2005), “Constrained inverse regression for incorporating prior information,” *Journal of the American Statistical Association*, 100, 204–211.
- Pankratz, V., de Andrade, M., and Therneau, T. (2005), “Random-effects Cox proportional hazards model: General variance components methods for time-to-event data,” *Genetic Epidemiology*, 28, 97–109.
- Qian, M. and Murphy, S. (2011), “Performance guarantees for individualized treatment rules,” *Annals of Statistics*, 39, 1180–1210.

- Ridgeway, G. (1999), “The state of boosting,” *Computing Science and Statistics*, 31, 172–181.
- Ripatti, S. and Palmgren, J. (2000), “Estimation of multivariate frailty models using penalized partial likelihood,” *Biometrics*, 56, 1016–1022.
- Robins, J. (2004), “Optimal structural nested models for optimal sequential decisions,” in *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data*.
- Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, K. H., Smeland, E. B., Giltneane, J. M., Hurt, E. M., Zhao, H., Averett, L., Yang, L., Wilson, W. H., Jaffe, E. S., Simon, R., Klausner, R. D., Powell, J., Duffey, P. L., Longo, D. L., Greiner, T. C., Weisenburger, D. D., Sanger, W. G., Dave, B. J., Lynch, J. C., Vose, J., Armitage, J. O., Montserrat, E., Lopez-Guillermo, A., Grogan, T. M., Miller, T. P., Leblanc, M., Ott, G., Kvaloy, S., Delabie, J., Holte, H., Krajci, P., Stokke, T., and Staudt, L. M. (2002), “The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma.” *New England Journal of Medicine*, 346, 1937–1947.
- Rubin, D. (1974), “Estimating causal effects of treatments in randomized and nonrandomized studies,” *Journal of educational Psychology*, 66, 688.
- (1978), “Bayesian inference for causal effects: The role of randomization,” *The Annals of Statistics*, 34–58.
- Schapire, R. E. (1990), “The strength of weak learnability,” *Machine Learning*, 5, 197–227.
- Schölkopf, B. and Smola, A. (2002), *Learning with kernels*, The MIT Press.
- Suykens, J. (2002), *Least squares support vector machines*, World Scientific Pub Co Inc.
- Therneau, T. (2003), “On mixed-effect Cox models, sparse matrices, and modeling data from large pedigrees,” *Mayo Clinic, Rochester, USA*.
- Tibshirani, R. (1996), “Regression shrinkage and selection via the Lasso,” *Journal of the Royal Statistical Society, Series B*, 58, 267–288.
- (1997), “The lasso method for variable selection in the Cox model,” *Statistics in Medicine*, 16, 385–395.
- van Houwelingen, H. C., Bruinsma, T., Hart, A. A. M., van’t Veer, L. J., and Wessels, L. F. A. (2006), “Cross-validated Cox regression on microarray gene expression data,” *Statistics in Medicine*, 25, 3201–3216.

- van Wieringen, W. N., Kun, D., Hampel, R., and Boulesteix, A.-L. (2009), “Survival prediction using gene expression data: A review and comparison,” *Computational Statistics & Data Analysis*, 53, 1590–1603.
- Vapnik, V. (1998), *Statistical learning theory*, Wiley-Interscience.
- Wang, H. (2009), “Forward regression for ultra-high dimensional variable screening,” *Journal of the American Statistical Association*, 104, 1512–1524.
- Wang, H. and Leng, C. (2007), “Unified LASSO estimation by least squares approximation,” *Journal of the American Statistical Association*, 102, 1039–1048.
- Witten, D. M. and Tibshirani, R. (2010), “Survival analysis with high-dimensional covariates.” *Statistical methods in medical research*, 19, 29–51.
- Xia, Y., Tong, H., Li, W. K., and Zhu, L.-X. (2002), “An adaptive estimation of dimension reduction space,” *Journal Of The Royal Statistical Society Series B*, 64, 363–410.
- Yin, X., Li, B., and Cook, R. D. (2008), “Successive direction extraction for estimating the central subspace in a multiple-index regression,” *Journal of Multivariate Analysis*, 99, 1733–1757.
- 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.
- (2007), “On the nonnegative garrote estimator,” *Journal of the Royal Statistical Society, Series B.*, 69, 143–161.
- Zhang, H. H. and Lu, W. (2007), “Adaptive Lasso for Cox’s proportional hazards model,” *Biometrika*, 94, 691–703.
- Zhao, Y., Kosorok, M., and Zeng, D. (2009), “Reinforcement learning design for cancer clinical trials,” *The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series*, 11.
- Zou, H. (2006), “The adaptive Lasso and its oracle properties,” *Journal of the American Statistical Association*, 101, 1418–1429.
- Zou, H. and Hastie, T. (2005), “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society, Series B*, 67, 301–320.