View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Purdue E-Pubs

Purdue University
Purdue e-Pubs

Open Access Dissertations | Theses and Dissertations

Fall 2014

Nonparametric variable selection and dimension reduction methods and their applications in pharmacogenomics

Jingyi Zhu
Purdue University

Follow this and additional works at: https://docs.lib.purdue.edu/open_access_dissertations

 Part of the Mathematics Commons

Recommended Citation
Zhu, Jingyi, "Nonparametric variable selection and dimension reduction methods and their applications in pharmacogenomics" (2014). *Open Access Dissertations*. 404.
https://docs.lib.purdue.edu/open_access_dissertations/404

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

Graduate School Form
(Revised /)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Jingyi Zhu

Entitled
NONPARAMETRIC VARIABLE SELECTION AND DIMENSION REDUCTION METHODS AND THEIR APPLICATIONS IN PHARMACOGENOMICS

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Jun Xie

Mary Ellen Bock

Lingsong Zhang

Michael Zhu

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification/Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Jun Xie

Approved by Major Professor(s):

Approved by: Jun Xie 10/09/2014

Head of the 'HSDUWPHQW Graduate Program Date

NONPARAMETRIC VARIABLE SELECTION AND DIMENSION REDUCTION METHODS AND THEIR APPLICATIONS IN PHARMACOGENOMICS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Jingyi Zhu

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2014

Purdue University

West Lafayette, Indiana

To my husband, Yuanzhe Xi, my son Jayden Xi,
my parents, Guoping Zhu and Jiajing Pu.

ACKNOWLEDGMENTS

First and foremost I would like to express my deepest gratitude to my advisor, Professor Jun Xie for her guidance and support during my Ph.D. study. I would never have been able to finish my dissertation without her help. Her dedication to research and friendly personality have impressed me deeply and motivated me towards the completion of my thesis work.

I am grateful to my committee members for their continuous support: Profs. Mary Ellen Bock, Lingsong Zhang, and Michael Zhu. Special thanks to Prof. Zhang, who is willing to serve as my examining committee at the last moment.

I also would like to thank some current and graduated Purdue graduate students for sharing this incredible journey with me. Special thanks to Yating Cao, Veavi Chang, Jing Dong, Will Eagan, Whitney Huang, and Vitara Pungpapong.

Lastly, I would like to express my sincere gratitude to my friends and family, especially to my husband and parents, who have provided meticulous care in my daily life, and supported me as always whenever I need them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Zhu, Jingyi Ph.D., Purdue University, December 2014. Nonparametric Variable Selection and Dimension Reduction Methods and Their Applications in Pharmacogenomics . Major Professor: Jun Xie.

Nowadays it is common to collect large volumes of data in many fields with an extensive amount of variables, but often a small or moderate number of samples. For example, in the analysis of genomic data, the number of genes can be very large, varying from tens of thousands to several millions, whereas the number of samples is several hundreds to thousands. Pharmacogenomics is an example of genomics data analysis that we are considering here. Pharmacogenomics research uses whole-genome genetic information to predict individuals' drug response. Because whole-genome data are high dimensional and their relationships to drug response are complicated, we are developing a variety of nonparametric methods, including variable selection using local regression and extended dimension reduction techniques, to detect nonlinear patterns in the relationship between genetic variants and clinical response.

High dimensional data analysis has become a popular research topic in the Statistics society in recent years. However, the nature of high dimensional data makes many traditional statistical methods fail, because most methods rely on the assumption that the sample size n is larger than the variable dimension p. Consequently, variable selection or dimension reduction is often the first step in high dimensional data analysis. Meanwhile, another important issue arises as the choice of an appropriate statistical modeling strategy for conducting variable selection or dimension reduction. It has been found from our studies that the traditional parametric linear model might not work well for detecting nonlinear patterns of relationships between predictors and response. The limitations of the linear model and other parametric

statistical approaches motivate us to consider nonparametric/nonlinear models for conducting variable selection or dimension reduction.

The thesis is composed of two major parts. In the first part, we develop a nonparametric predictive model of the response based on a small number of predictors, which are selected from a nonparametric forward variable selection procedure. We also propose strategies to identify subpopulations with enhanced treatment effects. In the second part, we develop an alternating least squares method to extend the classical Sliced Inverse Regression (SIR) [Li, 1991] to the context of high dimensional data. Both methods are demonstrated by simulation studies and a pharmacogenomics study of bortezomib in multiple myeloma [Mulligan *et al.*, 2007]. The proposed methods have favorable performances compared to other existing approaches in the literature.

1. NONPARAMETRIC VARIABLE SELECTION FOR PREDICTIVE MODELS AND SUBPOPULATIONS IN CLINICAL TRIALS

1.1 Introduction

Variable selection is often the first step in developing predictive models. There are many reasons for focusing on a subset of predictors: the desire to develop statistical procedures that are more efficient in making inferences, the interpretability of the estimated predictive model, and the concern of making the statistical procedures computationally effective and robust. The need of variable selection is stronger, when we have high dimensional data with a large number of variables. Suppose that we have a response variable Y, and a set of p predictors $X_1, \ldots, X_p$. The objective of variable selection is to examine the relationship between Y and a subset of $X_1, \ldots, X_p$. In sections 1.1.1 and 1.1.2, we review the existing variable selection methods in the context of linear models. Specifically, the relationship between Y and $X_1, \ldots, X_p$ is modeled as $Y = \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$, where ϵ is an error term following a standard normal distribution, and $\beta_1, \ldots, \beta_p$ are the regression coefficients that we want to estimate. Given a sample set of n subjects, $Y = (Y_1, \ldots, Y_n)$ and $\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix}$, the model can also be written in a matrix form, i.e., $\mathbf{Y} = \mathbf{X}\beta + \epsilon$.

1.1.1 Classical variable selection methods

Traditionally, there are two major types of variable selection methods. The first one is known as the best subset selection, which selects the best model among all possible combinations of the predictors based on some specific selection criterion. Examples of well-known selection criteria include the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Mallows' C_p. All of these three criteria consider a tradeoff between the goodness of fit of a model and its complexity. Specifically, AIC is aimed at selecting a model that minimizes the expected estimated Kullback-Leibler divergence of the fitted model from the true one [Akaike, 1973]. As an alternative to AIC, BIC is developed to maximize the posterior probability under the Bayesian framework and has a different representation of the model complexity [Schwarz, 1978]. Mallow's C_p is proposed to minimize the mean squared error of prediction [Mallows, 1973]. The best subset selection is known to be computationally expensive, which is impossible to implement when the dimension p grows large.

The other method is known as the heuristic variable selection procedure, which is often employed to select a subset of predictors in a sequential order. The best known examples of this procedure include forward selection, backward elimination, and stepwise selection. The forward selection procedure starts from the null model with no variable included, then adds the most significant variable to the model if its p-value is below some pre-determined significant level. Variables are continually added to the model one at a time until none of the remaining variables are significant when added to the model. In contrast, the backward elimination procedure is conducted in the opposite direction. It begins with the full model with all the variables included, and excludes the least significant variable from the model at a chosen significant level. This procedure continues to exclude variable from the model one at a time until all the remaining variables are statistically significant. The stepwise selection approach is a combination of forward selection and backward elimination, in the sense that it allows movement in either direction by adding or dropping variables at various steps.

It can either work as forward selection, but reconsider dropping variables already in the model, if they are no longer significant when other variables are added, or as backward elimination, but reconsider adding back variables excluded from the model earlier if they later appear to be significant. Compared to the best subset selection, the heuristic variable selection procedure is less computationally demanding, making it feasible for selecting subsets among a large number of predictors. However, it is not guaranteed to obtain the global optimal solution.

1.1.2 Regularization methods for variable selection

More recently, regularization methods have also been used as variable selection approaches, for example, Least Absolute Shrinkage and Selection Operator (LASSO), Smoothly Clipped Absolute Deviation (SCAD), elastic net, and Least Angle Regression (LARS). LASSO was proposed by Tibshirani [1996] in the context of linear models to minimize the residual sum of squares (RSS) subject to a L_1 penalty term controlled by a regularization parameter. Mathematically, the LASSO estimates can be obtained by minimizing the following constrained objective function

$$\hat{\beta}_{LASSO} = argmin_{\beta} \|Y - \mathbf{X}\beta\|^2 \qquad s.t. \sum_{i=1}^{p} |\beta_i| \leq s \quad ,$$

where s is a pre-determined positive constant. Or equivalently, we can find the LASSO estimates by solving the following optimization problem,

$$\hat{\beta}_{LASSO} = argmin_{\beta} \left\{ \|Y - \mathbf{X}\beta\|^2 + \lambda \sum_{i=1}^{p} |\beta_i| \right\} \quad ,$$

where λ is a nonnegative regularization parameter. As the regularization parameter λ increases, the coefficient estimates are shrunk towards zero and some of them become exactly zero, which can be excluded from the model. In the two extreme cases, if the regularization parameter λ equals zero, LASSO is equivalent to ordinary least squares (OLS). On the other hand, if the regularization parameter λ goes to infinity, all the coefficients are shrunk to be zero, thus no predictor is included in the model.

Elastic net was proposed by Zou and Hastie [2005] as an alternative to LASSO. Mathematically, the elastic net is trying to solve the following optimization problem,

$$\hat{\beta}_{elasticnet} = argmin_{\beta}\|Y - \mathbf{X}\beta\|^2 + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p} \|\beta_i\|^2 \quad ,$$

where λ_1 and λ_2 are two nonnegative regularization parameters for the L_1 and L_2 penalty terms respectively. Elastic net differs from LASSO in the sense that it combines both L_1 and L_2 penalty terms, where the L_1 penalty generates sparsity, and the L_2 penalty favors selection of a group of correlated predictors. Elastic net is known to have a grouping effect, where a group of significant predictors is selected together, whereas LASSO tends to select one predictor in a group but ignore the others. Fan and Li [2001] proposed SCAD, which uses a penalized likelihood approach to select significant predictors. The penalty function is specially defined to satisfy several good properties, such as symmetry and nonconcaveness, so that the resulting estimator is sparse, unbiased, and continuous. Fan and Li [2001] also proposed an oracle property in terms of penalized least squares. If a method satisfies the oracle property, then the coefficient estimates of the zero components in the model will converge to zero with probability tending to 1 and the coefficient estimates of the nonzero components can be obtained as if the true correct model is known in advance. Fan and Li [2001] argued that a good variable selection method should favor the oracle property. It is shown that SCAD satisfies the oracle property with a proper choice of the regularization parameter while both LASSO and elastic net do not. Efron *et al.* [2004] proposed LARS, which is a less greedy variable selection method compared to forward variable selection. The LARS algorithm is approximately implemented as follows. It starts with no variable in the model, and adds the variable which most correlates with the response. The algorithm then moves in the direction of the first selected predictor until some other variable is just as equally correlated with the current residual. After the entering of the second predictor in the model, it keeps moving in a direction such that the residual stays equally correlated with the first two predictors until the third variable enters the model with the largest correlation with the residual among

the remaining predictors. This procedure is repeated until some stopping criterion is satisfied.

The regularization methods conduct variable selection only in linear or parametric models. However, the linear relationship between Y and one or more predictors $\mathbf{X}$ is often too simple to be proper in the complicated data analysis. The traditional linear model does not work for detecting nonlinear patterns of the relationship between Y and $\mathbf{X}$. The limitations of the linear model and other parametric statistical approaches motivate our use of nonparametric methods to model the relationship between Y and $\mathbf{X}$.

1.1.3 Review on pharmacogenomics research

According to the definition of the American Medical Association (AMA), pharmacogenomics is the study of genetic variations that influence individual response to drugs. While a number of clinical and laboratory features such as age and disease index provide prognostic information, they may still be unable to define the highest risk patients most in need of novel therapies. It is anticipated that the pharmacogenomics research will help provide more precise prognostic and predictive tools in patient treatments. The pharmacogenomics research will also contribute towards facilitating the development of personalized medicine, which is tailored to the needs of different individuals. Figure 1.1 shows a simple example, which is adapted from http://psylab.idv.tw. When a group of patients are treated with a specific drug, they will usually have different drug responses; some would be good responders, some would be poor responders, and some would have adverse effects in the worst case. As shown in Figure 1.1, it can be found through pharmacogenomic studies that patients' genotypes are correlated with their corresponding drug responses. By utilizing the findings from the pharmacogenomics research, medical treatments are developed based on patients' genotypes, namely, personalized medicine.

Success in this important public health endeavor relies on efficient and accurate statistical and computational methods in the analysis of genomic data and clinical outcomes, which is essentially a predictive or regression problem. It is anticipated that the effect of genetic variants on drug response is highly combinatorial and non-linear. In a review article on bioinformatics challenges for genome-wide association studies [Moore *et al.*, 2010], the linear modeling framework is considered as a major limitation of the current studies. We aim to develop a nonparametric predictive model of clinical response using a large set of potential predictors, including the whole genome genetic information as well as standard clinical and laboratory features. The pharmacogenomics research provides a good application of the statistical methods for high dimensional data, such as variable selection and dimension reduction developed here.

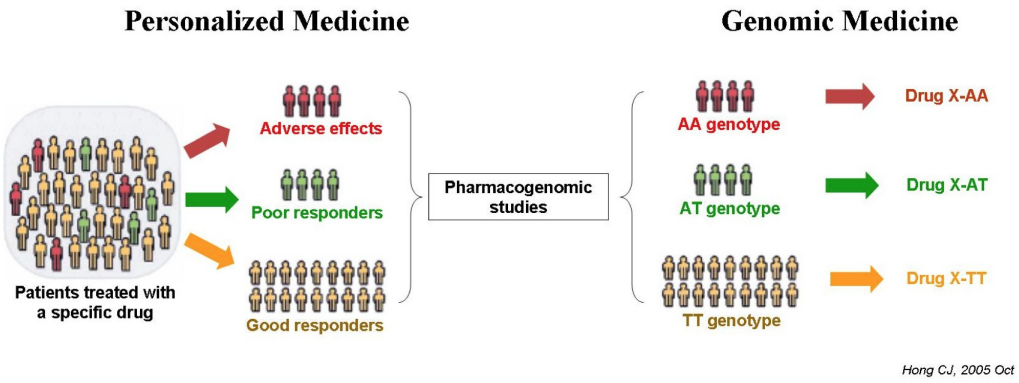

Figure 1.1. Illustration of contributions of the pharmacogenomics research towards the development of personalized medicine (cited from http://psylab.idv.tw).

1.1.4 Review on subpopulation with enhanced treatment effect

In most clinical trials, there is much heterogeneity among individual outcomes and the treatment effect may not be the same on all of the patients. If we could

determine which patients will respond better to the treatment, ideally ahead of time, but also possibly soon after the treatment is administered, the development and subsequent utilization of the therapy would be dramatically improved. It is substantively interesting but challenging to identify such patient subpopulations that will derive a more pronounced benefit from the active treatment than the rest of patients. More specifically, consider a clinical trial with patients' drug response and a large number of potential predictors, such as genetic information, clinical features, and demographic information. We want to develop statistical methods to select predictive covariates and consequently use them to define subpopulations with enhanced treatment effects.

The identification of subpopulation with enhanced treatment effect is recently a popular topic in clinical practice and medical research. It relates to the efforts in discovering patient-specific treatment strategy, or personalized medicine. Crump *et al.* [2006] conducted statistical tests for the heterogeneity of treatment effects across pre-specified patient subpopulations. Moineddin *et al.* [2008] proposed a multi-level random-effect model to identify subpopulations from patient baseline characteristics. Ruberg *et al.* [2010] and Foster *et al.* [2011] proposed to use a CART (Classification and Regression Tree) approach to select predictors and consequently to use their cut-off values to define subpopulations of patients. The tree splitting idea in CART was further explored by Lipkovich *et al.* [2011]. Zhang *et al.* [2012] used a regression model for the expected clinical response conditional on treatment and covariates. The parametric regression model defines an optimal treatment regime, or equivalently, subpopulations.

Here we propose a nonparametric method to model the expected response conditional on a small set of selected covariates. We intend to relax any parametric model assumptions, hence the method is not limited by misspecification of the regression model for the response. Our method is a combination of forward variable selection and nonparametric local regression. Forward variable selection is merely a heuristic procedure, but it is easy to implement and can obtain a subset of predictors with reasonably small prediction errors. Meanwhile, the use of nonparametric local regres-

sion has the major advantage that we make fewer assumptions about the functional form of the model for the clinical response. Our idea is in analogy with that of Zhang *et al.* [2012] but generalizes their parametric regression model to a nonparametric model. Specifically, Zhang *et al.* [2012] assumed a linear response model where the parameters were estimated through least or generalized least squares. After we developed the method, we found that a similar idea had been suggested by Storlie and Helton [2007] but under the context of reliability analysis. We demonstrate our variable selection approach using data simulated from a simplified yet realistic clinical trial. Our method has high accuracy in test data sets and performs comparably to the CART method. We also implement our method in a pharmacogenomics study of bortezomib in multiple myeloma [Mulligan *et al.*, 2007] and compare it with another existing method of linear predictive model. In the bortezomib example, our nonparametric model with three predictors achieves the same prediction power as a linear model with a large number of predictors (over 100).

1.2 Subpopulation definition in our method

Consider a randomized clinical trial. Each patient receives either an active treatment or placebo at random. Let $\mathbf{X} = (X_1, ..., X_p)$ denote a vector of p predictors (genetic biomarkers, demographics, etc.), and Y denote a clinical response. In principle, the clinical response Y has two components, Y_{trt} and $Y_{control}$, where Y_{trt} is the clinical response if a patient receives the active treatment, and $Y_{control}$ is the clinical response if a patient receives the placebo. We consider two types of treatment effects: the global treatment effect and the conditional treatment effect. The global treatment effect is $E(Y_{trt} - Y_{control})$, where $E(.)$ denotes the expectation of Y. The conditional treatment effect is $E(Y_{trt} - Y_{control}|\mathbf{X})$, where $E(.|\mathbf{X})$ denotes the conditional expectation of Y given $\mathbf{X}$.

Denote the sample space of $\mathbf{X}$ as $\mathcal{X}$. A partition of $\mathcal{X}$ defines subpopulations of patients. A subpopulation with an enhanced treatment effect is defined as a patient

group with covariates values in a subset of $\mathcal{X}$ that has a larger conditional treatment effect than the global treatment effect. Formally, the subpopulation can be represented as

$$S = \{\mathbf{X} \in \mathcal{X} : \ E(Y_{trt} - Y_{control}|\mathbf{X}) > E(Y_{trt} - Y_{control})\} \quad .$$

This subpopulation definition is similar to the idea of optimal treatment regime used in Zhang *et al.* [2012] except that Zhang *et al.* [2012] considered $E(Y_{trt} - Y_{control}|\mathbf{X})$ versus 0. We prefer to compare $E(Y_{trt} - Y_{control}|\mathbf{X})$ to the global treatment effect $E(Y_{trt} - Y_{control})$. In fact, without loss of generality, we assume $E(Y_{trt} - Y_{control}) \geq 0$ throughout the paper. Therefore, our definition of subpopulation is more rigorous. If we indeed have a clinical trial with $E(Y_{trt} - Y_{control}) < 0$, we will replace the inequality by $E(Y_{trt} - Y_{control}|\mathbf{X}) > 0$ in the above definition and modify our implementation procedure accordingly.

Given a data set, i.e., a randomized clinical trial with n patients, we estimate $E(Y_{trt} - Y_{control})$ by the difference of the sample means between the treatment and the control groups. We estimate the conditional expectations $E(Y_{trt} - Y_{control}|\mathbf{X})$ by the difference of two nonparametric functions of $\mathbf{X}$, one for $E(Y_{trt}|\mathbf{X})$ and the other for $E(Y_{control}|\mathbf{X})$. More specifically, we have

$$S = \{\mathbf{X} \in \mathcal{X} : \ \hat{g}_{trt}(\mathbf{X}) - \hat{g}_{control}(\mathbf{X}) > \bar{Y}_{trt} - \bar{Y}_{control}\} \quad ,$$

where $\hat{g}_{trt}(.)$ denotes the nonparametric estimate of $E(Y_{trt}|\mathbf{X})$ in the treatment group, and $\hat{g}_{control}(.)$ denotes the nonparametric estimate of $E(Y_{control}|\mathbf{X})$ in the control group, $\bar{Y}_{trt}$ and $\bar{Y}_{control}$ are the sample means of Y in the treatment and the control groups, respectively. Note that we are modeling the treatment and the control groups separately, instead of considering a combined response model with covariates $\mathbf{X}$, a treatment variable, and their interactions. In fact, once we relax the parametric model assumption, the model for the treatment group and that for the control group are arbitrary functions and in different functional forms. Therefore, the interactions of treatment and covariates on treatment effects are automatically incorporated. In

the following, we develop a nonparametric variable selection approach to estimate $E(Y_{trt}|\mathbf{X})$ and $E(Y_{control}|\mathbf{X})$, which provide predictive models for the response and also derive the subpopulation S.

1.3 Variable selection via LOESS

We first describe the method of variable selection in the context of nonparametric models. Let $y_i, i = 1, \cdots, n$, denote n measurements of the response variable Y. Let $\mathbf{x}_{ij}, i = 1, \cdots, n$ and $j = 1, \cdots, p$, denote n observations of p potential predictors. Without loss of generality we assume that the data of each predictor, $x_{1j}, \cdots, x_{nj}$, have been normalized so that all predictors have the same scale, for example, with mean 0 and standard deviation 1.

1.3.1 Estimation via LOESS given a fixed set of predictors

Assume a given subset of multiple predictors, $(X_1, X_2, ..., X_d)$, where d is the number of predictors in the model and is often limited to four. Let $\mathbf{x}_i = (x_{i1}, \cdots, x_{id}), i = 1, \cdots, n$, be n measurements of the selected predictors. Assume a model of the form

$$y_i = g(\mathbf{x}_i) + \epsilon_i \quad ,$$

where $g(.)$ is an unknown smooth function and ϵ_i's are i.i.d. error terms with mean 0 and finite variance σ^2 and independent of $\mathbf{x}_i$. The model assumption implies that $E(Y|\mathbf{X}) = g(\mathbf{X})$. In our previous subpopulation discussions we have two such nonparametric models, one for Y_{trt} and the other for $Y_{control}$. Cleveland [1979] proposed the locally weighted scatterplot smoothing (LOWESS), a local regression method for a response variable Y on a single predictor X. It was further generalized to multivariate predictors, known as LOESS [Cleveland and Devlin, 1988], to model the relationship between a response variable Y and multiple predictors. With LOESS we can estimate a large class of smooth functions without being restricted to a specific class of parametric functions. The estimate of g at a single point $\mathbf{x}$ uses all neighbors

around $\mathbf{x}$, where the neighbors are decided by a span parameter. Each neighbor is then weighted according to its distance from $\mathbf{x}$ by a kernel function. A linear or quadratic function of $\mathbf{x}$ is fitted to Y using weighted least squares.

1.3.2 Forward selection criterion

We want to conduct forward variable selection using LOESS. Predictors are added to the model one by one if they are statistically significant. The local regression method implies that

$$\hat{g}(\mathbf{x}) = \sum_{i=1}^{n} l_i(\mathbf{x}) y_i \quad .$$

That is, the LOESS estimate $\hat{g}(\mathbf{x})$, is a linear combination of the observed response y_i where $l_i(\mathbf{x})$ depends on the observed predictor values in a neighborhood of $\mathbf{x}$, but not on y_i. This form of local estimator will be specifically shown in Section 1.4. Let $\hat{y}_i = \hat{g}(\mathbf{x}_i)$ be the fitted values, $\hat{\epsilon}_i = y_i - \hat{y}_i$ be the residuals, and denote $y = (y_1, \cdots, y_n)^t$, $\hat{y} = (\hat{y}_1, \cdots, \hat{y}_n)^t$, $\hat{\epsilon} = (\hat{\epsilon}_1, \cdots, \hat{\epsilon}_n)^t$. Since each $\hat{y}_i$ is a linear combination of y we have $\hat{y} = Ly$ where L is an $n \times n$ matrix and $\hat{\epsilon} = (I - L)y$ where I is the $n \times n$ identity matrix. Suppose that one predictor has been selected and $L_1 y$ is the vector of its fitted values. We consider adding a second predictor into the model and let $L_2 y$ be the fitted values using the two predictors. We want to test a null hypothesis H_0 of one predictor for y against an alternative hypothesis H_a of two predictors using a nonparametric F-test. More specifically, let $y^t R_1 y = y^t (I - L_1)(I - L_1)^t y$ and $y^t R_2 y = y^t (I - L_2)(I - L_2)^t y$ be the residual sum of squares of the two fits. Under H_0, we have the following test statistic

$$\hat{F} = \frac{(y^t R_1 y - y^t R_2 y)/\upsilon_1}{(y^t R_2 y)/\delta_1} \quad ,$$

which approximately follows an F distribution with the degrees of freedom υ_1^2/υ_2 and δ_1^2/δ_2, where $\upsilon_1 = trace(R_1 - R_2)$, $\upsilon_2 = trace(R_1 - R_2)^2$, $\delta_1 = trace(R_2)$, and $\delta_2 = trace(R_2)^2$ [Cleveland and Devlin, 1988]. The degrees of freedom are obtained by generalizing the F-test statistic of parametric models to a nonparametric case.

Another definition of the test statistic was used in Storlie and Helton [2007], with different degrees of freedom defined by υ_1 and δ_1 for the F distribution under H_0. However, our simulations show that Storlie and Helton [2007] test statistic tends to over-select variables, including more predictors in the model than needed.

We select the first predictor by comparing a null hypothesis of a constant model to an alternative hypothesis of one predictor for Y. If the most significant single predictor has a p-value less than a cutoff, e.g., 0.01, it is added into the model. Next we consider adding a second predictor into the model. We conduct F-tests for all possible second predictors. If none of the second predictors is significant, we end with a model with only one predictor. Otherwise, we select the most significant one and extend the model to two predictors. We continue using the nonparametric F-test as the criterion to select significant predictors and this procedure stops if no predictor is found to be significant.

1.3.3 Smoothing parameter selection

Nonparametric methods, including LOESS, use a smoothing parameter to control potential over-fitting of local regression. The smoothing parameter here is the proportion of the neighbor points out of all data points that are used to fit $g(\mathbf{x})$ at $\mathbf{x}$. It is referred to as the span parameter α. If α is too small insufficient data fall within the neighborhood resulting in an over-fitting with large variance. On the other hand, if α is too large the local regression may not fit data well resulting in a fit with large bias. Thus the span parameter must be chosen to compromise the bias-variance trade-off.

Commonly used criteria of selecting span parameters in general nonparametric techniques, e.g., smoothing splines, include AIC and Generalized Cross Validation.

In our method, we use an extended version of AIC, known as AIC_C. It was first introduced by Hurvich and Tsai [1989] for a linear model,

$$\begin{aligned} AIC_C &= log(\hat{\sigma}^2) + \frac{1 + d/n}{1 - (d+2)/n} \\ &= log(\hat{\sigma}^2) + 1 + \frac{2(d+1)}{n-d-2}, \end{aligned}$$

where $\hat{\sigma}^2$ is the average of residual sum of squares and d is the number of variables included in the model. Hurvich *et al.* [1998] further generalized AIC_C to the context of nonparametric regression for span parameter selection. The AIC_C score for a local estimate with the smoothing parameter α is

$$\begin{aligned} AIC_C(\alpha) &= log(\hat{\sigma}^2) + \frac{1 + tr(L_\alpha)/n}{1 - \{tr(L_\alpha) + 2\}/n} \\ &= log(\hat{\sigma}^2) + 1 + \frac{2\{tr(L_\alpha) + 1\}}{n - tr(L_\alpha) - 2}, \end{aligned}$$

where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{g}_\alpha(x_i))^2$ is the estimated error variance and L_α is the estimation matrix L as defined in Section 1.3.2 but depends on α. It was shown in the simulation study by Hurvich *et al.* [1998] that compared to the criterion of Generalized Cross Validation or AIC, the use of AIC_C avoided the large variability and the tendency to undersmooth. AIC_C is also easy to apply in practice since it is a function of L only through its trace.

Ideally an optimal span parameter is chosen where the AIC_C score is minimized. However, it is found from our study that AIC_C scores have several different patterns as shown in Figure 1.2. For example, the reduction of AIC_C scores becomes negligible near the upper boundary of the span parameter. Therefore minimizing AIC_C tends to choose larger span parameters which may not be necessary. Here we define a modified criterion that selects an optimal span parameter $\hat{\alpha}$ which is at least 0.2 and satisfies

$$\hat{\alpha} = \min_{\alpha}\{\operatorname*{argmin}_{\alpha} AIC_C, \operatorname*{argmax}_{\alpha} \triangle AIC_C, \arg_{\alpha}\{\triangle AIC_C = 0\}\} \quad . \tag{1.1}$$

Looking for α that is the root of $\triangle AIC_C = 0$ is an alternative criterion to minimize AIC_C. We also study the change of AIC_C, i.e., $\triangle AIC_C$, with large changes being

favorable. In implementation, we consider a grid of the span parameter between 0.2 and 0.8, as suggested in Cleveland and Devlin [1988], and with an increment of 0.01. We select the smallest α value in the range of 0.2 to 0.8, which either corresponds to the minimum of AIC_C or the maximum change of AIC_C.

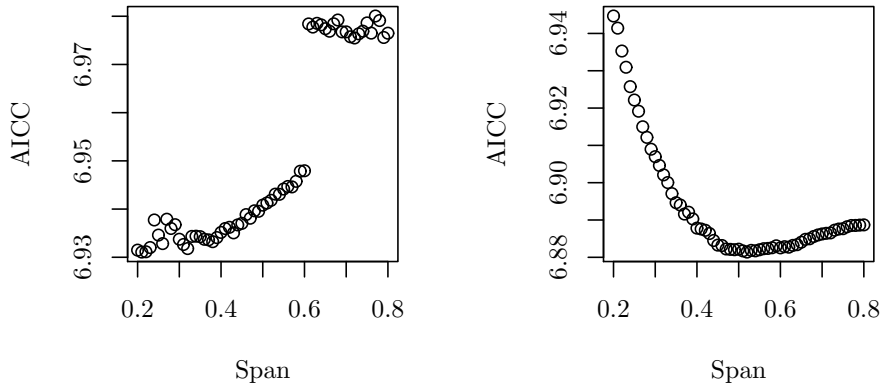

Figure 1.2. Illustration of different patterns of AIC_C versus the span parameter.

In our variable selection procedure, we first use (1.1) to select an optimal span parameter for each candidate predictor $X_i, i = 1, \cdots, p$. Let $\hat{\alpha}_1$ be the optimal span parameter for the most significant predictor at the first step. In the following steps, the span parameter is chosen to be the maximum value between 0.2 and the power of $\hat{\alpha}_1$, i.e., $\hat{\alpha}_1^d$, where d is the number of predictors selected into the model and is up to 4. This is a simple rule to consider cubic neighborhoods from multi-dimensional predictors and use the same $\hat{\alpha}_1$ value for all predictors added into the model. For instance, assume we have $\hat{\alpha}_1 = 0.8$ and $d = 3$. We will use $\hat{\alpha}_1^3 = 0.512$ to define a neighborhood when we estimate $g(\mathbf{x})$ for any given value $\mathbf{x}$ in the 3-dimensional space $\mathbf{R}^3$. More specifically, we use the amount of $0.512n$ observations around $\mathbf{x}$ to estimate $g(\mathbf{x})$, where Euclidean distance defines the neighbor points around $\mathbf{x}$. This amount

of neighbor points will further be weighted by a multivariate kernel functin, e.g., a tricube kernel in $\mathbf{R}^3$, according to their Euclidean distance to $\mathbf{x}$. An appropriately chosen span parameter and a small number of predictors, i.e., $d \leq 4$, help prevent overfitting of the nonparametric model.

1.3.4 Predictive model for subpopulation

We fit predictive models for the treatment and the control groups separately after selecting significant predictors for these two groups respectively. Let $\mathbf{X}_{trt}$ be the significant predictors for the treatment group and $\mathbf{X}_{control}$ be the significant predictors for the control group. Since it is possible that either $\mathbf{X}_{trt}$ or $\mathbf{X}_{control}$ is $\emptyset$, we define a subset of the sample space $\mathcal{X}$ in three different ways.

1. $\mathbf{X}_{trt} \neq \emptyset$, *but* $\mathbf{X}_{control} = \emptyset$.
   The subpopulation is defined as
   $$S = \{\mathbf{X}_{trt} \in \mathcal{X} : \ \hat{g}_{trt}(\mathbf{X}_{trt}) > \bar{Y}_{trt}\}.$$

2. $\mathbf{X}_{trt} = \emptyset$, *but* $\mathbf{X}_{control} \neq \emptyset$.
   The subpopulation is defined as
   $$S = \{\mathbf{X}_{control} \in \mathcal{X} : \ \hat{g}_{control}(\mathbf{X}_{control}) < \bar{Y}_{control}\}.$$

3. *Both of* $\mathbf{X}_{trt}$ *and* $\mathbf{X}_{control}$ *are* $\neq \emptyset$.
   Let $\mathbf{X} = (\mathbf{X}_{trt}, \mathbf{X}_{control})$. The subpopulation is defined as
   $$S = \{\mathbf{X} \in \mathcal{X} : \ \hat{g}_{trt}(\mathbf{X}_{trt}) - \hat{g}_{control}(\mathbf{X}_{control}) > \bar{Y}_{trt} - \bar{Y}_{control}\}.$$

Patients, whose covariates values are within the subset S such defined, correspond to a subpopulation with enhanced treatment effects. These patiens will benefit the most from the treatment and we should design a treatment regime to specifically assign them the treatment.

1.4 Property of local regression

In this section we provide properties of local regression after we select d variables, where d is the number of significant predictors identified by the nonparametric variable selection and $d \ll p$. We first show that the local function estimate $\hat{g}(\mathbf{x})$ is linear in $\mathbf{Y} = (Y_1, \cdots, Y_n)^t$. Recall the nonparametric model

$$y_i = g(\mathbf{x}_i) + \epsilon_i \quad .$$

Suppose $\mathbf{x} = (x_1, \cdots, x_d)^t$ is a point in $\mathbb{R}^d$, where we want to estimate $g(\mathbf{x})$. Given the observed data $\{X_{ij}\}_{i=1,\cdots,n, j=1,\cdots,d}$, let

$$\mathbf{X_x} = \begin{pmatrix} 1 & X_{11} - x_1 & \cdots & X_{1d} - x_d \\ 1 & X_{21} - x_1 & \cdots & X_{2d} - x_d \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} - x_1 & \cdots & X_{nd} - x_d \end{pmatrix}$$

be a matrix centered at $\mathbf{x}$ and $\mathbf{B} = \alpha \mathbf{I}_d$ be the bandwidth matrix with α as the smoothing parameter. We consider LOESS as a class of kernel-type nonparametric regression estimators, which is generally studied in Fan and Gijbels [1996]. Given a kernel function $K(\mathbf{u})$, for example, the tricube kernel,

$$W(\mathbf{u}) = \begin{cases} (1 - |\mathbf{u}|^3)^3, & \text{if } |\mathbf{u}| \leq 1; \\ 0, & \text{otherwise} \end{cases}$$

we define $K_{\mathbf{B}}(\mathbf{u}) = \dfrac{1}{|\mathbf{B}|} K(\mathbf{B}^{-1}\mathbf{u})$. Furthermore let $\mathbf{W_x} = diag\{K_{\mathbf{B}}(\mathbf{X}_i - \mathbf{x})\}$ denote the $n \times n$ diagonal matrix of weights where $\mathbf{X}_i = (X_{i1}, \cdots, X_{id})^T$. When we use local linear estimate, the estimated function $\hat{g}(.)$ has a linear form with an intercept β_0 and a slope vector β_1.

Lemma 1 *Let $\mathbf{x} = (x_1, \cdots, x_d)^T$ be a point in $\mathbb{R}^d$. Then, $\hat{g}(\mathbf{x})$ is a linear combination of the response $\mathbf{Y}$. That is, there exists a vector $\mathbf{l}(\mathbf{x}) = \{l_i(x)\}_{i=1}^n$ such that $\hat{g}(\mathbf{x}) = \sum_{i=1}^n l_i(x) Y_i$. Furthermore $\mathbf{l}(\mathbf{x})$ has the following representation*

$$\mathbf{l}(\mathbf{x})^T = \mathbf{e}_1^T (\mathbf{X_x}^T \mathbf{W_x} \mathbf{X_x})^{-1} \mathbf{X_x}^T \mathbf{W_x},$$

where $\mathbf{e}_1 = (1, 0, \cdots, 0)^T$ *is the* $d \times 1$ *unit vector.*

Proof The solution of the coefficients $\beta = \{\beta_0, \beta_1\}$ corresponds to minimizing $\sum_{i=1}^{n}\{Y_i - \beta_0 - \beta_1^T(\mathbf{X}_i - \mathbf{x})\}^2 K_{\mathbf{B}}(\mathbf{X}_i - \mathbf{x})$. This is equivalent to solving a weighted least-squares problem. The resulting coefficient estimates have the following form

$$\hat{\beta} = (\mathbf{X}_\mathbf{x}^T \mathbf{W}_\mathbf{x} \mathbf{X}_\mathbf{x})^{-1} \mathbf{X}_\mathbf{x}^T \mathbf{W}_\mathbf{x} \mathbf{Y}.$$

The local linear estimate at $\mathbf{x}$ is just the intercept $\hat{\beta}_0$. If we use the matrix representation, the local linear estimate at $\mathbf{x}$ is

$$\begin{aligned}\hat{g}(\mathbf{x}) &= (1, 0, \cdots, 0)^T (\mathbf{X}_\mathbf{x}^T \mathbf{W}_\mathbf{x} \mathbf{X}_\mathbf{x})^{-1} \mathbf{X}_\mathbf{x}^T \mathbf{W}_\mathbf{x} \mathbf{Y} \\ &= \mathbf{e}_1^T (\mathbf{X}_\mathbf{x}^T \mathbf{W}_\mathbf{x} \mathbf{X}_\mathbf{x})^{-1} \mathbf{X}_\mathbf{x}^T \mathbf{W}_\mathbf{x} \mathbf{Y}\end{aligned}$$

■

In the following, we want to prove the asymptotic consistency of a LOESS estimate at a fixed point $\mathbf{x} = (x_1, \cdots, x_d)^t$. Let f be the d-variate marginal density function of $\mathbf{X}$. We obtain the following theorem about the consistency of the local regression estimate. The basic idea is that, the tricube kernel function that we use and the smoothing parameter selected by AIC_C satisfy the regularity assumptions for the general theorem of nonparametric estimation. The regularity assumptions also guarantee that both the bias and variance of $\hat{g}(\mathbf{x})$ go to zero as the sample size $n \to \infty$.

Assume the following regularity conditions [Ruppert and Wand, 1994].

Assumption 1 *The kernel* K *is a compactly supported, bounded kernel such that* $\int \mathbf{u}\mathbf{u}^T K(\mathbf{u}) d\mathbf{u} = \mu_2(K)\mathbf{I}$*, where* $\mu_2(K) \neq 0$ *is scalar and* $\mathbf{I}$ *is the* $d \times d$ *identity matrix. In addition, all odd-order moments of* K *vanish, that is,* $\int u_1^{l_1} \cdots u_d^{l_d} K(\mathbf{u}) d\mathbf{u} = 0$ *for all nonnegative integers* $l_1, \cdots, l_d$ *such that their sum is odd.*

Assumption 2 *The point* $\mathbf{x}$ *is in the support of the density function* f*, i.e., supp*(f)*. At* $\mathbf{x}$*,* f *is continuous and continuously differentiable and all second-order derivatives of* g *are continuous. Recall that* σ^2 *is the variance of the error term. Also,* $f(\mathbf{x}) > 0$ *and* $\sigma^2 > 0$*.*

Assumption 3 *The sequence of bandwidth matrices* $\mathbf{B}$ *is such that* $n^{-1}|\mathbf{B}|^{-1}$ *and each entry of* $\mathbf{B}\mathbf{B}^T$ *tends to zero as* $n \to \infty$ *with* $\mathbf{B}\mathbf{B}^T$ *remaining symmetric and positive definite. Also, there is a fixed constant* L *such that the condition number of* $\mathbf{B}\mathbf{B}^T$ *(i.e., the ratio of its largest to its smallest eigenvalue) is at most* L *for all* n.

Lemma 2 *[Ruppert and Wand, 1994] Let* $\mathbf{x}$ *be a fixed element in the interior of supp*(f). *Assume that Assumption 1-3 hold. Then,*

$$E\{\hat{g}(\mathbf{x}) - g(\mathbf{x})|\mathbf{X}_1, \cdots, \mathbf{X}_n\} = \frac{1}{2}\mu_2(K)tr\{H(\mathbf{x})\mathbf{B}\mathbf{B}^T\} + o_p\{tr(\mathbf{B}\mathbf{B}^T)\},$$

where $H(\mathbf{x})$ *is the Hessian matrix of* g *at* $\mathbf{x}$,
and

$$Var\{\hat{g}(\mathbf{x})|\mathbf{X}_1, \cdots, \mathbf{X}_n\} = \frac{1}{n|\mathbf{B}|}R(K)\frac{\sigma^2}{f(\mathbf{x})}\{1 + o_p(1)\},$$

where $R(K) = \int K^2(\mathbf{u})d\mathbf{u}$.

Theorem 1.4.1 ***(Consistency)*** *Assume that the three regularity assumptions hold. Then at each continuous point* $\mathbf{x}$ *in the interior of the support of the density function* f, *the LOESS estimator* $\hat{g}(\mathbf{x})$ *is asymptotically unbiased and consistent, i.e., for each* $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\hat{g}(\mathbf{x}) - g(\mathbf{x})| > \epsilon) = 0$$

Proof To prove Theorem 1.4.1 consider the conditional mean squared error of $\hat{g}(\mathbf{x})$,

$$MSE(\hat{g}(\mathbf{x})|\mathbf{X}_1, \cdots, \mathbf{X}_n) = Var(\hat{g}(\mathbf{x})|\mathbf{X}_1, \cdots, \mathbf{X}_n) + (Bias(\hat{g}(\mathbf{x})|\mathbf{X}_1, \cdots, \mathbf{X}_n))^2.$$

According to Lemma 1.4 we have

$$\begin{aligned} Bias(\hat{g}(\mathbf{x})|\mathbf{X}_1, \cdots, \mathbf{X}_n) &= \frac{1}{2}\mu_2(K)tr\{H(\mathbf{x})\mathbf{B}\mathbf{B}^T\} \\ &= \frac{1}{2d}\mu_2(K)tr(\mathbf{B}\mathbf{B}^T)\sum_{i=1}^{d}\frac{\partial^2 g}{\partial x_i^2} \to 0 \quad as \quad n \to \infty, \end{aligned}$$

since $tr(\mathbf{B}\mathbf{B}^T) \to 0$ if each entry of $\mathbf{B}\mathbf{B}^T$ tends to zero as $n \to \infty$. Thus the LOESS estimator $\hat{g}(\mathbf{x})$ is asymptotically unbiased. Similarly,

$$Var(\hat{g}(\mathbf{x})|\mathbf{X}_1, \cdots, \mathbf{X}_n) = \frac{1}{n|\mathbf{B}|} R(K) \frac{\sigma^2}{f(\mathbf{x})} \to 0 \quad as \quad n \to \infty.$$

Hence, $\hat{g}(\mathbf{x})$ is consistent. ■

1.5 Results

1.5.1 Simulation examples

We use data simulated from a simplified yet realistic clinical trial to demonstrate our procedure. The data is from an open challenge of data analysis posted online by Eli Lily & Company's statistical group. It consists of $n = 322$ patients, each with a response Y for the clinical outcome and 99 continuous predictors $X_1, X_2, ..., X_{99}$. There is another treatment index variable indicating whether a patient received an active treatment or placebo, as randomly assigned at the beginning of the trial. There are two data sets generated by two different response models. In the first data set Y only depends on the predictor X_{19} and the ideal subpopulation with enhanced treatment effect is $X_{19} > -0.22$. In the second data set, Y depends on two predictors X_{30} and X_{43} and the ideal subpopulation with enhanced treatment effect is $X_{30} > -0.42$ and $X_{43} > -0.29$.

In our exploratory data analysis we find that some predictors have extreme values which may dramatically affect the model fitting. Therefore, we apply a $5IQR$ rule (5 times interquartile range) to detect possible outliers and exclude them from the following analysis. For the first data set we identify X_{19} as the most significant predictor in the treatment group but no significant predictor in the control group. Figure 1.3 shows that X_{19} can be clearly identified from the LOESS fit with a very small p-value, less than 10^{-5}, but it becomes insignificant in a linear model for Y. In addition, if we consider both the treatment and the control arms and fit a combined linear model with X_{19} and the treatment variable and their interaction, then only a

main treatment effect is significant but not X_{19} nor the interaction. These results indicate that a linear model is not able to identify any significant predictors for the clinical response.

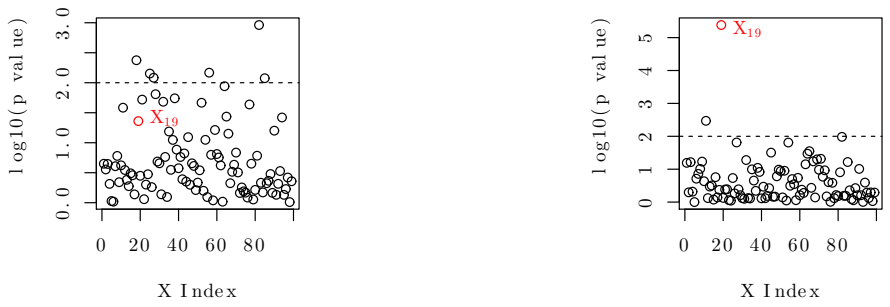

Figure 1.3. Comparison of variable selection between simple linear regression (SLR) and LOESS. The p-values are obtained by SLR (left panel) and LOESS (right panel) from the treatment group of the first data set. The predictive covariate X_{19} is marked in both plots but is only significant according to LOESS.

Applying our method in the first data set, we select one predictor, X_{19}, for the treatment group. The span parameter is chosen as $\alpha = 0.2$ according to the AIC_C criterion (1.1). No significant predictor is selected for the model of the control group. Figure 1.4 shows a plot of nonparametric predictive model for the response in the treatment group. The curve represents the predicted LOESS function versus X_{19}, with the dashed horizontal line as the sample mean of Y in the treatment group. Whenever the LOESS curve crosses the horizontal line with large values, it defines a subpopulation with enhanced treatment effect.

As shown in Figure 1.4, we identify subpopulations with enhanced treatment effects according to $S = \{\mathbf{X}_{trt} \in \mathcal{X} : \ \hat{g}_{trt}(\mathbf{X}_{trt}) > \bar{Y}_{trt}\}$. To stabilize the prediction results, we further apply a refinement procedure to validate or discard the identified subpopulations. Specifically, in this example we first construct a number of nonoverlapping subpopulation intervals. For each nonoverlapping interval, if it contains a small number of observations (less than 4) or the length of the interval is small (less

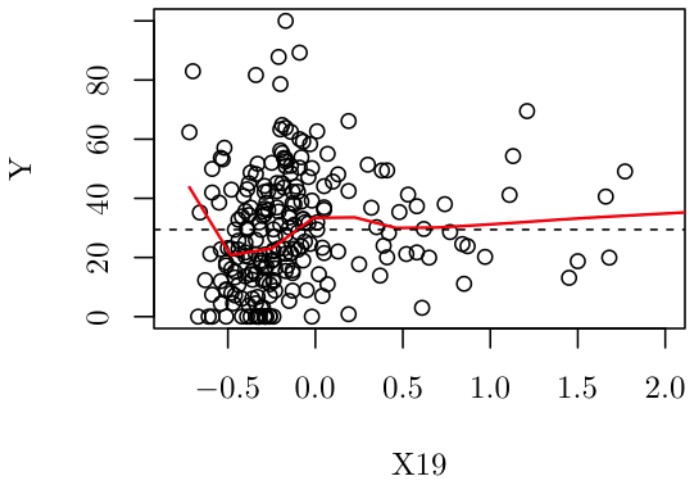

Figure 1.4. Subpopulation identification plot of the first data set in a zoomed-in region not including large values of the predictor. The curve represents the predicted LOESS curve, with the dashed horizontal line as the sample mean of Y in the treatment group.

than 0.1), we discard the corresponding interval. Otherwise we retain this interval. Consequently, the subpopulation with enhanced treatment effect identified by our procedure is $X_{19} \in [-0.23, 0.49]$ and $X_{19} \in [0.63, 10.85]$ for the first data set. A careful examination of Figure 1.4 actually indicates that the fitted curve is above the reference line for almost all values of $X_{19} \geq -0.23$. Therefore, we may use the whole set $X_{19} \geq -0.23$ as the subpopulation.

For the second data set we identify X_{30} as the most significant predictor with a p-value 3.67×10^{-5} for the treatment group at the first step of variable selection. The span parameter is chosen as $\alpha = 0.5$ according to the AIC_C criterion (1.1). The predictor X_{43} has a p-value 0.00212 at the first step of variable selection but it becomes insignificant with a p-value 0.23 at the second step of variable selection, after X_{30} is already in the model. Therefore, we only select one predictor X_{30} for the treatment group. No significant predictor is selected for the model of the control group. We identify a subpopulation as $X_{30} \geq -0.37$.

For these simulated studies, we know the ideal subpopulations hence use the truth to assess our method and compare it with the CART approach. Table 1.1 shows the prediction errors for the first data set, where the identified subpopulation of $X_{19} \in [-0.23, 0.49]$ and $X_{19} \in [0.63, 10.85]$ is compared to the ideal subpopulation defined by $X_{19} > -0.22$. We have an overall accuracy of 95% (Sensitivity=94%, Specificity=95%). If we use the whole set $X_{19} \geq -0.23$ to define the subpopulation, as indicated in Figure 1.4, then we obtain an overall accuracy of 98% (Sensitivity=100%, Specificity=95%). These results are comparable to that of CART, as provided online with the simulated data. Table 1.2 shows the prediction errors for the second data set. The identified subpopulation by our procedure is compared to the ideal subpopulation defined by $X_{30} > -0.42$ and $X_{43} > -0.29$. Table 1.2 also reports the prediction errors of the CART method. Although we only select one predictor X_{30} for the second data set, the performance of our procedure is promising with an overall accuracy of 82% (Sensitivity=94%, Specificity=70%), whereas the CART model has an overall accuracy of 83% (Sens=100%, Spec=66%). Our nonparametric method with few se-

lected variables provides good results of subpopulations in these examples, compared to both the truth and a competing method.

To assess robustness of our proposed procedure we further implement 10-fold cross validation. We randomly divide the data set into 10 subsets/folds, where nine folds of the data form a training set to select variables and fit a model, and the remaining one fold serves as a test set for prediction errors. We repeat the process 10 times with each of the 10 folds as the test set. The predictions of whether an observation belongs to the subpopulation are then combined together for these 10 simulations. Table 1.3 shows the results of cross validation for the first data set. Our proposed method has an overall accuracy of 92% (Sensitivity=89%, Specificity=95%), indicating good performances in replications. Table 1.4 shows the prediction errors of the second data set using 10-fold cross validation. The overall accuracy of 79% is lower than 84% when we use all data to fit the predictive model (shown in Table 1.2). However, the result of 10-fold cross validation shows that the procedure is reliable in multiple replicates of the simulation.

1.5.2 Application in a pharmacogenomics example

We implement our method in a pharmacogenomics study of bortezomib in multiple myeloma [Mulligan *et al.*, 2007]. Multiple myeloma is an incurable malignancy and bortezomib is the first therapeutic proteasome inhibitor tested in humans for treating relapsed multiple myeloma. As the new active agent bortezomib is a therapeutic choice in addition to the standard chemotherapy, there is a need to reliably identify the patient population that will mostly benefit from the therapy. The data set of this study is available at Gene Expression Omnibus (GSE9782). There are four clinical trials conducted in multiple centers in the United States, Canada, Europe, and Israel from June 2002 to October 2003, denoted as trial 024, trial 025, trial 039, and trial 040, with a total of 264 patients. The clinical outcome is a five-level variable denoting a patient response after a therapy, ranging from progressive disease,

Table 1.1
Subpopulation prediction table of the first data set.

		Ideal		
		In S	Not in S	Total
Identified	In S	148	8	156
	Not in S	9	157	166
Total		157	165	322

Sensitivity: 94%

Specificity: 95%

Positive Predictive Value: 95%

Negative Predictive Value: 95%

Accuracy: 95%

Table 1.2
Comparison of subpopulation prediction table of the second data set. The upper panel is the result from our method, and the lower panel is the result from Eli Lily's regression tree model.

		Ideal		
		In S	Not in S	Total
Identified	In S	149	49	198
	Not in S	10	114	124
Total		159	163	322

Sensitivity: 94%

Specificity: 70%

Positive Predictive Value: 75%

Negative Predictive Value: 92%

Accuracy: 82%

		Ideal		
		In S	Not in S	Total
Identified	In S	159	55	214
	Not in S	0	108	108
Total		159	163	322

Sensitivity: 100%

Specificity: 66%

Positive Predictive Value: 74%

Negative Predictive Value: 100%

Accuracy: 83%

Table 1.3
Subpopulation prediction table of the first data set by using 10-fold cross validation.

		Ideal		
		In S	Not in S	Total
Identified	In S	140	8	148
	Not in S	17	157	174
Total		157	165	322

Sensitivity: 89%

Specificity: 95%

Positive Predictive Value: 95%

Negative Predictive Value: 90%

Accuracy: 92%

Table 1.4
Subpopulation prediction table of the second data set by using 10-fold cross validation.

		Ideal		
		In S	Not in S	Total
Identified	In S	135	45	180
	Not in S	24	118	142
Total		159	163	322

Sensitivity: 85%

Specificity: 72%

Positive Predictive Value: 75%

Negative Predictive Value: 83%

Accuracy: 79%

no change, minimal response, partial response to complete response. The potential predictors include 44,928 gene expression values.

In the clinical trials, most patients (169) received the new treatment bortezomib but only 70 patients in trial 039 received chemotherapy as controls. Therefore, in this study we focus on analysing only the treatment group, i.e., the novel therapy bortezomib. The goal here is to construct predictive models for the clinical response in the treatment group and identify patient subpopulations who respond the best to the treatment, instead of comparing the treatment and the control groups.

In the exploratory study of Mulligan *et al.* [2007], the five-category clinical response was simplified to two levels: progressive disease (PD) and response (R), excluding no change (NC) patients. Since the data is high dimensional with 44,928 genes but 264 patients, they applied a two-stage gene filtering method in which only the 9200 gene probe sets with the strongest between-sample variance relative to their in-sample replicate variance were retained. Among the 9200 genes only the top 100 differentially expressed genes with respect to clinical response (PD vs. R) by t-tests were used as predictors in the predictive model. The data were also divided into a training data set for trials 025 and 040, and a test data set for trial 039. Trial 024 was not used in the analysis due to a very small number of patients (7) with evaluable response. A linear predictor classifier [Wright *et al.*, 2003] was developed on the training data and was used to classify each patient to be either PD or R in the test data. The classifier is based on a linear combination of the 100 predictors with each being weighted by its t-test score.

In our procedure we use the original five-level clinical responses as the response variable Y and encode the ordinal categories by values 1-5 such that progressive disease (PD) is coded as 1, no change (NC) is coded as 2, minimal response (MR) is coded as 3, partial response (PR) is coded as 4 and complete response (CR) is coded as 5. We use the same gene filtering method of Mulligan *et al.* [2007] to keep only the top 100 differentially expressed genes for further analysis. We also divide the data

into a training data for trials 025 and 040 and a test data for trial 039. There are 91 observations in the training set and 71 in the test set.

We first conduct the nonparametric forward variable selection on the training data to identify a small number of significant genes. Three genes are selected by the forward variable selection when we set the significance level at 0.05. At the first step gene 219233_s_at is selected with p-value 8.83×10^{-6} from the F-test. At the second step gene 212240_s_at is selected with p-value 0.0076 from the F-test. At the third step gene 200017_at is selected with the corresponding p-value of the F-test to be 0.042. Gene 219233_s_at, known as gasdermin B, may play a role as secretory or metabolic product involved in secretory pathway. Gene 212240_s_at, known as phosphoinositide 3-kinase regulatory subunit 1, has several important biological functions and is necessary for the insulin-stimulated increase in glucose uptake and glycogen synthesis in insulin-sensitive tissues. Gene 200017_at, known as ribosomal protein S27a, is a component of the 40S subunit of the ribosome. Identification of these genes provides a hint for further studies in myeloma and the treatment effect of bortezomib. They will also be used to define subpopulations that benefit from bortezomib.

A predictive model with the three genes is fitted by LOESS. The span parameter is chosen to be 0.8 at the first step of the variable selection according to AIC_C. The span parameter is about 0.6, as the square of 0.8, at the second step and 0.5 ($0.8^3 = 0.512$) at the third step. The predictive model is used to predict patient response in the test data. If the predicted clinical response value is less than or equal to 2 it is classified as nonresponse (NR), which includes both PD and NC, otherwise it is classified as response (R), which includes MR, PR and CR. A subpopulation can be defined as $S = \{\mathbf{X}_{trt} \in \mathcal{X} : \ \hat{g}_{trt}(\mathbf{X}_{trt}) > 2\}$, which corresponds to the patient subgroup that responds to the bortezomib treatment.

As demonstration, Figure 1.5 shows two plots of the nonparametric LOESS fit for patient response in the test data, one projected on the two-dimensional space consisting of the first predictor 219233_s_at and the response Y, the other projected on the three-dimensional space consisting of the first two predictors, 219233_s_at and

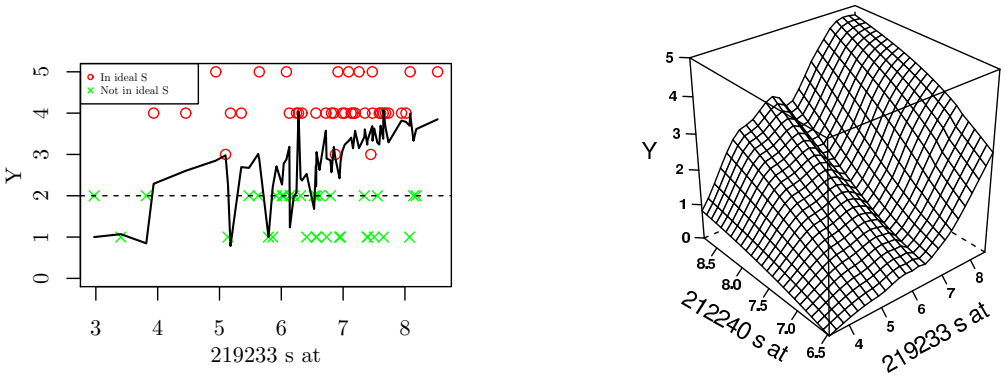

Figure 1.5. Predictive model and subpopulation identification of the bortezomib data set. Left: Only the first significant predictor is used in the plot, although three predictors are selected. The patients who are responders are labelled by circles and the patients who are non-responders are labelled by crosses. The curve represents the LOESS curve, but projected on the first predictor, with the dashed horizontal line as the cutoff value 2. Right: The first two significant predictors are used to show the LOESS curve projected on the three-dimensional space.

Table 1.5
Comparison of prediction table of PD vs. R for the bortezomib data in the pharmacogenomics example. The upper panel is the result from our method and the lower panel is the result from Mulligan *et al.* [2007].

		Actual		
		R	PD	Total
Predicted	R	37	13	50
	PD	1	2	3
Total		38	15	53

Sensitivity: 97%

Specificity: 13%

Positive Predictive Value: 74%

Negative Predictive Value: 67%

Accuracy: 74%

		Actual		
		R	PD	Total
Predicted	R	35	10	45
	PD	3	5	8
Total		38	15	53

Sensitivity: 92%

Specificity: 33%

Positive Predictive Value: 78%

Negative Predictive Value: 63%

Accuracy: 75%

212240_s_at, and Y. In the two-dimensional plot, the patients who are responders are labelled by circles and the patients who are non-responders (NC and PD) are labelled by crosses. The dashed horizontal line displays the cutoff value 2. The patient subpopulation is defined to be $\{\mathbf{X} : \hat{g}(\mathbf{X}) > 2\}$, where $\hat{g}(.)$ is the predictive LOESS function. We compare the prediction result of PD vs. R based on our predictive model on the test data with the result from Mulligan *et al.* [2007] as shown in Table 1.5. In summary, we identify 37 out of the 38 patients who are responders to the bortezomib treatment and only one patient, who is a responder to the treatment, is incorrectly classified as progressive disease (PD). Two out of the 15 patients who have PD to the treatment are correctly classified, but the other 13 patients are incorrectly classified as responders to the treatment. The overall accuracy of our prediction result is 74%, which is comparable to 75% of Mulligan *et al.* [2007]'s result. Thus, with a nonparametric model a small number of predictors can achieve the same prediction power as a linear model with a much larger number of predictors (100). Applying to a new patient our model will predict whether or not it has a more pronounced benefit from the bortezomib treatment.

1.6 Discussion

We perform a nonparametric forward variable selection procedure to identify significant predictive covariates among a large set of potential predictors which may include standard clinical and laboratory features and whole genome gene expression measurements. Forward variable selection is merely a heuristic procedure but it has the advantage of easy implementation in practice. It is not limited by high dimensional predictors because it starts with a constant model with only an intercept and will end up with selecting a small number of significant predictors. The local regression method, LOESS, is computationally efficient, especially when we consider a small number of significant predictors (≤ 4). In practice we recommend checking for outliers before doing any analysis. We also need to pay attention to fitting a local

regression in the boundary of data. While local linear regression has been shown to provide a simple and effective way of modeling slopes in the boundary region and reduce bias compared to other kernel estimates (e.g., Nadaraya-Watson estimate) [Hastie and Loader, 1993], the mean squared error of the local linear regression may still be big when data are sparse and of high curvature in the boundary. Fortunately, the boundary effect is not severe in our examples. In general, visualization of the data and their LOESS fitting in the boundary is always recommended.

While variable selection helps to build a predictive model based on a small number of significant predictors, an alternative approach is to consider that interesting features of high-dimensional data are retrievable from low-dimension projections. In the next chapter, we explore dimension reduction techniques in analysis of high dimensional but low sample size data.

2. EXTENSION OF SLICED INVERSE REGRESSION FOR HIGH DIMENSIONAL BUT LOW SAMPLE SIZE DATA

2.1 Introduction

High dimensional data analysis often suffers from the curse of dimensionality [Bellman, 1961], because any given size of data becomes sparse as the dimension increases. In fact, the amount of data needed for an appropriate inference, either estimation or prediction, grows exponentially with the dimension of variables. In practice, a sample size is often comparable to or even less than the dimension of variables. Therefore, direct application of many traditional statistical methods becomes problematic when the dimension is large. In addition, large sample theory does not hold any more when we have a relatively small sample size. We have discussed variable selection methods in the previous chapter. As an alternative to variable selection methods, dimension reduction techniques can be employed, which assume that high dimensional features can be extracted from their low dimensional subspace. We are studying dimension reduction methods in this chapter. Again, suppose that we have a response variable Y, and a p-dimensional predictor $\mathbf{X} = (X_1, \ldots, X_p)^T$, where p is a large number. Consider a general regression model,

$$Y = f(\beta_1 \mathbf{X}, \ldots, \beta_d \mathbf{X}, \epsilon), \tag{2.1}$$

where the β's are unknown row vectors, ϵ is an error term independent of $\mathbf{X}$, and f is an arbitrary unknown function on $\mathbf{R}^{d+1}$. In practice, d is often a much smaller number compared to the dimension p. If Model 2.1 holds, the projection of the p-dimensional predictor $\mathbf{X}$ onto a d-dimensional subspace, i.e., $\beta_1 \mathbf{X}, \ldots, \beta_d \mathbf{X}$, captures all the information we need about predicting the response variable Y, or equivalently, the distribution of Y is independent of $\mathbf{X}$ given $\beta_1 \mathbf{X}, \ldots, \beta_d \mathbf{X}$. For many problems of

interest, for example, classification and subpopulation identification, the dimension reduction directions, β's, are more important than the actual function $f(.)$. We do not need to estimate the general function $f(.)$ but can still obtain a direction for $\mathbf{X}$ to predict Y. In the pharmacogenomics example which we will revisit in the application section of this chapter, we will conduct a classification analysis, to predict response versus nonresponse patients, where we only need good estimators of β's. We can also incorporate dimension reduction techniques in identifying subpopulation. Accordingly, how to obtain the estimates of β's is the key issue of dimension reduction methods. For simplicity, we often work on the standardized scale of $\mathbf{X}$, denoted as $\mathbf{Z} = \Sigma_{\mathbf{x}}^{-1/2}(\mathbf{X} - E(\mathbf{X}))$. Model 2.1 can be rewritten as

$$Y = f(\eta_1\mathbf{Z}, \ldots, \eta_d\mathbf{Z}, \epsilon), \tag{2.2}$$

where $\eta_i = \beta_i\Sigma_{\mathbf{x}}^{1/2}, i = 1, \ldots, d$ denotes the standardized dimension reduction direction. Note that Model 2.1 does not specify the functional form of f, which can be a very general function of the predictors.

In order to find β's in (2.1), several dimension reduction methods have been developed since early 1990s. Li [1991] first proposed Sliced Inverse Regression (SIR) which conducted an eigenvalue decomposition of the conditional covariance matrix $cov(E(\mathbf{X}|Y))$. A detailed review of Li [1991]'s SIR can be found in Section 2.1.1. Following Li [1991]'s work, Cook and Weisberg [1991] proposed Sliced Average Variance Estimation (SAVE), which used the second order moment $var(\mathbf{X}|Y)$ instead of the first order moment $E(\mathbf{X}|Y)$. Li [1992] proposed principal Hessian direction (pHd), where the dimension reduction directions were estimated through finding the eigenstructure of a sample Hessian matrix. Bura and Cook [2001] proposed Parametric Inverse Regression (PIR) which assumed a multivariate linear model for the p inverse regressions and fitted smooth parametric curves. Cook and Ni [2005] developed Inverse Regression Estimation (IRE) by minimizing a quadratic discrepancy function.

In addition, the Sufficient Dimension Reduction (SDR) theory was established (Cook [1994a],Cook [1998]), which introduced the concept of sufficiency in dimension reduction methods. With respect to regression problems, a reduction $R(\mathbf{X})$ of the

p-dimensional predictor $\mathbf{X}$ is sufficient if the conditional distribution of Y given $\mathbf{X}$ is the same as the distribution of Y given $R(\mathbf{X})$. Hence, SDR seeks to replace the p-dimensional predictor vector with its projection onto a subspace of the predictor space without loss of information on $Y|\mathbf{X}$. In terms of statistical terminology, SDR is to find a d-dimensional subspace $\mathbf{S}$ such that

$$Y \perp \mathbf{X}|P_{\mathbf{S}}\mathbf{X}, \tag{2.3}$$

where $\perp$ indicates independence, and $P_{\mathbf{S}}$ represents a projection operator in the standard inner product. The projection subspace $\mathbf{S}$ satisfying (2.1) is called a dimension reduction subspace for $Y|\mathbf{X}$. There exist many dimension reduction subspaces satisfying (2.1), and the intersection of all these subspaces also satisfies (2.1) under mild conditions [Cook, 1996]. Such intersection is defined as the central subspace, denoted as $\mathbf{S}_{Y|\mathbf{X}}$, and its dimension, denoted as $d = dim(\mathbf{S}_{Y|\mathbf{X}})$. The central subspace can be interpreted as the unique minimal subspace that preserves all the original information of $Y|\mathbf{X}$. Hence, the estimation of the central subspace becomes the main interest of the SDR theory. We will focus on extension of Li [1991]'s dimension reduction method in this dissertation, which is a simple and effective approach to obtain the central subspace $\mathbf{S}_{Y|\mathbf{X}}$.

2.1.1 Sliced inverse regression

Li [1991] introduced Sliced Inverse Regression (SIR) as an effective dimension reduction method of estimating the central subspace. Unlike many traditional statistical methods where the response variable Y is regressed against the predictor $\mathbf{X}$, SIR considers the opposite way by regressing $\mathbf{X}$ against Y. As Y varies, $E(\mathbf{X}|Y)$ forms an inverse regression curve centered at $E(\mathbf{X})$, and the main idea of SIR is to investigate the trajectory of the inverse regression curve. Before we describe the method in detail, it is necessary to introduce the following condition which is a fundamental probabilistic assumption required by SIR.

Condition 1 *(Condition 3.1; Li [1991]) For any b in $\mathbb{R}^p$, the conditional expectation $E(b\mathbf{X}|\beta_1\mathbf{X},\ldots,\beta_d\mathbf{X})$ is linear in $\beta_1\mathbf{X},\ldots,\beta_d\mathbf{X}$; that is, there exist some constants $c_0, c_1, \ldots, c_d$, $E(b\mathbf{X}|\beta_1\mathbf{X},\ldots,\beta_d\mathbf{X}) = c_0 + c_1\beta_1\mathbf{X} + \cdots + c_d\beta_d\mathbf{X}$.*

Condition 1 is commonly known as the linearity condition, which is satisfied when the distribution of $\mathbf{X}$ is elliptically symmetric, for example, the normal distribution [Li, 1991]. It is not a severe restriction because most low dimensional projections of high dimensional data are close to normal [Hall and Li, 1993].

Theorem 2.1.1 *(Theorem 3.1; Li [1991]) Under Model 2.1 and Condition 1 the centered inverse regression curve $E(\mathbf{X}|Y) - E(\mathbf{X})$ is contained in the linear subspace spanned by $\beta_i\Sigma_{\mathbf{X}}(i = 1,\ldots,d)$.*

Corollary 1 *(Corollary 3.1; Li [1991]) Under Model 2.2 and Condition 1 the standardized inverse regression curve $E(\mathbf{Z}|Y)$ is contained in the linear subspace generated by the standardized dimension reduction directions $\eta_i(i = 1,\ldots,d)$.*

Corollary 1 indicates that any vector that is orthogonal to the space spanned by $\eta_1,\ldots,\eta_d$ is a degenerate direction for $cov(E(\mathbf{Z}|Y))$. Therefore, we can use the eigenvectors of the covariance matrix $cov(E(\mathbf{Z}|Y))$ to estimate the standardized dimension reduction directions.

Remark 1 *Corollary 1 implies that eigenvectors of $cov(E(\mathbf{Z}|Y))$ is contained in the central subspace $\mathbf{S}_{Y|\mathbf{Z}}$. As a result, we are able to obtain a proper subset of the central subspace, which still yields important information about predicting the response Y. Cook [2004] further assumed a coverage condition so that the subspace spanned by $E(\mathbf{Z}|Y)$ was equivalent to $\mathbf{S}_{Y|\mathbf{Z}}$. Cook [2004] also pointed out that this condition was common in regression analysis based on SIR.*

The implementation of the SIR algorithm on the standardized $\mathbf{Z}$ scale is as follows.

1. *Obtain the standardized $\mathbf{Z}$ scale such that $\mathbf{Z} = \hat{\Sigma}_{\mathbf{x}}^{-1/2}(\mathbf{X} - \bar{\mathbf{X}})$, where $\hat{\Sigma}_{\mathbf{x}}$ is the sample covariance matrix of $\mathbf{X}$.*

2. *Divide range of Y into H slices as equally as possible.* Let n_h be the number of observations in slice h.

3. *Within each slice, compute the sample mean of $\mathbf{Z}$, which is denoted as $\bar{\mathbf{Z}}_h = n_h^{-1} \sum_{(i) \in slice_h} \mathbf{Z}_{(i)}$.*

4. *Compute the weighted covariance matrix denoted as $\hat{\gamma}$ for the slice means of $\mathbf{Z}$, whose weights are determined by the slice sizes:*

$$\hat{\gamma} = n^{-1} \sum_{h=1}^{H} n_h \bar{\mathbf{Z}}_h \bar{\mathbf{Z}}_h^T.$$

5. *Conduct the eigenvalue decomposition of $\hat{\gamma}$, and record its eigenvalues and eigenvectors.*

6. *The d eigenvectors associated with the largest d eigenvalues are the solution to the standardized dimension reduction directions, $\hat{\eta}_1, \ldots, \hat{\eta}_d$. The estimates of the dimension reduction directions $\beta_1, \ldots, \beta_d$ can then be obtained by transforming them back to the $\mathbf{X}$ scale, i.e., $\hat{\beta}_i = \hat{\Sigma}_{\mathbf{x}}^{-1/2} \hat{\eta}_i, i = 1, \ldots, d$.*

Theorem 2.1.1 also implies that we can obtain the estimates of the dimension reduction directions directly instead of through the standardized $\mathbf{Z}$ scale. Since $cov(E(\mathbf{X}|Y))$ is degenerate in any direction that is orthogonal to $\beta_i \Sigma_{\mathbf{x}}, i = 1, \ldots, d$, we can solve the following generalized eigenvalue decomposition problem,

$$\hat{\Gamma} \beta_i = \lambda_i \hat{\Sigma}_{\mathbf{x}} \beta_i, \quad i = 1, \ldots, d \quad , \tag{2.4}$$

where $\hat{\Gamma} = n^{-1} \sum_{y=1}^{h} n_y (\bar{\mathbf{X}}_y - \bar{\mathbf{X}})(\bar{\mathbf{X}}_y - \bar{\mathbf{X}})^T$ is the empirical covariance matrix for the slice means of $\mathbf{X}$ weighted by the slice sizes, $\lambda_i, i = 1, \ldots, d$ are the largest d eigenvalues of $\hat{\Gamma}$ relative to $\hat{\Sigma}_{\mathbf{x}}$, and $\beta_i, i = 1, \ldots, d$ are the d eigenvectors associated with $\lambda_i, i = 1, \ldots, d$.

2.1.2 High dimensional cases

SIR was originally designed to work only for the classical data format when the sample size n is larger than the predictor dimension p. Specifically, SIR uses sample covariance matrices and requires standardization of the predictors $\mathbf{X}$. In case of high dimensional analysis when the sample size n is less than the dimension p, the sample covariance matrix of $\mathbf{X}$ as a $p \times p$ high dimensional matrix is not of full rank and neither a consistent estimator of the true population covariance matrix of $\mathbf{X}$. Recently, work has been done to overcome the difficulties of SIR for high dimensional analysis. Zhong et al. [2005] proposed a regularized SIR (RSIR) approach by shrinking the sample covariance matrix of $\mathbf{X}$ towards a $p \times p$ identity matrix. This new covariance estimate of $\mathbf{X}$ becomes nonsingular, which is then used to substitute the sample covariance matrix of $\mathbf{X}$ to solve the generalized eigenvalue decomposition problem. Li and Yin [2008] proposed another regularized SIR approach by solving a constrained optimization problem based on the least squares formulation of SIR. Specifically, their objective function is a combination of an equivalent form of the least squares function derived from the original predictor scale and a L_2 type penalty term, and the basis estimates of the central subspace are obtained through alternating least squares. An L_1 type penalty term is further incorporated with the L_2 type penalty term for the purpose of achieving basis estimates and variable selection simultaneously. More discussions about these two methods can be found in Section 2.2. Wu et al. [2008] proposed a nonlinear SIR (kSIR) method by using kernel methods and regularization techniques. Essentially the kSIR method maps the original predictor space to a possibly infinite-dimensional Hilbert space by exploiting a Mercer kernel. The original SIR method by Li [1991] can then be applied on the mapped predictors to obtain the dimension reduction directions. In addition, they also regularize the sample covariance matrix of the mapped predictors towards the identity matrix due to a possibly ill-conditioning problem in practice.

2.1.3 Basic ideas of our proposed method

Motivated by Cook [2004] and Li and Yin [2008], our proposed method is derived on the basis of least squares formulation of SIR and we estimate the dimension reduction directions through alternating least squares. Bernard-Michel et al. [2008] showed that Li and Yin [2008]'s regularized SIR estimate was either zero or did not exist in practice. To overcome the issue, we start from the least squares function on the original $\mathbf{X}$ scale instead of the standardized $\mathbf{Z}$ scale proposed by Cook [2004]. Moreover, we modify the L_2 type penalty term as suggested by Bernard-Michel et al. [2008]. We will introduce our proposed method in detail in the next section.

Remark 2 *For a common research problem there are often multiple approaches for a solution. In many cases, one approach is more general than the others in terms of fewer assumptions, so that the approach can be applied in an extended scenario. What we are doing here is one of the examples that demonstrates the general applicability of the least squares method. Specifically, in situations where the sample size n is larger than the variable dimension p, our proposed alternating least squares method is equivalent to Zhong et al. [2005]'s RSIR. However, when the sample size n is less than the variable dimension p, our method is more general than Zhong et al. [2005]'s RSIR, because we overcome the limitation that the regularized covariance estimate is a poor estimate of the population covariance matrix.*

Remark 3 *The equivalence between the alternating least squares method and the original SIR [Li, 1991] is under a good condition that the covariance matrices, both the original data covariance matrix and the conditional covariance matrix, are well-conditioned and well-estimated. This is not the case when we have $n < p$. We claim that the alternating least squares method is more general. In fact, the matrix decomposition method cannot be applied when we have small sample sizes, because the standard estimate of a population covariance matrix is not consistent. The proposed alternating least squares method is able to avoid the estimation of the conditional co-*

variance matrix, but use iterative least squares regression for the bases of the reduced dimension space.

2.2 Extended sliced inverse regression for high-dimensional data

In this section, we start with a discussion of extending SIR for high dimensional but low sample size data through matrix decomposition. We then review the least squares formulation of SIR proposed by Cook [2004] and introduce our proposed method in detail. After that, we discuss strategies of selecting two important parameters for our proposed method, including the tuning parameter and the structural dimension.

2.2.1 Matrix decomposition methods for high dimensional data

Recall that the central subspace $\mathbf{S}_{Y|\mathbf{X}}$ is estimated by SIR through matrix decomposition. Formula (2.4) shows that the estimated dimension reduction directions correspond to the eigenvectors associated with the largest d eigenvalues of $\hat{\Gamma}$ relative to $\hat{\Sigma}_{\mathbf{x}}$, where $\hat{\Gamma}$ is the sample covariance matrix for the slice means of $\mathbf{X}$ weighted by the slice sizes, and $\hat{\Sigma}_{\mathbf{x}}$ is the sample covariance matrix of $\mathbf{X}$. For a classical data format when $n > p$, the sample covariance matrix is often used to estimate the population covariance matrix. It is shown that the sample covariance matrix is a consistent estimate of the population covariance matrix and has an optimal convergence rate of $n^{-1/2}$ when the dimension p is fixed and does not depend on the sample size n. However, the sample covariance matrix has several undesirable properties when p is large.

1. *As mentioned before, the sample covariance matrix is not of full rank when $n < p$, thus its inverse does not exist.*

2. *Even if the sample covariance matrix is invertible, the expected value of its inverse is a biased estimate for the theoretical inverse.*

3. *Unless $p/n \to 0$, the eigenvalues of the sample covariance matrix are more spread out than the population eigenvalues, even asymptotically [Johnstone, 2001].* Consider a simple case where samples of size n are obtained from a p-dimensional multivariate normal distribution with the mean vector $\vec{\mu}$ and the population covariance matrix as a $p \times p$ identity matrix $\mathbf{I}$. Marĉenko and Pastur [1967] showed that if $p/n \to c \in (0, 1)$, then the empirical distribution of the eigenvalues of the sample covariance matrix is supported on $((1 - \sqrt{c})^2, (1 + \sqrt{c})^2)$. Thus the larger p/n, the more spread out the eigenvalues.

4. *The sample eigenvectors are not consistent when p is large [Johnstone and Lu, 2004].*

To overcome the limitation of the sample covariance matrix for high dimensional but low sample size data, one possible approach of extending SIR is to find a better estimate of the population covariance matrix. Ledoit and Wolf [2004] proposed a well-conditioned estimator of the population covariance matrix when the dimension p was relatively large, where a well-conditioned estimator was referred to that the operation of its matrix inversion did not amplify the estimation error. Specifically, the estimator Σ_{Ledoit} has the following form,

$$\Sigma_{Ledoit} = \rho_1 \mathbf{I} + \rho_2 \mathbf{S}, \tag{2.5}$$

where $\mathbf{I}$ is the $p \times p$ identity matrix, $\mathbf{S}$ is the $p \times p$ sample covariance matrix, and ρ_1 and ρ_2 are two positive parameters which control the amount of shrinkage of the sample covariance matrix towards the identity matrix. Not only is this estimator invertible when $n < p$, but also it is more accurate than the sample covariance matrix asymptotically. It is also shown in Ledoit and Wolf [2004] that this estimator is an optimal convex linear combination of the identity matrix and the sample covariance matrix in terms of the quadratic loss when both n and p go to infinity. Zhong et al. [2005] developed a regularized SIR (RSIR) method to overcome the singularity issue.

Specifically, they replaced the sample covariance matrix $\hat{\Sigma}_{\mathbf{x}}$ with $\hat{\Sigma}_{\mathbf{x}} + s\mathbf{I}$, where s is a prescribed nonnegative regularization parameter. It is actually a special case of Ledoit and Wolf [2004]'s well conditioned estimator if $\rho_1 = s$ and $\rho_2 = 1$ in (2.5).

When the sample covariance matrix $\hat{\Sigma}_{\mathbf{x}}$ is nonsingular, (2.4) can be modified to the usual eigenvalue problem

$$\hat{\Sigma}_{\mathbf{x}}^{-1}\hat{\Gamma}\beta_i = \lambda_i\beta_i, \tag{2.6}$$

and this equivalent problem can be solved by many highly efficient and stable numerical linear algebra algorithms. However, $\hat{\Sigma}_{\mathbf{x}}$ is indeed singular in the context of high dimensional but low sample size data. Under such circumstances, the modification from (2.4) to (2.6) is impossible and there is actually not a complete set of the eigenvalues for the original problem. In some cases, the missing eigenvalues are treated as "infinite". In other cases, the entire problem may be considered as poorly posed. In order to be able to solve this problem numerically, people usually add some perturbation to $\hat{\Sigma}_{\mathbf{x}}$ to make it nonsingular and replace the original generalized eigenvalue problem with a nearby well posed problem. Zhong et al. [2005]'s regularized SIR method is developed based on this technique. But how to choose the perturbation is not an easy task and often requires some deep understanding of the problem itself. For example, the common choice of the perturbation is to add $\lambda\mathbf{I}$ to $\hat{\Sigma}_{\mathbf{x}}$ where $\lambda \in \mathbb{R}$ is a parameter to control the degree of the perturbation. If a large λ is chosen, (2.4) becomes much easier to solve, but the computed eigenvalues are not reliable and far away from the original problem. On the contrary, a small λ may cause the computed eigenvalues grow unboundedly. Therefore, the optimal λ has to be found by a successive of numerical tests, which is computationally expensive. In analogy, how to appropriately balance between $\mathbf{I}$ and $\mathbf{S}$ in (2.5) is found to be difficult in practice.

In addition to the type of regularized matrix estimators of Ledoit and Wolf [2004], we can consider employing other estimation methods for a high dimensional covariance matrix. Bickel and Levina [2008] proposed to regularize the sample covariance matrix by hard thresholding, such that the resulting estimator satisfies $T_{hard}(\hat{\Sigma}_{\mathbf{x}}) = [\sigma_{ij}1\left(|\sigma_{ij}| \geq s\right)]$, where s denotes a positive hard thresholding param-

eter. Further, Rothman *et al.* [2010] proposed a class of generalized thresholding operators for large covariance matrices, including hard thresholding [Bickel and Levina, 2008], soft thresholding, SCAD [Fan and Li, 2001], and adaptive LASSO [Zou, 2006]. In particular, if we regularize the sample covariance matrix by soft thresholding, the corresponding estimator has the form $T_{soft}(\hat{\Sigma}_{\mathbf{x}}) = \left[sign(\sigma_{ij})\left(|\sigma_{ij}| - \lambda\right)_{+}\right]$, where λ denotes a positive soft thresholding parameter. Even though Rothman *et al.* [2010] have shown that the generalized thresholding of the sample covariance matrix has good theoretical properties, for example, an optimal rate of convergence can be achieved with respect to the spectral norm, it is not guaranteed that this covariance estimator is positive definite, which is desirable for the generalized eigenvalue decomposition. Rothman [2012] proposed a sparse positive definite covariance estimator by solving a convex optimization problem. However, our simulation study indicates that the computation of this covariance estimator is much slower than both the hard and soft thresholdings, and the performance of the estimated dimension reduction directions is merely comparable to that of hard or soft thresholdings.

The basis estimates of the central subspace $\mathbf{S}_{Y|\mathbf{X}}$ can be obtained by substituting the sample covariance matrix of $\mathbf{X}$ in (2.4) with any of these high dimensional covariance estimators. We have conducted a simple simulation study to compare the performances between our proposed method and extended SIR through matrix decomposition, which includes Zhong et al. [2005]'s regularized SIR and the extended SIR through hard and soft thresholding the sample covariance matrices. We assume that the predictors $\mathbf{X}$ follow i.i.d. standard normal distribution, and the response variable Y only linearly depends on the first four predictors of $\mathbf{X}$. The simulation is repeated for 50 times, and we compute the corresponding canonical correlations between the true and estimated projected directions for the first dimension reduction direction. Figures 2.4 and 2.5 in Section 2.4 illustrate the canonical correlations between the estimated and true projected directions to assess the performances of the extended SIR through matrix decomposition compared to our proposed method. It can be seen that Zhong et al. [2005]'s regularized SIR method works similarly as our

proposed method, whereas regularizing $\hat{\Sigma}_{\mathbf{x}}$ through hard or soft thresholding does not work well. We also notice that the regularization of $\hat{\Sigma}_{\mathbf{x}}$ through hard or soft thresholding fails to retain the positive definiteness in several complicated simulated examples, thus does not work for the extension of SIR for high dimensional but low sample size data.

2.2.2 Least squares formulation of SIR

Suppose that we have a response variable Y and a set of p predictors $\mathbf{X} = (X_1, \ldots, X_p)$, and there are n independently and identically distributed (i.i.d.) samples of $(\mathbf{X}, Y)$. Let $\mathbf{Z}$ denote the standardized scale of $\mathbf{X}$ such that $\mathbf{Z} = \Sigma_{\mathbf{x}}^{-1/2}(\mathbf{X} - E(\mathbf{X}))$.

For simplicity, let us first consider the situation when we have the predictors in the standardized scale $\mathbf{Z}$. In the procedure of SIR, we divide the data into h nonoverlapping slices according to the ordered values of Y. Let $\vec{Z}_{yj}$ denote the jth observation of $\mathbf{Z}$ in the yth slice, $y = 1, \ldots, h, j = 1, \ldots, n_y$, where n_y is the sample size of the yth slice. For each observation $\vec{Z}_{yj}$, it hovers around its population mean $E(\mathbf{Z}|y)$. Recall that the standardized inverse regression curve, i.e., $E(\mathbf{Z}|Y)$, falls into the central subspace $\mathbf{S}_{Y|\mathbf{Z}}$ spanned by the columns of the $p \times d$ basis matrix $\eta = (\eta_1, \ldots, \eta_d)$. This indicates that $E(\mathbf{Z}|y) = \eta\rho_y$, where ρ_y is a d-dimensional column vector of coefficients. Therefore, The following model holds,

$$\begin{aligned} \vec{Z}_{yj} &= E(\mathbf{Z}|y) + \vec{e}_{yj} \\ &= \eta\rho_y + \vec{e}_{yj} \quad , \end{aligned}$$

where $\vec{e}_{yj}$ is an error term with $E(\vec{e}_{yj}) = \vec{0}$. The problem of estimating the dimension reduction directions, η, can be formulated as a least squares problem, with a corresponding loss function as follows,

$$L_d(\mathbf{B}, C_y) = \sum_{y=1}^{h} \sum_{j=1}^{n_y} \|\vec{Z}_{yj} - \mathbf{B}C_y\|^2 \tag{2.7}$$

over $\mathbf{B} \in \mathbb{R}^{p\times d}$ and $\mathbf{C} = (C_1, \ldots, C_h) \in \mathbb{R}^{d\times h}$. The minimum solutions of $\mathbf{B}$ and $\mathbf{C}$ are the desired η and ρ, respectively. Moreover, minimizing (2.7) is equivalent to minimizing

$$\tilde{L}_d(\mathbf{B}, C_y) = \sum_{y=1}^{h} f_y \|\bar{Z}_y - \mathbf{B}C_y\|^2 \quad , \tag{2.8}$$

where $f_y = \dfrac{n_y}{n}$ denotes the weight/proportion of slice y, and $\bar{\mathbf{Z}}_y$ denotes the average of $\mathbf{Z}$ in the yth slice. In fact, we have

$$\begin{aligned} L_d(\mathbf{B}, C_y) &= \sum_{y=1}^{h}\sum_{j=1}^{n_y} \|\vec{Z}_{yj} - \mathbf{B}C_y\|^2 \\ &= \sum_{y=1}^{h}\sum_{j=1}^{n_y} \|(\vec{Z}_{yj} - \bar{Z}_y) + (\bar{Z}_y - \mathbf{B}C_y)\|^2 \\ &= \sum_{y=1}^{h}\sum_{j=1}^{n_y} \|\vec{Z}_{yj} - \bar{Z}_y\|^2 + \sum_{y=1}^{h} n_y \|\bar{Z}_y - \mathbf{B}C_y\|^2 \quad , \end{aligned}$$

where the first term is a constant and the second term is proportional to (2.8). The solution $\hat{\mathbf{B}}$ that minimizes (2.8) forms an estimate of the basis of $S_{Y|\mathbf{Z}}$.

However, as mentioned before, if the sample size n is less than the variable dimension p, we cannot work on the standardized $\mathbf{Z}$ scale since the standardization requires the inverse of $\Sigma_{\mathbf{x}}$ whose sample estimate $\hat{\Sigma}_{\mathbf{x}}$ becomes singular. Alternatively, Li and Yin [2008] provided the least-squares formulation of SIR in the original predictor $\mathbf{X}$ scale. Let $\beta = (\beta_1, \ldots, \beta_d)$ denote the $p \times d$ basis matrix of the central subspace $\mathbf{S}_{Y|\mathbf{X}}$, which is related to η as $\eta = \Sigma_{\mathbf{x}}^{1/2}\beta$. If we change the variable in (2.8) by substituting $\bar{Z}_y$ with $\hat{\Sigma}_{\mathbf{x}}^{-1/2}\bar{X}_y$ where $\bar{X}_y$ is the average of $\mathbf{X}$ in the yth slice, it leads to the following loss function on the original $\mathbf{X}$ scale,

$$G(\mathbf{A}, \mathbf{C}) = \sum_{y=1}^{h} \hat{f}_y\{(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - \hat{\Sigma}_{\mathbf{x}}\mathbf{A}C_y\}^T \times \hat{\Sigma}_{\mathbf{x}}^{-1}\{(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - \hat{\Sigma}_{\mathbf{x}}\mathbf{A}C_y\} \quad , \tag{2.9}$$

where $\bar{\mathbf{X}}$ denotes the global average of $\mathbf{X}$, $\mathbf{A} \in \mathbb{R}^{p\times d}$, and $\mathbf{C} = (C_1, \ldots, C_h) \in \mathbb{R}^{d\times h}$. The solution $\hat{\mathbf{A}}$ that minimizes (2.9) forms an estimate of the basis of $S_{Y|\mathbf{X}}$.

Further, Li and Yin [2008] derived a modified form of $G(\mathbf{A}, \mathbf{C})$:

$$\tilde{G}(\mathbf{A}, \mathbf{C}) = \sum_{y=1}^{h} \hat{f}_y \|(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - \hat{\Sigma}_{\mathbf{x}}\mathbf{A}C_y\|^2 \quad . \tag{2.10}$$

Li and Yin [2008] showed that $\tilde{G}(\mathbf{A}, \mathbf{C})$ and $G(\mathbf{A}, \mathbf{C})$ had the same minimum solution but $\tilde{G}(\mathbf{A}, \mathbf{C})$ is easier to examine than $G(\mathbf{A}, \mathbf{C})$, as $\tilde{G}(\mathbf{A}, \mathbf{C})$ has the inverse covariance matrix removed in its formulation.

2.2.3 Penalized alternating least squares method

Based on the least squares formulation of SIR, Li and Yin [2008] proposed an alternating least squares method to obtain the basis estimates of $\mathbf{S}_{Y|\mathbf{X}}$ when $n < p$. They added a L_2-type penalty $\tau vec(\mathbf{A})^T vec(\mathbf{A})$ to the least squares function (2.10) with τ as a tuning parameter and $vec(.)$ as an operator that stacks all columns of the matrix to a single vector. For a fixed τ, their algorithm alternates between minimizing $\mathbf{A}$ and $\mathbf{C}$ until convergence. However, we find that there exist several shortcomings in their method. First, Bernard-Michel et al. [2008] showed that Li and Yin [2008]'s estimator either does not exist or is zero in theory. Second, we argue that the equivalence between (2.9) and (2.10) does not necessarily hold when a penalty is added. This can be illustrated in the following simple example. Suppose $x_0 = argmin f_1(x) = argmin f_2(x)$, where f_1 represents the function G in (2.9) and f_2 represents the function $\tilde{G}$ in (2.10). After adding a penalty, e.g., $g(x)$, we generally do not have

$$argmin\{f_1(x) + g(x)\} = argmin\{f_2(x) + g(x)\} \quad .$$

In fact, penalty in terms of norm can be considered as constrained minimization, which is

$$min\{f_1(x)\} \quad \text{s.t.} \quad norm(x) < \tau \quad ,$$

and

$$min\{f_2(x)\} \quad \text{s.t.} \quad norm(x) < \tau \quad .$$

When τ is smaller than x_0, these two minima may not be the same. Third, it is found in our simulation study that the convergence rate of Li and Yin [2008]'s algorithm tends to be extremely slow.

Consequently, we go back to the original minimization function $G(\mathbf{A}, \mathbf{C})$ instead of the modified form $\tilde{G}(\mathbf{A}, \mathbf{C})$ and add a new penalty term as suggested by Bernard-Michel et al. [2008]. Specifically, our least squares objective function is

$$H_\tau(\mathbf{A}, \mathbf{C}) = G(\mathbf{A}, \mathbf{C}) + \tau \sum_{y=1}^{h} \hat{f}_y \|\mathbf{A}C_y\|^2 \quad . \tag{2.11}$$

Remark 4 *It is noted that H_τ is invariant to bijective transformations, i.e.,*

$$H_\tau(\mathbf{AM}, \mathbf{M}^{-1}\mathbf{C}) = H_\tau(\mathbf{A}, \mathbf{C}) \quad ,$$

for all regular $d \times d$ matrix $\mathbf{M}$.

For a fixed τ, we follow the alternating least-squares idea from Li and Yin [2008] and derive a new alternating least-squares algorithm for (2.11). Recall that

$$H_\tau(\mathbf{A}, \mathbf{C}) = \sum_{y=1}^{h} \hat{f}_y \{(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - \hat{\Sigma}_{\mathbf{x}} \mathbf{A} C_y\}^T \times \hat{\Sigma}_{\mathbf{x}}^{-1} \{(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - \hat{\Sigma}_{\mathbf{x}} \mathbf{A} C_y\} + \tau \sum_{y=1}^{h} \hat{f}_y \|\mathbf{A}C_y\|^2 \quad .$$

For the specific values of $\mathbf{A} = \mathbf{0}$ and $\mathbf{C} = \mathbf{0}$, we have

$$H_\tau(\mathbf{0}, \mathbf{0}) = \sum_{y=1}^{h} \hat{f}_y \{\bar{\mathbf{X}}_y - \bar{\mathbf{X}}\}^T \times \hat{\Sigma}_{\mathbf{x}}^{-1} \{\bar{\mathbf{X}}_y - \bar{\mathbf{X}}\} \quad .$$

Since $H_\tau(\mathbf{0}, \mathbf{0})$ does not involve with either $\mathbf{A}$ or $\mathbf{C}$, minimizing (2.11) is equivalent to $\min_{\mathbf{A},\mathbf{C}} \tilde{H}_\tau(\mathbf{A}, \mathbf{C})$ for a fixed τ, where $\tilde{H}_\tau(\mathbf{A}, \mathbf{C}) = H_\tau(\mathbf{A}, \mathbf{C}) - H_\tau(\mathbf{0}, \mathbf{0})$. We can easily

show that by considering the difference, $\tilde{H}_\tau(\mathbf{A},\mathbf{C})$, $\hat{\Sigma}_{\mathbf{x}}^{-1}$ disappears from $\tilde{H}_\tau(\mathbf{A},\mathbf{C})$. Applying Kronecker product operations, we can then rewrite $\tilde{H}_\tau(\mathbf{A},\mathbf{C})$ as follows.

$$
\begin{aligned}
\tilde{H}_\tau(\mathbf{A},\mathbf{C}) =& \sum_{y=1}^{h} \hat{f}_y C_y^T \mathbf{A}^T(\hat{\Sigma}_{\mathbf{x}}+\tau\mathbf{I}_p)\mathbf{A}C_y - 2\sum_{y=1}^{h}\hat{f}_y(\bar{\mathbf{X}}_y-\bar{\mathbf{X}})^T\mathbf{A}C_y \\
=& \sum_{y=1}^{h} \hat{f}_y\{(\hat{\Sigma}_{\mathbf{x}}+\tau\mathbf{I}_p)^{\frac{1}{2}}\mathbf{A}C_y\}^T\{(\hat{\Sigma}_{\mathbf{x}}+\tau\mathbf{I}_p)^{\frac{1}{2}}\mathbf{A}C_y)\} \\
& - 2\sum_{y=1}^{h}\hat{f}_y(\bar{\mathbf{X}}_y-\bar{\mathbf{X}})^T\mathbf{I}_p\mathbf{A}C_y \\
=& \sum_{y=1}^{h} \hat{f}_y\{(C_y^T\otimes(\hat{\Sigma}_{\mathbf{x}}+\tau\mathbf{I}_p)^{\frac{1}{2}})\mathrm{vec}(\mathbf{A})\}^T\{(C_y^T\otimes(\hat{\Sigma}_{\mathbf{x}}+\tau\mathbf{I}_p)^{\frac{1}{2}})\mathrm{vec}(\mathbf{A})\} \\
& - 2\sum_{y=1}^{h}\hat{f}_y(\bar{\mathbf{X}}_y-\bar{\mathbf{X}})^T(C_y^T\otimes\mathbf{I}_p)\mathrm{vec}(\mathbf{A}) \\
=& \sum_{y=1}^{h} \hat{f}_y\mathrm{vec}(\mathbf{A})^T(C_y\otimes(\hat{\Sigma}_{\mathbf{x}}+\tau\mathbf{I}_p)^{\frac{1}{2}})(C_y^T\otimes(\hat{\Sigma}_{\mathbf{x}}+\tau\mathbf{I}_p)^{\frac{1}{2}})\mathrm{vec}(\mathbf{A}) \\
& - 2\sum_{y=1}^{h}\hat{f}_y(\bar{\mathbf{X}}_y-\bar{\mathbf{X}})^T(C_y^T\otimes\mathbf{I}_p)\mathrm{vec}(\mathbf{A}) \\
=& \sum_{y=1}^{h} \hat{f}_y\mathrm{vec}(\mathbf{A})^T(C_yC_y^T\otimes(\hat{\Sigma}_{\mathbf{x}}+\tau\mathbf{I}_p))\mathrm{vec}(\mathbf{A}) \\
& - 2\sum_{y=1}^{h}\hat{f}_y\mathrm{vec}(\mathbf{A})^T(C_y\otimes\mathbf{I}_p)(\bar{\mathbf{X}}_y-\bar{\mathbf{X}}) \\
=& \mathrm{vec}(\mathbf{A})^T(\mathbf{C}\mathbf{D}_f\mathbf{C}^T\otimes(\hat{\Sigma}_{\mathbf{x}}+\tau\mathbf{I}_p))\mathrm{vec}(\mathbf{A}) - 2\mathrm{vec}(\mathbf{A})^T(\mathbf{C}\mathbf{D}_f\otimes\mathbf{I}_p)\tilde{\mathbf{Y}} \quad ,
\end{aligned}
$$

where $\otimes$ is the Kronecker product, $\tilde{\mathbf{Y}} = \mathrm{vec}(\bar{\mathbf{X}}_1-\bar{\mathbf{X}},\ldots,\bar{\mathbf{X}}_h-\bar{\mathbf{X}})$, and $\mathbf{D}_f = diag(\hat{f}_1,\ldots,\hat{f}_h)$.

As a result, given $\mathbf{A}$, the solution of $\mathbf{C}$ can be obtained as follows,

$$
\begin{aligned}
\hat{\mathbf{C}} &= (\hat{C}_1,\ldots,\hat{C}_h), \quad \text{with} \\
\hat{C}_y &= (\mathbf{A}^T(\hat{\Sigma}_{\mathbf{x}}+\tau\mathbf{I}_p)^2\mathbf{A})^{-1}\mathbf{A}^T(\hat{\Sigma}_{\mathbf{x}}+\tau\mathbf{I}_p)(\bar{\mathbf{X}}_y-\bar{\mathbf{X}}), \quad y=1,\ldots,h \quad .
\end{aligned}
$$

Furthermore, given $\mathbf{C}$, the solution of $\mathbf{A}$ is

$$
\mathrm{vec}(\mathbf{A}) = \{\mathbf{C}\mathbf{D}_f\mathbf{C}^T\otimes(\hat{\Sigma}_{\mathbf{x}}+\tau\mathbf{I}_p)\}^{-1}(\mathbf{C}\mathbf{D}_f\otimes\mathbf{I}_p)\tilde{\mathbf{Y}} \quad .
$$

We alternate between minimizing $\mathbf{A}$ and $\mathbf{C}$ until some convergence criterion is satisfied. The details of implementing our penalized alternating least squares method are shown as follows.

1. *Input an initial value of A denoted as A_0, the dimension d and the regularization parameter τ.*

2. *At the i^{th} iteration, $i = 1, \ldots$:*

   2.1 *Given A_{i-1}, update C_i as*

   $$\hat{C} = (\hat{C}_1, \ldots, \hat{C}_h), \quad \textit{with}$$
   $$\hat{C}_y = (A^T(\hat{\Sigma}_{\mathbf{x}} + \tau \mathbf{I}_p)^2 A)^{-1} A^T(\hat{\Sigma}_{\mathbf{x}} + \tau \mathbf{I}_p)(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}), \quad y = 1, \ldots, h \quad .$$

   2.2 *Given C_i, update A_i as*

   $$vec(A) = \{C\mathbf{D}_f C^T \otimes (\hat{\Sigma}_{\mathbf{x}} + \tau \mathbf{I}_p)\}^{-1}(C\mathbf{D}_f \otimes \mathbf{I}_p)\tilde{Y} \quad ,$$

   where $\otimes$ is the Kronecker product, $\tilde{Y} = vec(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}, \ldots, \bar{\mathbf{X}}_h - \bar{\mathbf{X}})$, and $\mathbf{D}_f = diag(\hat{f}_1, \ldots, \hat{f}_h)$.

3. *Repeat 2. until the difference between two successive objective function values is negligible.*

2.2.4 Tuning parameter selection

In this section we develop a strategy to select the tuning parameter τ for our new alternating least squares method. Recall that our least squares objective function is

$$H_\tau(\mathbf{A}, \mathbf{C}) = G(\mathbf{A}, \mathbf{C}) + \tau \sum_{y=1}^{h} f_y \|\mathbf{A}C_y\|^2 \quad .$$

Assume the theoretic $\Sigma_{\mathbf{x}}$ and $\Sigma_{\mathbf{x}}^{-1}$ are known and $\Sigma_{\mathbf{x}}^{-1}$ can be decomposed as $\Sigma_{\mathbf{x}}^{-1} = \mathbf{L}^T\mathbf{L}$. We can rewrite $H_\tau(\mathbf{A}, \mathbf{C})$ in the following form.

$$\begin{aligned} H_\tau(\mathbf{A}, \mathbf{C}) =& \sum_{y=1}^{h} f_y\{\mathbf{L}((\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - \Sigma_{\mathbf{x}}\mathbf{A}C_y)\}^T\{\mathbf{L}((\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - \Sigma_{\mathbf{x}}\mathbf{A}C_y)\} \\ &+ \tau\sum_{y=1}^{h} f_y\|\mathbf{A}C_y\|^2 \\ =& \sum_{y=1}^{h} f_y\|\mathbf{L}((\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - \Sigma_{\mathbf{x}}\mathbf{A}C_y)\|^2 + \tau\sum_{y=1}^{h} f_y\|\mathbf{A}C_y\|^2 \end{aligned}$$

Let $\zeta = (\zeta_1, \ldots, \zeta_h)$, where $\zeta_y = \sqrt{f_y}\mathbf{A}C_y, y = 1, \ldots, h$. We have

$$\begin{aligned} H_\tau(\mathbf{A}, \mathbf{C}) =& \sum_{y=1}^{h} f_y\|\mathbf{L}((\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - \Sigma_{\mathbf{x}}\mathbf{A}C_y)\|^2 + \tau\sum_{y=1}^{h}\zeta_y^T\zeta_y \\ =& \|\tilde{\mathbf{W}}^{1/2}(\mathbf{I}_h \otimes \mathbf{L})\tilde{Y} - (\mathbf{I}_h \otimes \mathbf{L}^{-T})vec(\zeta)\|^2 + \tau vec(\zeta)^T vec(\zeta) \quad , \end{aligned}$$

where $\otimes$ is the Kronecker product, $\tilde{Y} = vec(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}, \ldots, \bar{\mathbf{X}}_h - \bar{\mathbf{X}})$, $\tilde{\mathbf{W}}^{1/2} = \mathbf{D}_f^{1/2} \otimes \mathbf{I}_p$, $\mathbf{I}_h$ is a $h \times h$ identity matrix and $\mathbf{D}_f = diag(f_1, \ldots, f_h)$.

This formulation is analogous to a ridge regression. More specifically, $H_\tau(\mathbf{A}, \mathbf{C})$ can be considered as a loss function corresponding to a ridge regression, where the response variable is $\tilde{\mathbf{W}}^{1/2}(\mathbf{I}_h \otimes \mathbf{L})\tilde{Y}$, and the predictor matrix is $\mathbf{I}_h \otimes \mathbf{L}^{-T}$. According to Golub et al. [1979], a good ridge parameter $\hat{\lambda}$ for solving the ridge regression problem $\frac{1}{n}\|y - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2$ can be chosen according to certain criterion, e.g., the generalized cross validation (GCV), which is defined as

$$V(\lambda) = \frac{1}{n}\|(\mathbf{I} - \mathbf{M}(\lambda))y\|^2 / \left[\frac{1}{n}Trace(\mathbf{I} - \mathbf{M}(\lambda))\right]^2 ,$$

where $\mathbf{M}(\lambda) = \mathbf{X}(\mathbf{X}^T\mathbf{X} + n\lambda\mathbf{I})^{-1}\mathbf{X}^T$.

In particular, for our ridge regression problem, λ corresponds to $\frac{1}{n}\tau$, $\mathbf{M}(\lambda)$ corresponds to $\mathbf{S}(\tau)$, where $\mathbf{S}(\tau) = (\mathbf{I}_h \otimes \mathbf{L}^{-T})(\mathbf{I}_h \otimes \Sigma_{\mathbf{x}} + \tau\mathbf{I}_{ph})^{-1}(\mathbf{I}_h \otimes \mathbf{L}^{-1})$, and $\mathbf{I}_{ph}$ is a $ph \times ph$ identity matrix, and n corresponds to ph. The GCV criterion can be defined as

$$GCV(\tau) = \frac{\|(\mathbf{I}_{ph} - \mathbf{S}(\tau))\tilde{\mathbf{W}}^{1/2}(\mathbf{I}_h \otimes \mathbf{L})\tilde{Y}\|^2}{ph(1 - Trace(\mathbf{S}(\tau))/ph)^2} \quad .$$

If $\Sigma_{\mathbf{x}}$ is decomposed as $\Sigma_{\mathbf{x}} = \mathbf{L}^*\mathbf{L}^{*T}$, then $\mathbf{L}^* = \mathbf{L}^{-1}$ and $\mathbf{L}^{*T} = \mathbf{L}^{-T}$. Consequently, the GCV criterion can be modified as follows by simply replacing $\mathbf{L}$ with $\mathbf{L}^{*-1}$,

$$GCV(\tau) = \frac{\|(\mathbf{I}_{ph} - \mathbf{S}(\tau))\tilde{\mathbf{W}}^{1/2}(\mathbf{I}_h \otimes \mathbf{L}^{*-1})\tilde{Y}\|^2}{ph(1 - Trace(\mathbf{S}(\tau))/ph)^2} \quad ,$$

where $\mathbf{S}(\tau) = (\mathbf{I}_h \otimes \mathbf{L}^{*T})(\mathbf{I}_h \otimes \Sigma_{\mathbf{x}} + \tau\mathbf{I}_{ph})^{-1}(\mathbf{I}_h \otimes \mathbf{L}^*)$.

In analogy, AIC and BIC can also be used to choose our tuning parameter τ, which are defined as

$$AIC = ph\log(\|(\mathbf{I}_{ph} - \mathbf{S}(\tau))\tilde{\mathbf{W}}^{1/2}(\mathbf{I}_h \otimes \mathbf{L}^{*-1})\tilde{Y}\|^2/ph) + 2Trace(\mathbf{S}(\tau)) \quad ,$$
$$BIC = ph\log(\|(\mathbf{I}_{ph} - \mathbf{S}(\tau))\tilde{\mathbf{W}}^{1/2}(\mathbf{I}_h \otimes \mathbf{L}^{*-1})\tilde{Y}\|^2/ph) + Trace(\mathbf{S}(\tau))\log(ph) \quad .$$

In our application, we use the sample covariance matrix $\hat{\Sigma}_{\mathbf{x}}$ to estimate $\Sigma_{\mathbf{x}}$. We already know that the sample covariance matrix is a bad estimate for a large population covariance matrix, thus it is not desirable to directly use it in the matrix decomposition procedure for estimating the dimension reduction directions. On the other hand, the selection of the tuning parameter τ is a relatively minor issue for our proposed method, and we intend to use the sample covariance matrix as a simple covariance estimate here. If we conduct an eigenvalue decomposition on $\hat{\Sigma}_{\mathbf{x}}$ such that $\hat{\Sigma}_{\mathbf{x}} = \mathbf{Q}\Lambda\mathbf{Q}^T$, then $\mathbf{L}^*$ satisfies the form $\mathbf{L}^* = \mathbf{Q}\Lambda^{1/2}$. $\mathbf{L}^{*-1}$ can be obtained by computing the pseudo-inverse of $\mathbf{L}^*$.

2.2.5 Choice of dimension for the subspace

The dimension d of the central subspace $\mathbf{S}_{Y|\mathbf{X}}$ is the other important parameter to be defined besides the tuning parameter τ. In this section we will discuss two criteria for determining the dimension d. The dimension d is called the structural dimension of the central subspace in the SDR theory.

Let $\Gamma = cov(E(\mathbf{X}|Y))$ and $\hat{\Gamma}$ as its sample estimator. Because Γ is a $p \times p$ matrix of rank d, the smallest $p - d$ eigenvalues of Γ should be zero. As a result, we conduct the eigenvalue decomposition of $\hat{\Gamma}$ and the number of nonzero eigenvalues of $\hat{\Gamma}$, i.e.,

$\tilde{d}$, can be considered as a rough estimate of the dimension d. It is obvious that if we add an identity matrix $\mathbf{I}_p$ to $\hat{\Gamma}$, $\tilde{d}$ is also equal to the number of eigenvalues of $\hat{\Gamma} + \mathbf{I}_p$ which are greater than 1. Zhu et al. [2006] proposed a BIC-type criterion to obtain a more accurate estimate of d, which is defined as

$$G(k) = \frac{n}{2} \sum_{i=1+min(\tilde{d}+k)}^{p} (log\hat{\nu}_i + 1 - \hat{\nu}_i) - \frac{C_n k(2p-k+1)}{2} \quad , \tag{2.12}$$

where $k \in \{0, 1, \ldots, p-1\}$, $\hat{\nu}_1, \ldots, \hat{\nu}_{\tilde{d}}$ are the largest $\tilde{d}$ eigenvalues of $\hat{\Gamma} + \mathbf{I}_p$, and C_n is some penalty constant. The estimated dimension $\hat{d}$ is the one that maximizes (2.12) over $(0, 1, \ldots, p-1)$.

Alternatively, we propose a method to estimate the dimension of the subspace by measuring the canonical correlations between the h centered sliced means of $\mathbf{X}$, i.e., $\bar{X}_1 - \bar{X}, \ldots, \bar{X}_h - \bar{X}$, and $\hat{\beta}_i \hat{\Sigma}_{\mathbf{x}}, i = 1, \ldots, d$. Recall that Theorem 2.1.1 indicates that $E(\mathbf{X}|Y) - E(\mathbf{X})$ is a linear combination of $\beta_i \Sigma_{\mathbf{x}}, i = 1, \ldots, d$. If the dimension d is appropriately chosen, the largest canonical correlation between $E(\mathbf{X}|Y) - E(\mathbf{X})$ and $\beta_i \Sigma_{\mathbf{x}}, i = 1, \ldots, d$ should be high. In practice, we substitute $E(\mathbf{X}|Y) - E(\mathbf{X})$ and $\Sigma_{\mathbf{x}}$ with their sample estimates respectively. Specifically, let $m \in (1, \ldots, \tilde{d})$ denote a candidate structural dimension value. After obtaining the estimated dimension reduction directions $\hat{\beta}_1, \ldots, \hat{\beta}_m$ for a given m, we calculate the following statistic, denoted as T,

$$T = \rho_1 \left\{ (\bar{X}_1 - \bar{X}, \ldots, \bar{X}_h - \bar{X}), (\hat{\beta}_1 \hat{\Sigma}_{\mathbf{x}}, \ldots, \hat{\beta}_m \hat{\Sigma}_{\mathbf{x}}) \right\} \quad ,$$

where $\rho_1 \{.\}$ represents the first/maximum canonical correlation. Ideally, we prefer to choose the smallest possible dimension m with a large value of T. Thus we plot T against m to choose a proper dimension d.

2.3 Discussion of Properties

2.3.1 Local convergence

In this section, we want to show the algorithmatic convergence, i.e., the solution of our alternating least squares algorithm converges to the local minimum of our objective function $H_\tau(\mathbf{A}, \mathbf{C})$.

Suppose that we are now at the n^{th} iteration such that we have obtained $\mathbf{A}^{(n)}$ and $\mathbf{C}^{(n)}$. Based on our alternating least squares algorithm, we first update the solution of $\mathbf{C}$ at the $(n+1)^{th}$ iteration, i.e., $\mathbf{C}^{(n+1)}$. We know

$$H_\tau(\mathbf{A}^{(n)}, \mathbf{C}^{(n+1)}) \leq H_\tau(\mathbf{A}^{(n)}, \mathbf{C}^{(n)}) \quad , \tag{2.13}$$

since $\mathbf{C}^{(n+1)}$ is the least squares solution that minimizes $H_\tau(\mathbf{A}^{(n)}, \mathbf{C})$ for any $\mathbf{C}$. After updating $\mathbf{C}$ at the $(n+1)^{th}$ iteration, we then need to update the solution of $\mathbf{A}$ at the same iteration. Similarly, the following inequality holds.

$$H_\tau(\mathbf{A}^{(n+1)}, \mathbf{C}^{(n+1)}) \leq H_\tau(\mathbf{A}^{(n)}, \mathbf{C}^{(n+1)}) \quad , \tag{2.14}$$

since $\mathbf{A}^{(n+1)}$ is the solution that minimizes $H_\tau(\mathbf{A}, \mathbf{C}^{(n+1)})$ for any $\mathbf{A}$. By combining (2.13) and (2.14), we then have

$$H_\tau(\mathbf{A}^{(n+1)}, \mathbf{C}^{(n+1)}) \leq H_\tau(\mathbf{A}^{(n)}, \mathbf{C}^{(n)}) \quad . \tag{2.15}$$

Formula (2.15) suggests that the value of our objective function $H_\tau(\mathbf{A}, \mathbf{C})$ decreases with each iteration and the solution of $\mathbf{A}$ reaches the local minimum when the convergence criterion is satisfied.

2.3.2 Discussion of Statistical Property

Suppose that we work on the original $\mathbf{X}$ scale. Without loss of generality, we also assume that $\mathbf{X}$ are already centered such that $E(\mathbf{X}) = 0$. The data are divided into h nonoverlapping slices according to the ordered values of Y. Let $\vec{X}_{yj}$ denote the jth observation in the yth slice, $y = 1, \ldots, h, j = 1, \ldots, n_y$, where n_y is the sample size of

the yth slice. We have used $\mathbf{A} \in \mathbb{R}^{p\times d}$ to denote the true basis matrix of the central subspace $\mathbf{S}_{Y|\mathbf{X}}$ so that $\mathbf{S}_{Y|\mathbf{X}}$ is spanned by the columns of $\mathbf{A}$. Recall that based on the least squares formulation of SIR, we solve the following optimization problem up to a constraint:

$$(\hat{\mathbf{A}}, \hat{\mathbf{C}}) = argmin_{\mathbf{A},\mathbf{C}} \sum_{y=1}^{h} \sum_{j=1}^{n_y} \|\vec{X}_{yj} - \hat{\Sigma}_{\mathbf{x}} \mathbf{A} C_y\|^2 \quad , \tag{2.16}$$

where $\hat{\mathbf{A}}$ is the estimated $p \times d$ basis matrix of the central subspace $\mathbf{S}_{Y|\mathbf{X}}$ and $\hat{\mathbf{C}} = (\hat{C}_1, \ldots, \hat{C}_h)$ estimates the corresponding $d \times h$ coefficient matrix of $\mathbf{C} = (C_1, \ldots, C_h)$. According to Theorem 2.1.1, for a slice $y, y = 1, \ldots, h$, we have

$$E(\mathbf{X}|y) = \Sigma_{\mathbf{x}} \mathbf{A} C_y \quad , \tag{2.17}$$

where $\Sigma_{\mathbf{x}}$ denotes the population covariance matrix of $\mathbf{X}$.

Definition 2.3.1 *Let P denote a joint probability distribution model of the predictors $\mathbf{X}$ and the response Y. Suppose $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$ and are functions of $(\mathbf{X}, Y)$. We define a distance between $\mathbf{u}$ and $\mathbf{v}$ as*

$$d(\mathbf{u}, \mathbf{v}) = E\|\mathbf{u} - \mathbf{v}\|^2 \quad , \tag{2.18}$$

where $E(.)$ represents the expectation with respect to the joint distribution P.

We now consider the following minimization problem:

$$\min_{\xi \in \mathbf{S}_{Y|\mathbf{X}}} d(\mathbf{X}, \xi) \quad , \tag{2.19}$$

under the distance defined in (2.18). In the following we want to show that the objective function of least squares for the dimension reduction directions is related to a risk function of the estimated dimension reduction directions.

Lemma 3 *Under Definition 2.3.1, $E(\mathbf{X}|y)$ is an orthogonal projection of $\mathbf{X}$ onto the central subspace $\mathbf{S}_{Y|\mathbf{X}}$.*

Proof Formula (2.19) is equivalent to minimizing $E_Y E_{\mathbf{X}|Y}\|\mathbf{X}-\xi\|^2$ for any $\xi \in \mathbf{S}_{Y|\mathbf{X}}$. Obviously, the theoretical solution of ξ is $\xi_0 \equiv E(\mathbf{X}|y) = \Sigma_{\mathbf{x}}\mathbf{A}C_y$ as $E(\mathbf{X}|y) \in \mathbf{S}_{Y|\mathbf{X}}$ from Theorem 2.1.1. Therefore, $\xi_0 = E(\mathbf{X}|y)$ is an orthogonal projection of $\mathbf{X}$ on $\mathbf{S}_{Y|\mathbf{X}}$. ■

Based on (2.17), consider the following model

$$\vec{X}_{yj} = \Sigma_{\mathbf{x}}\mathbf{A}C_y + \vec{e}_{yj}, \quad j = 1, \ldots, n_y, \tag{2.20}$$

where $\vec{e}_{yj}$ represents the error term, with $E(\vec{e}_{yj}) = \vec{0}$. We further assume $Var(\vec{e}_{yj}) = \sigma^2 I_p$. Then, $E\|\vec{X}_{yj} - \Sigma_{\mathbf{x}}\mathbf{A}C_y\|^2 = p\sigma^2$.

Proposition 2.3.1 *For a given slice y, the following relationship holds,*

$$d(\vec{X}_{yj}, \xi) = d(\xi, \xi_0) + p\sigma^2 \quad ,$$

where ξ is any p-dimensional random column vector and $\xi_0 = \Sigma_{\mathbf{x}}\mathbf{A}C_y$, which is the projection of $\vec{X}_{yj}$ on $\mathbf{S}_{Y|\mathbf{X}}$.

Proof Clearly, we have $E\|\vec{X}_{yj} - \xi_0\|^2 = p\sigma^2$ and $E((\vec{X}_{yj} - \xi_0)^T(\xi - \xi_0)) = 0$ from Lemma 3. Thus,

$$\begin{aligned} d(\vec{X}_{yj}, \xi) &= E(\|\vec{X}_{yj} - \xi\|^2) \\ &= E(\|\vec{X}_{yj} - \xi_0 + \xi_0 - \xi\|^2) \\ &= E\|\xi - \xi_0\|^2 + E\|\vec{X}_{yj} - \xi_0\|^2 - 2E((\vec{X}_{yj} - \xi_0)^T(\xi - \xi_0)) \\ &= d(\xi, \xi_0) + p\sigma^2. \end{aligned}$$

■

Remark 5 *Proposition 2.3.1 indicates that the population least squares function is the risk function (the mean squared error of the estimate) under quadratic loss plus a constant. In application, since the distribution of P is unknown, we instead work on the empirical least squares function. Moreover, in the case of high dimensional but low sample size data, we need to further work on a constrained least squares problem, i.e., incorporating a L_2-type penalty term to obtain the solution.*

2.4 Applications

2.4.1 Simulated examples

In this section, we present five simulated examples to validate our proposed method. Denote the true dimension reduction directions as β, and the estimated dimension reduction directions as $\hat{\beta}$. Canonical correlations between the true and projected dimension reduction directions, i.e., $\mathbf{X}\beta$ and $\mathbf{X}\hat{\beta}$ are used as a measurement to assess the simulation performance. Both Example 1 and Example 2 assume that the predictors are independent from each other; the response variable Y has a linear relationship with the predictors $\mathbf{X}$ in Example 1 while the response model in Example 2 includes a main effect and an interaction term. Examples 3 and 4 assume that the predictors are correlated, and the response variable Y has a nonlinear relationship with the predictors $\mathbf{X}$. The predictors $\mathbf{X}$ are generated from a four factor model in Example 5, and the response model is a simple linear model.

Example 1 *The data are generated from the following model,*

$$Y = X_1 + X_2 + X_3 + X_4 + \sigma_0 \epsilon \quad ,$$

where both the predictors $X = (X_1, \ldots, X_p)^T$ *and the error term* ϵ *follow the standard normal distribution, and the parameter* σ_0 *defines a signal to noise ratio, which is set to be* 0.4*.*

The sample size n is chosen to be 100 and the predictor dimension p to be 500. The central subspace is spanned by $\beta_1 = (1, 1, 1, 1, 0, \ldots, 0)^T$, and the true structural dimension is $d = 1$. We implement our proposed method, Li and Yin [2008]'s method, and Zhong et al. [2005]'s regularized SIR method and then compare their performances of finding the dimension reduction directions. Figure 2.1 shows the coefficients $\hat{\beta}$'s for all predictors in one simulation. It indicates that all of the three methods have very similar estimation results. We also repeat the simulation 50 times, and compute the corresponding canonical correlations between the true and estimated

projection directions. Figure 2.2 shows the scatterplot matrix of canonical correlations for the first dimension reduction direction, where dots above the diagonal line indicate higher canonical correlations. Figure 2.3 shows the boxplot of these canonical correlations. We find that the result of our proposed method is comparable to that of Li and Yin [2008]. However, Li and Yin [2008]'s method suffers from an extremely slow convergence rate. Li and Yin [2008]'s method has not converged after 100 iteration steps. We do not think Li and Yin [2008]'s method is favorable and will skip the simulations for Li and Yin [2008]'s method for Examples 2-5. On the other hand, Zhong et al. [2005]'s regularized SIR method results in a few poor canonical correlations compared to the other two methods.

Example 2 *The data are generated from the following model,*

$$Y = X_1 + X_1 \times X_2 + \sigma_0 \epsilon \quad ,$$

where both the predictors $X = (X_1, \ldots, X_p)^T$ *and the error term* ϵ *follow the standard normal distribution, and the parameter* σ_0 *defines a signal to noise ratio, which is set to be* 0.4*.*

The sample size n is chosen to be 100 and the predictor dimension p to be 500. The central subspace is spanned by $\beta_1 = (1, 0, 0, \ldots, 0)^T$ and $\beta_2 = (0, 1, 0, \ldots, 0)^T$, and the true structural dimension is $d = 2$. We repeat the simulations 50 times and compute the corresponding canonical correlations between the true and estimated projection directions for both our proposed method and Zhong et al. [2005]'s regularized SIR. Figure 2.6 shows the scatterplot matrix of canonical correlations for the first and second dimension reduction directions, as well as the average canonical correlations of these two directions, where dots above the diagonal line indicate higher canonical correlations. Figure 2.7 shows the boxplot of these canonical correlations. The simulation results indicate that the performances of our proposed method and Zhong et al. [2005]'s regularized SIR method are comparable, with our method slightly better.

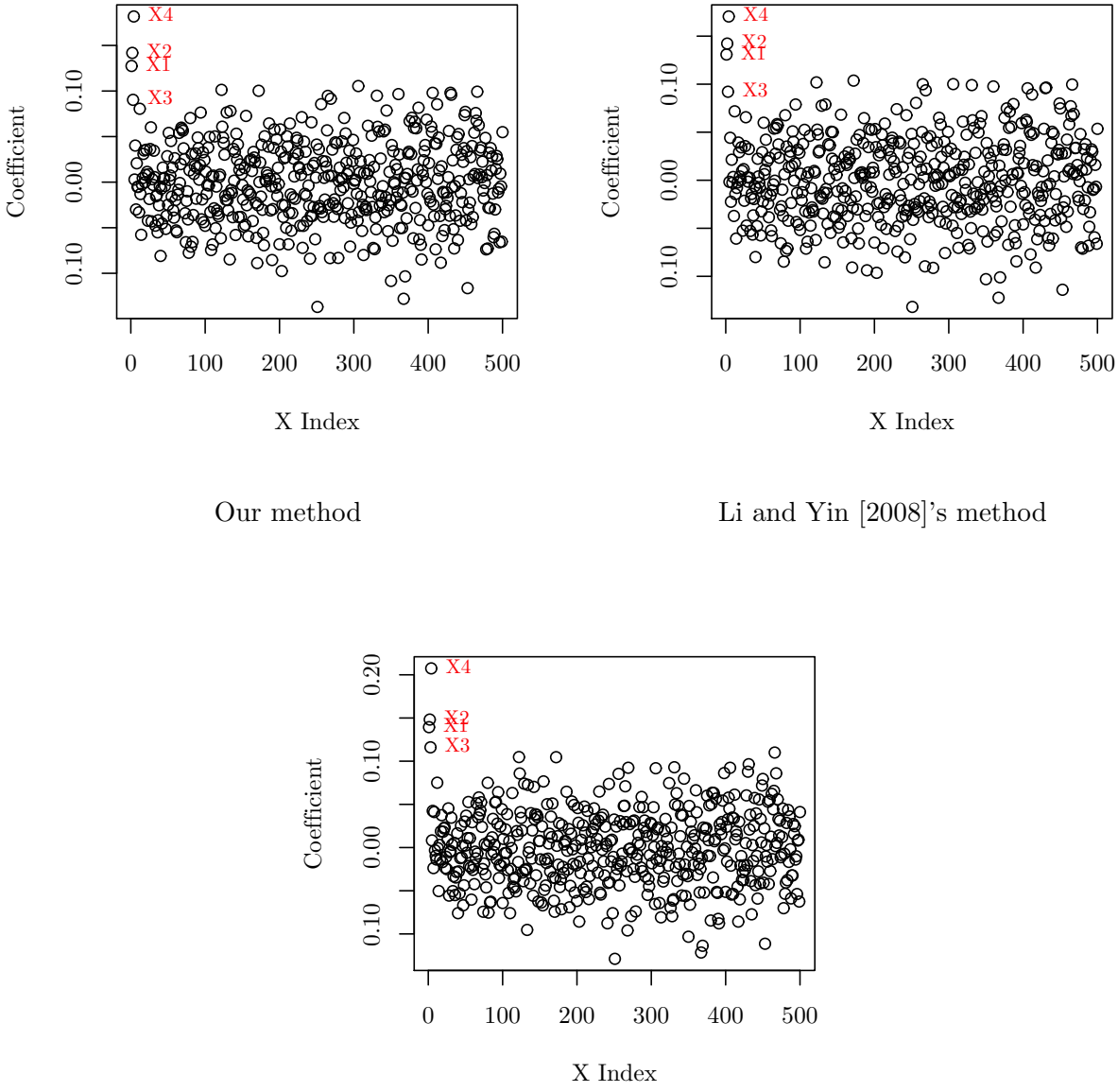

Figure 2.1. Comparison of coefficient estimates for the first dimension reduction direction in Example 1. The upper panel displays the estimates from our method (left) and Li and Yin [2008]'s method (right), and the lower panel illustrates the estimate from Zhong et al. [2005]'s RSIR.

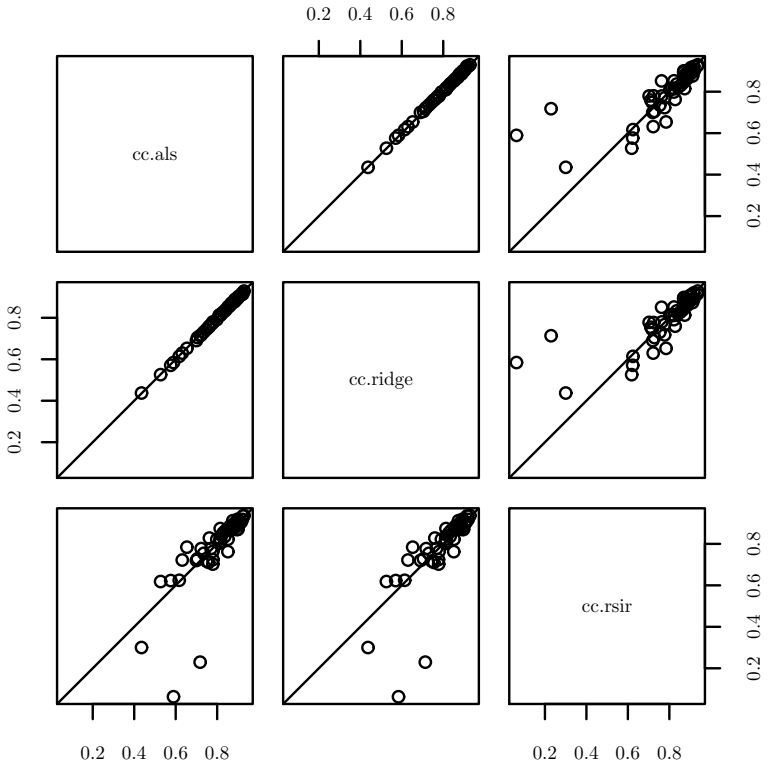

Figure 2.2. Scatterplot comparing canonical correlations between the estimated and true projected directions for Example 1, with the reference diagonal line passing through the origin. "cc.als" represents the canonical correlations from our alternating least squares method, "cc.ridge" represents the canonical correlations from Li and Yin [2008]'s method, and "cc.rsir" represents the canonical correlations from Zhong et al. [2005]'s RSIR method.

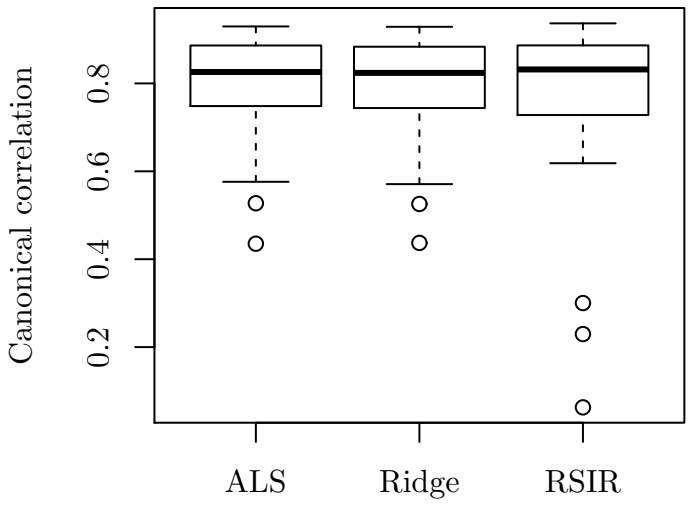

Figure 2.3. Boxplot comparing canonical correlations between the estimated and true projected directions for Example 1. "ALS" represents our alternating least squares method, "Ridge" represents Li and Yin [2008]'s method, and "RSIR" represents Zhong et al. [2005]'s regularized SIR method.

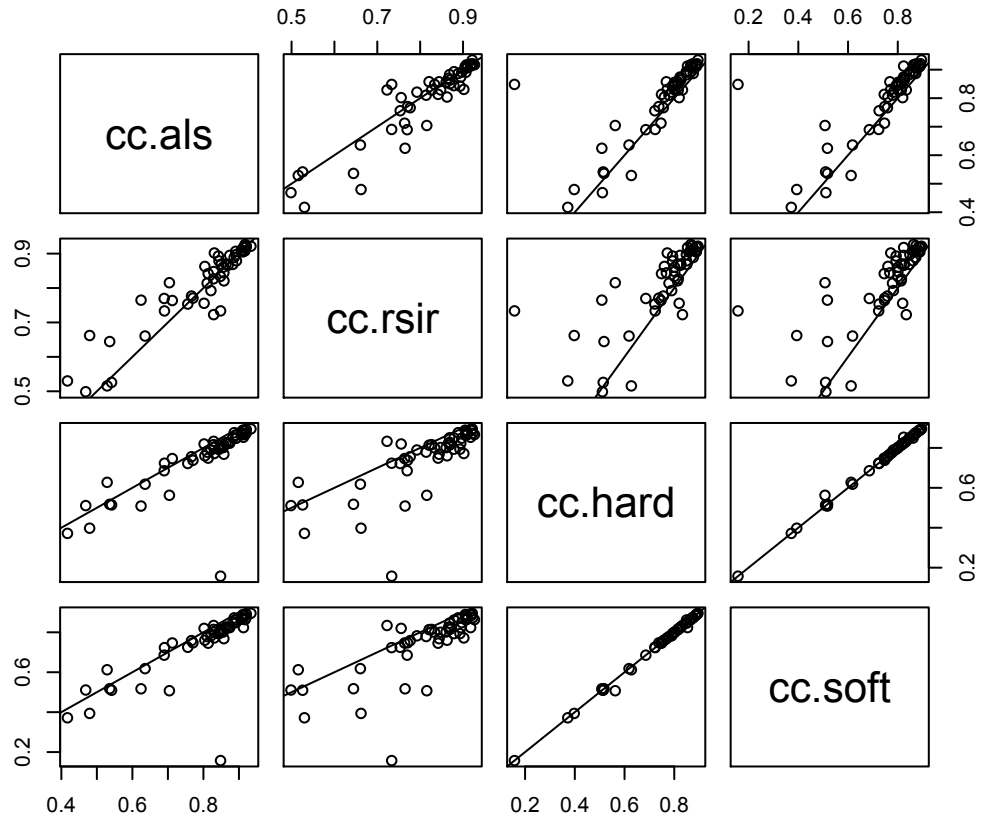

Figure 2.4. Scatterplot comparing canonical correlations between the estimated and true projected directions for Example 1. "cc.als" represents the canonical correlations from our alternating least squares method, "cc.rsir" represents the canonical correlations from RSIR [Zhong et al., 2005], "cc.hard" represents the canonical correlations based on the hard thresholding estimator of the covariance matrix, and "cc.soft" represents the canonical correlations based on the soft thresholding estimator of the covariance matrix.

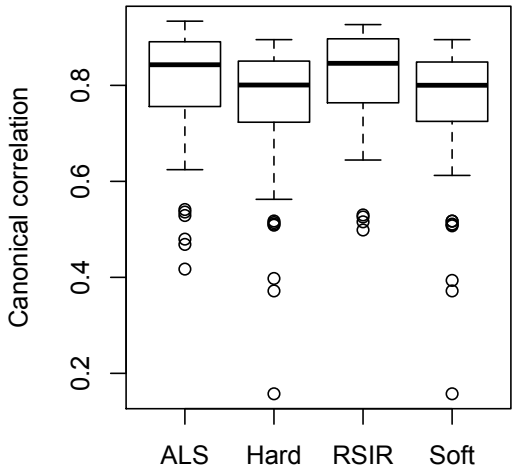

Figure 2.5. Boxplot comparing canonical correlations between the estimated and true projected directions for Example 1. "ALS" represents our alternating least squares method, "Hard" represents hard thresholding, "RSIR" represents Zhong et al. [2005]'s regularized SIR method, and "Soft" represents soft thresholding.

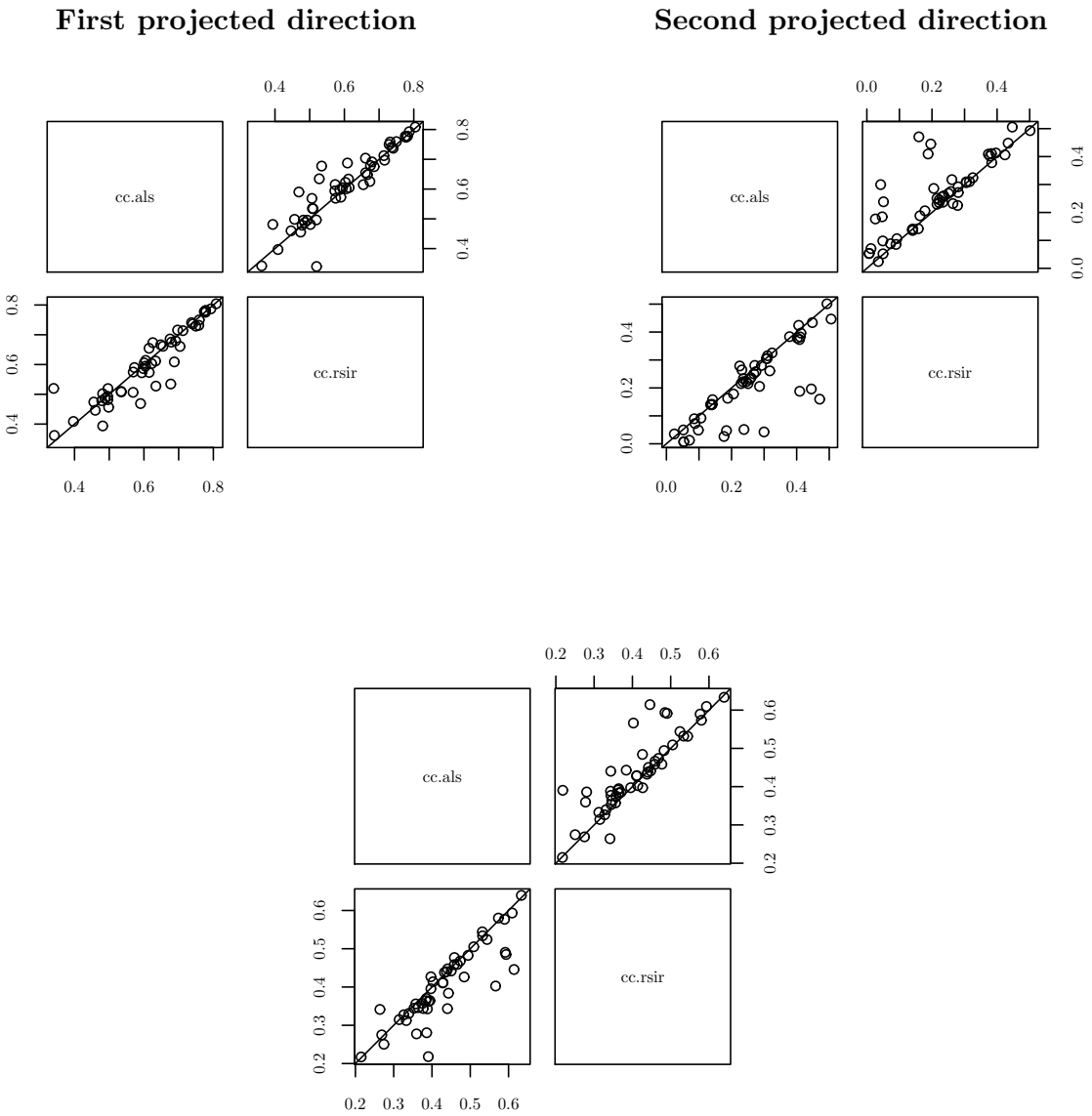

Figure 2.6. Scatterplot comparing canonical correlations between the estimated and true projected directions for Example 2. "cc.als" represents the canonical correlations from our alternating least squares method, and "cc.rsir" represents the corresponding canonical correlations from RSIR [Zhong et al., 2005].

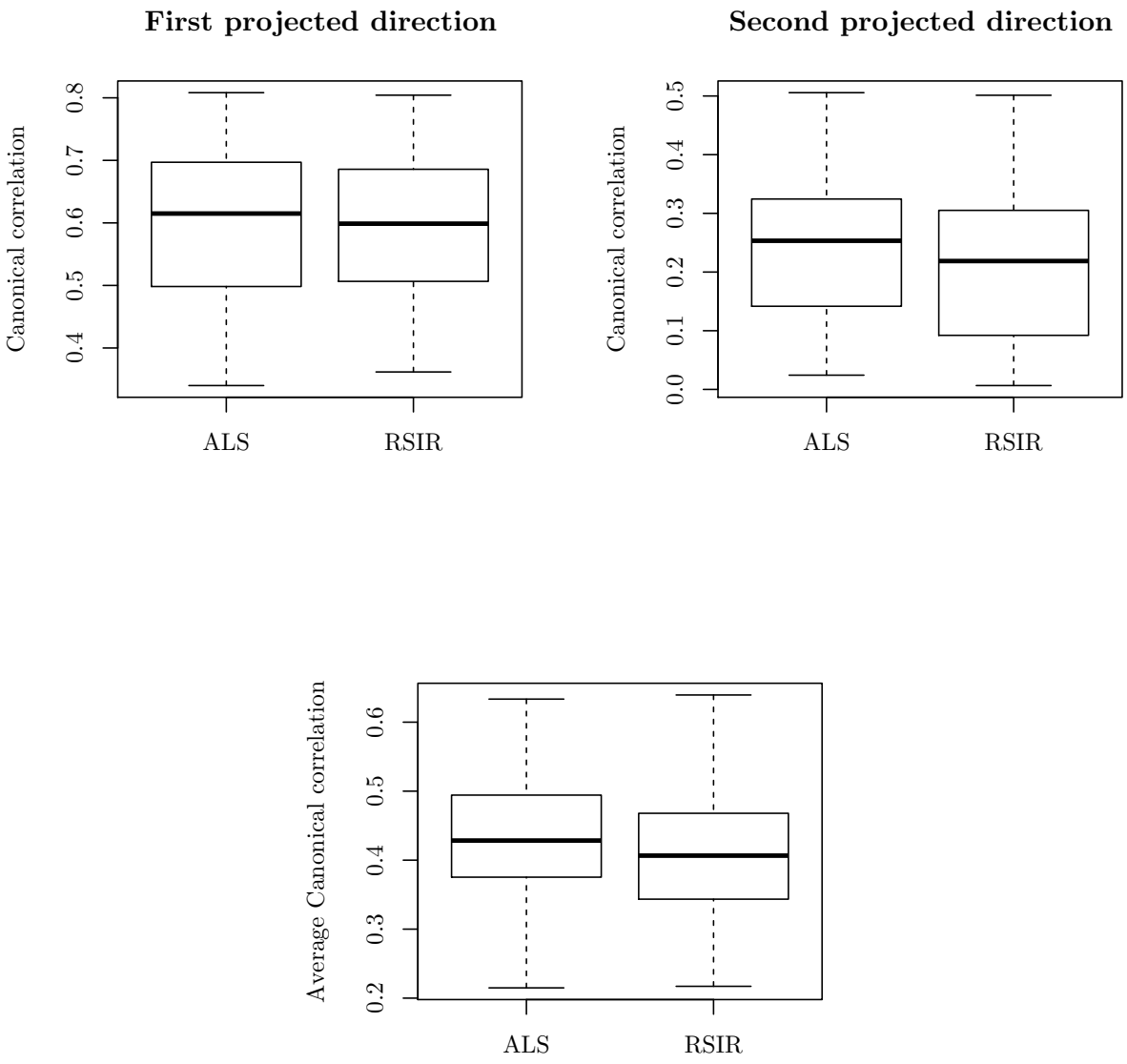

Figure 2.7. Boxplot comparing canonical correlations between the estimated and true projected directions for Example 2. "ALS" represents our alternating least squares method, and "RSIR" represents Zhong et al. [2005]'s regularized SIR method.

Example 3 *The data are generated from the following model,*

$$\beta_1 = \frac{(1,1,1,0,\ldots,0)^T}{\sqrt{3}} \quad , \qquad U = \beta_1^T \mathbf{X} \quad ,$$
$$Y = 2U + U^2 + \sigma_0 \epsilon \quad ,$$

where the predictors $X = (X_1, \ldots, X_p)^T$ *follow a multivariate normal distribution with zero mean and covariance matrix* $\Sigma_{\mathbf{x}} = \left\{\sigma^2 \rho^{|i-j|}\right\}_{i,j=1,\ldots,p}$*, the error term* ϵ *follows the standard normal distribution, and the parameter* σ_0 *defines a signal to noise ratio, which is set to be* 0.4*.*

The sample size n is chosen to be 100 and the predictor dimension p to be 500. The central subspace is spanned by $\beta_1 = (1,1,1,0,\ldots,0)^T$, and the true structural dimension is $d = 1$. We vary the value of ρ for different types of the covariance matrix. First, assume that $\sigma^2 = 1$ and $\rho = 0.5$, where the predictors are moderately correlated. Second, assume that $\sigma^2 = 1$ and $\rho = 0.8$, where the predictors are highly correlated. We repeat the simulations 50 times and compute the corresponding canonical correlations between the true and estimated projection directions for both our proposed method and Zhong et al. [2005]'s regularized SIR. Figures 2.8 and 2.9 illustrate the canonical correlations when the predictors are moderately correlated ($\rho = 0.5$), and the canonical correlations are visualized in Figures 2.10 and 2.11 when the predictors are highly correlated ($\rho = 0.8$). More specifically, Figures 2.8 and 2.10 show the scatterplot matrix of canonical correlations for the first dimension reduction direction, where dots above the diagonal line indicate higher canonical correlations, and Figures 2.9 and 2.11 show the boxplot of these canonical correlations. These plots indicate that the performance of our proposed method is similar to Zhong et al. [2005]'s regularized SIR method in Example 3.

Example 4 *The data are generated from the following model,*

$$\beta_1 = 10^{-1/2}(1,1,1,1,1,1,1,1,1,1,0,\ldots,0)^T,$$
$$\beta_2 = 10^{-1/2}(1,-1,1,-1,1,-1,1,-1,1,-1,0,\ldots,0)^T,$$
$$Y = \frac{\beta_1^T \mathbf{X}}{0.5 + (\beta_2^T \mathbf{X} + 1.5)^2} + \sigma_0 \epsilon,$$

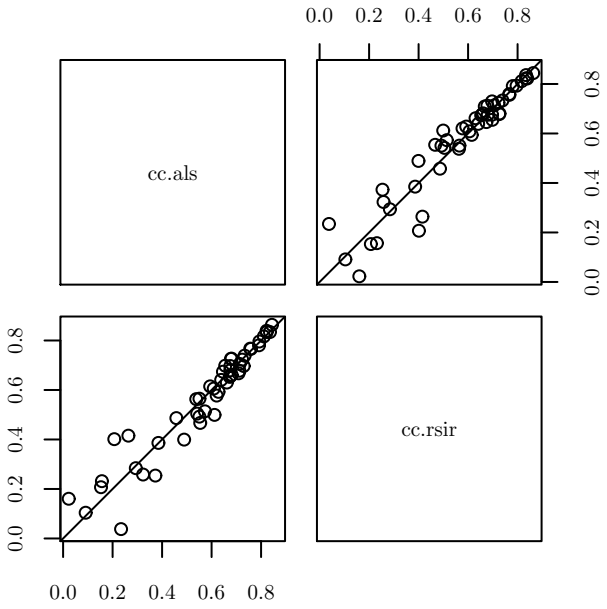

Figure 2.8. Scatterplot comparing canonical correlations between the estimated and true projected directions when the predictors are moderately correlated ($\rho = 0.5$) in Example 3. "cc.als" represents the canonical correlations from our alternating least squares method, and "cc.rsir" represents the corresponding canonical correlations from RSIR [Zhong et al., 2005].

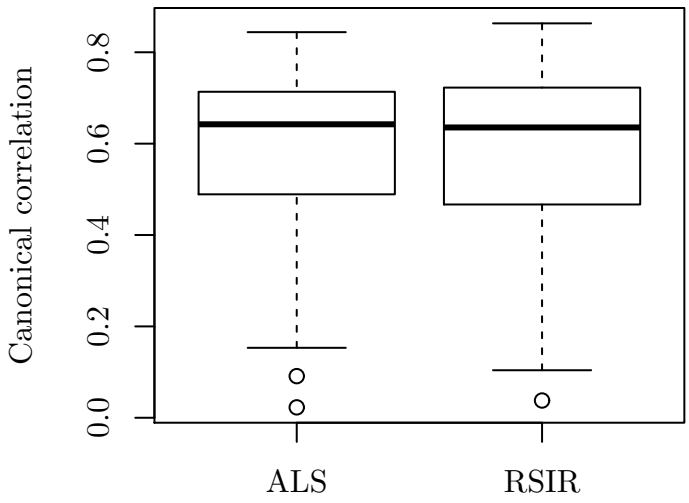

Figure 2.9. Boxplot comparing canonical correlations between the estimated and true projected directions when the predictors are highly correlated ($\rho = 0.5$) in Example 3. "ALS" represents our alternating least squares method, and "RSIR" represents Zhong et al. [2005]'s regularized SIR method.

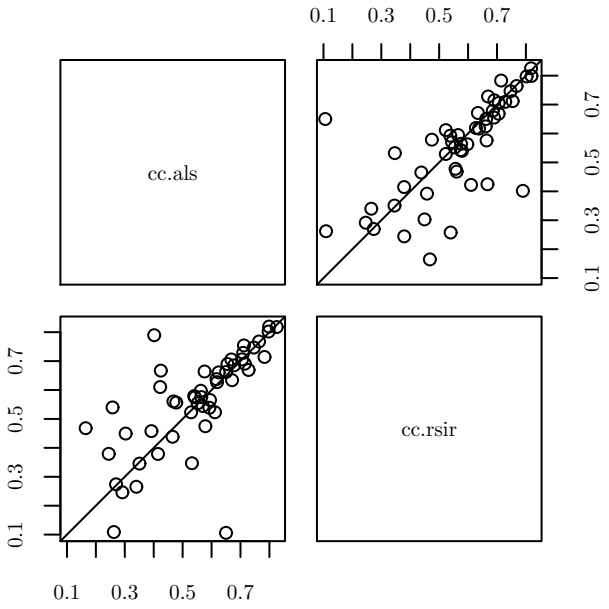

Figure 2.10. Scatterplot comparing canonical correlations between the estimated and true projected directions when the predictors are highly correlated ($\rho = 0.8$) in Example 3. "cc.als" represents the canonical correlations from our alternating least squares method, and "cc.rsir" represents the corresponding canonical correlations from RSIR [Zhong et al., 2005].

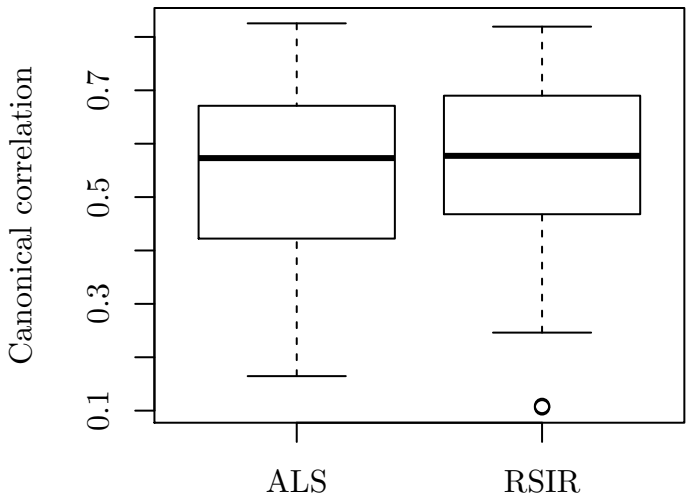

Figure 2.11. Boxplot comparing canonical correlations between the estimated and true projected directions when the predictors are moderately correlated ($\rho = 0.8$) in Example 3. "ALS" represents our alternating least squares method, and "RSIR" represents Zhong et al. [2005]'s regularized SIR method.

where the predictors $X = (X_1, \ldots, X_p)^T$ *follow a multivariate normal distribution with zero mean and covariance matrix* $\Sigma_{\mathbf{x}} = \left\{\sigma^2 \rho^{|i-j|}\right\}_{i,j=1,\ldots,p}$, *the error term* ϵ *follows the standard normal distribution, and the parameter* σ_0 *defines a signal to noise ratio, which is set to be* 0.5.

The sample size n is chosen to be 100 and the predictor dimension p to be 500. The central subspace is spanned by $\beta_1 = 10^{-1/2}(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, \ldots, 0)^T$ and $\beta_2 = 10^{-1/2}(1, -1, 1, -1, 1, -1, 1, -1, 1, -1, 0, \ldots, 0)^T$, and the true structural dimension is $d = 2$. Similar to the previous example, we assume $\sigma^2 = 1$ but consider two values for ρ, 0.5 and 0.8. We repeat the simulations 50 times and compute the corresponding canonical correlations between the true and estimated projection directions for both our proposed method and Zhong et al. [2005]'s regularized SIR. Figures 2.12 and 2.13 illustrate the canonical correlations when the predictors are moderately correlated ($\rho = 0.5$), and the canonical correlations are visualized in Figures 2.14 and 2.15 when the predictors are highly correlated ($\rho = 0.8$). More specifically, Figures 2.12 and 2.14 show the scatterplot matrix of average canonical correlations for the first and second dimension reduction directions, where dots above the diagonal line indicate higher canonical correlations. Figures 2.13 and 2.15 show the boxplot of these canonical correlations. These plots indicate that the performance of our proposed method is also similar to Zhong et al. [2005]'s regularized SIR method in Example 4.

Example 5 *The data are generated from the following model [Johnstone, 2006],*

$$X_{ij} = \sum_{\nu=1}^{4} b_{j\nu} f_{\nu i} + e_{ij}, \quad i = 1, \ldots, n, j = 1, \ldots, p,$$

$$Y = X_1 + X_2 + X_3 + X_4 + \sigma_0 \epsilon,$$

where $b_{j\nu} \sim N(0.6, 0.4^2)$, $f_{\nu i} \sim N(0, 0.01257^2)$, $e_{ij} \sim N(0, 0.0671^2)$, ϵ *is the i.i.d. error term following a standard normal distribution, and the parameter* σ_0 *defines a signal to noise ratio, which is set to be* 0.5.

The sample size n is chosen to be 80 and the predictor dimension p to be 100. The central subspace is spanned by $\beta_1 = (1, 1, 1, 1, 0, \ldots, 0)^T$, and the true structural

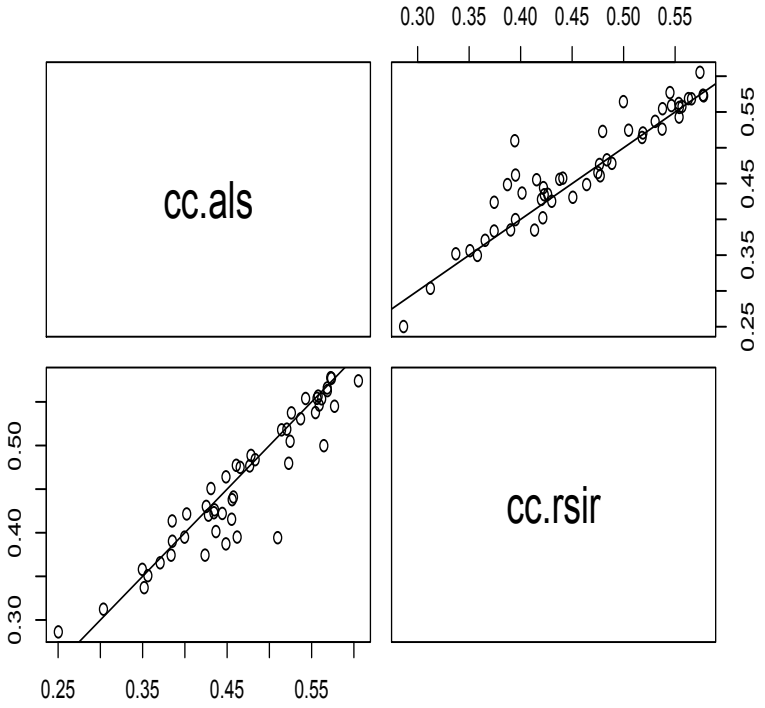

Figure 2.12. Scatterplot comparing canonical correlations between the estimated and true projected directions when the predictors are moderately correlated ($\rho = 0.5$) in Example 4. "cc.als" represents the canonical correlations from our alternating least squares method, and "cc.rsir" represents the corresponding canonical correlations from RSIR [Zhong et al., 2005].

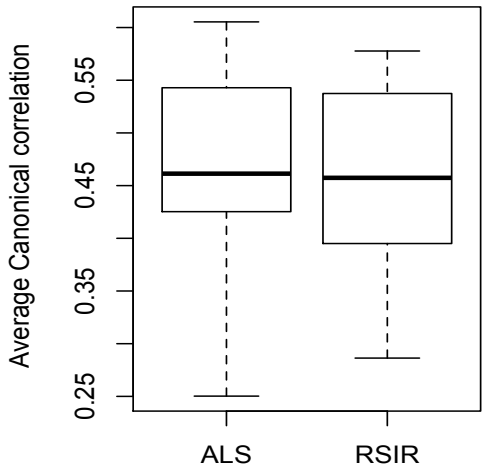

Figure 2.13. Boxplot comparing canonical correlations between the estimated and true projected directions when the predictors are moderately correlated ($\rho = 0.5$) in Example 4. "ALS" represents our alternating least squares method, and "RSIR" represents Zhong et al. [2005]'s regularized SIR method.

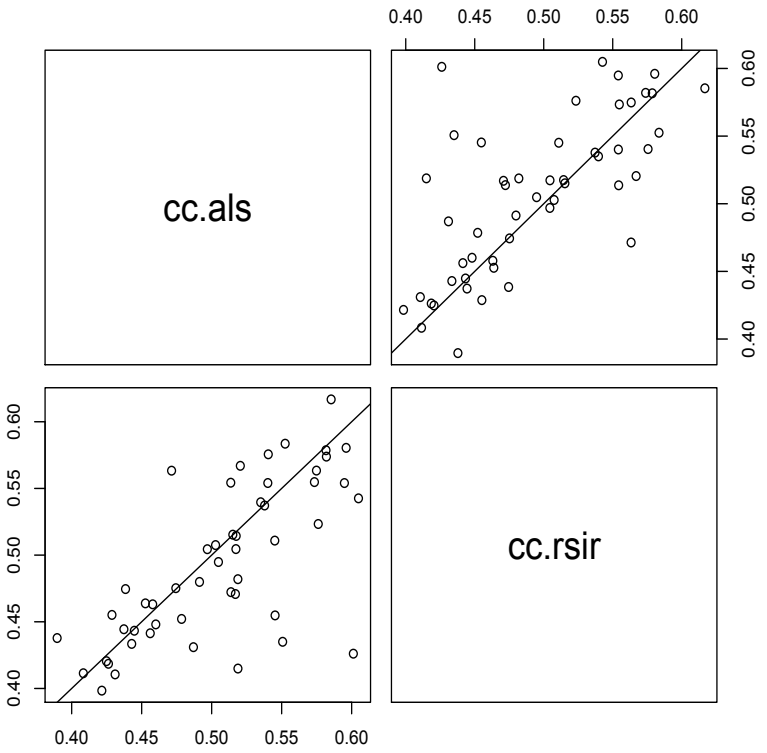

Figure 2.14. Scatterplot comparing canonical correlations between the estimated and true projected directions when the predictors are highly correlated ($\rho = 0.8$) in Example 4. "cc.als" represents the canonical correlations from our alternating least squares, and "cc.rsir" represents the corresponding canonical correlations from RSIR [Zhong et al., 2005].

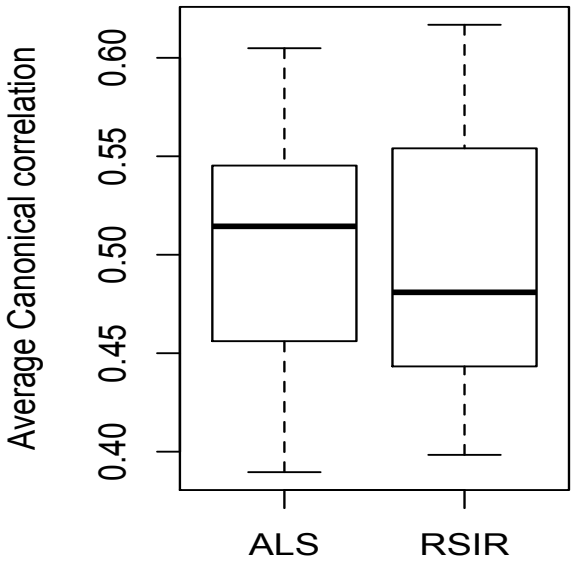

Figure 2.15. Boxplot comparing canonical correlations between the estimated and true projected directions when the predictors are highly correlated ($\rho = 0.8$) in Example 4. "ALS" represents our alternating least squares method, and "RSIR" represents Zhong et al. [2005]'s regularized SIR method.

dimension is $d = 1$. We repeat 50 simulations for both of our alternating least squares method and Zhong et al. [2005]'s regularized SIR, and compute the corresponding canonical correlations between the true and estimated projection directions for both our proposed method and Zhong et al. [2005]'s regularized SIR. Figure 2.16 shows the scatterplot of canonical correlations for the first dimension reduction direction, where dots above the diagonal line indicate higher canonical correlations, and Figure 2.17 shows the boxplot of these canonical correlations.

Figures 2.16 and 2.17 indicate that our proposed method outperforms Zhong et al. [2005]'s regularized SIR. It can be seen that in one of our simulations, the canonical correlation between the estimated and true projection directions is approximately 0.6 for our method, whereas it is around 0 for Zhong et al. [2005]'s regularized SIR. In this example the predictors $\mathbf{X}$ are constructed from a factor model of four independent factors as described in Johnstone [2006]. It is shown by Johnstone [2006] that there is a severe overestimation issue for the nonzero eigenvalues of the population covariance matrix of $\mathbf{X}$, thus the sample covariance matrix $\hat{\Sigma}_{\mathbf{x}}$ becomes a very bad estimate. Zhong et al. [2005]'s regularized SIR adds a simple perturbation ($s\mathbf{I}_p$) to this bad estimate to obtain a well conditioned covariance estimate. However, the unstable performance of Zhong et al. [2005]'s method indicates that the perturbation seems to be inappropriate for this example.

2.4.2 Application to the pharmacogenomics data

We implement our proposed alternating least squares method in the pharmacogenomics study of bortezomib in multiple myeloma [Mulligan *et al.*, 2007]. The data set has been used in Section 4 of Chapter 1, thus we will skip the detailed description here. Recall that the data set consists of a five-level clinical response ranging from progressive disease, no change, minimal response, partial response to complete response, $44,928$ gene expression values and other clinical features. There are 264 patients in four clinical trials.

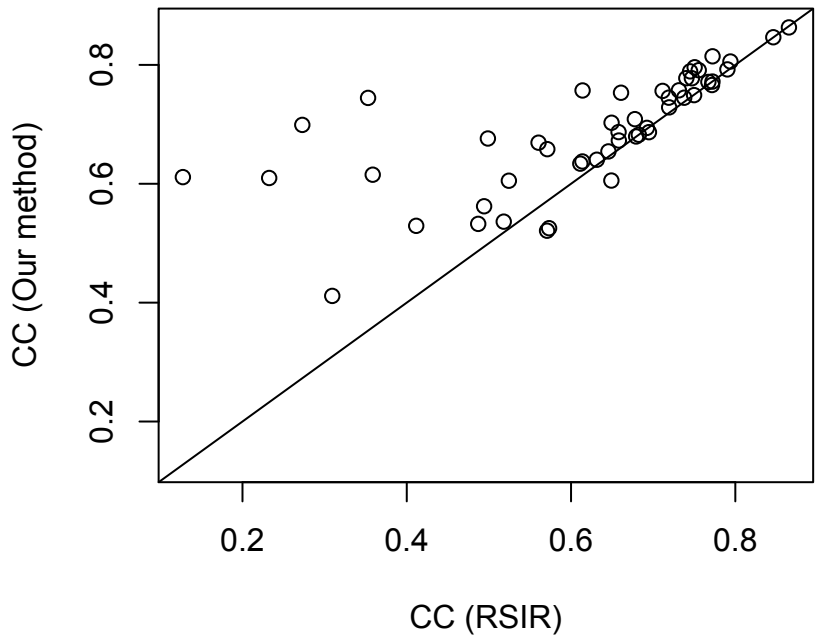

Figure 2.16. Scatterplot comparing canonical correlations between the estimated and true projected directions for Example 5. The vertical axis represents the canonical correlations from our alternating least squares method, and the horizontal axis represents the corresponding canonical correlations from RSIR [Zhong et al., 2005].

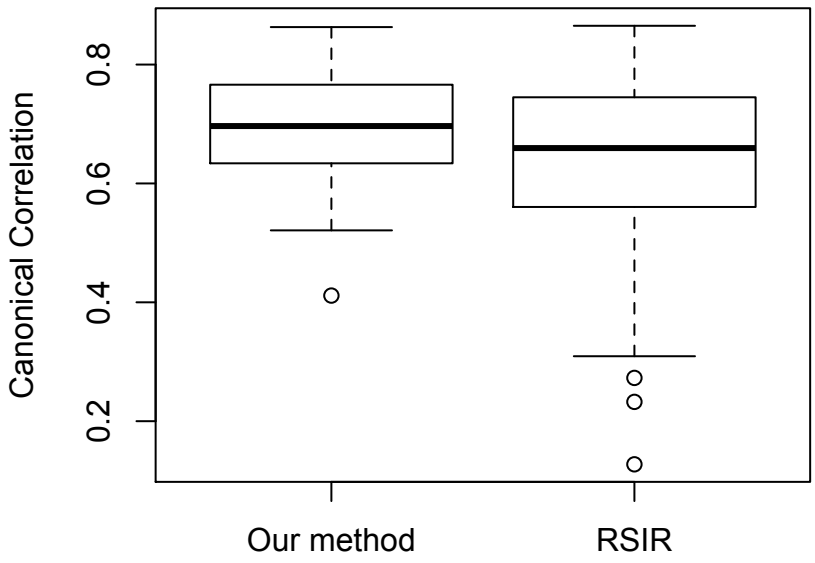

Figure 2.17. Boxplot comparing canonical correlations between the estimated and true projected directions for Example 5.

Mulligan *et al.* [2007] simplified the five-level clinical response to two levels (progressive disease (PD) and response (R), excluding no change (NC) patients) and used a linear discriminant analysis to predict PD and R. Specifically, the $44,928$ genes were preselected to retain the top 100 genes with the largest difference between PD and R. A linear combination of these 100 genes were used to build a classifier function for the simplified two-level response Y. Two of the four trials were used as a training set and one trial as a test set such that there were 91 observations in the training set and 71 in the test set.

In our procedure we treat the original five-level clinical response as the response variable Y, and use the same set of the top 100 differentially expressed genes as predictors. The training set is sliced into 5 pieces according to the ordinal values of Y, and we employ techniques in Sections 2.2.4 and 2.2.5 to determine the tuning parameter τ and the structural dimension d. Specifically, τ is chosen to be 5 and d is determined to be 1. We then conduct dimension reduction through our proposed method to estimate the dimension reduction direction β's. The linear discriminant analysis (LDA) is applied on the projected direction $\beta\mathbf{X}$ for the training set to build a classifier for Y. The classifier is then used to make predictions on the test set and the predicted result is compared to Mulligan *et al.* [2007]'s classification result as shown in Table 2.1. In summary, we identify 37 out of the 38 patients who are responders to the bortezomib treatment and only one patient, who is a responder to the treatment, is incorrectly classified as progressive disease (PD). Six out of the 15 patients who have PD to the treatment are correctly classified, but the other 9 patients are incorrectly classified as responders to the treatment. The overall accuracy of our prediction result is 81%, which is 6% higher compared to 75% of Mulligan *et al.* [2007]'s result. Thus, by applying our proposed dimension reduction method which extends SIR to high dimensional data, the one-dimensional projected direction, $\beta\mathbf{X}$, leads to better prediction than the other linear combination of the predictors.

Table 2.1
Comparison of prediction table of PD vs. R for the bortezomib data in the pharmacogenomics example. The upper panel is the result from our alternating least squares method and the lower panel is the result from Mulligan *et al.* [2007].

		Actual		
		R	PD	Total
Predicted	R	37	9	46
	PD	1	6	7
Total		38	15	53

Sensitivity: 97%

Specificity: 40%

Positive Predictive Value: 80%

Negative Predictive Value: 86%

Accuracy: 81%

		Actual		
		R	PD	Total
Predicted	R	35	10	45
	PD	3	5	8
Total		38	15	53

Sensitivity: 92%

Specificity: 33%

Positive Predictive Value: 78%

Negative Predictive Value: 63%

Accuracy: 75%

2.5 Conclusion and discussion

Sliced Inverse Regression (SIR) is an effective dimension reduction method in prediction, where a response variable is assumed to depend on a large number of predictors through an unknown function. The original SIR extracts features of high dimensional predictors from their low dimensional projections and is based on eigenvalue decomposition of the conditional sample covariance matrix. The difficulty of estimating large covariance matrices and their inverses limits the application of SIR.

In our work we develop a new alternating least squares method based on the least squares formulation of SIR. We borrow the idea of alternating least squares from Li and Yin [2008], but solve a different constrained optimization problem with a modified L_2 type penalty term suggested by Bernard-Michel et al. [2008]. Our proposed method is an iterative method, and it is shown that the solutions of our algorithm converge to the local minimum. We also successfully sidestep the difficult problem of estimating large covariance matrices and their inverses. Both simulation examples and the application in a pharmacogenomics study of bortezomib in multiple myeloma demonstrate the effectiveness of our method. By overcoming the limitation of SIR for high dimensional data, our method brings the conventional dimension reduction technique back to date for the challenges of high dimensional data analysis.

LIST OF REFERENCES

LIST OF REFERENCES

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory*, 267-281.

Bellman, R.E. (1961) Adaptive control processes: a guided tour. *Princeton University Press*.

Bernard-Michel C., Gardes L. and Girard S. (2008) A note on sliced inverse regression with regularizations, *Biometrics*, 64, 982-986.

Bickel, P. and Levina, E. (2008) Covariance regularization by thresholding, *Annals of Statistics*, 36, 2577-2604.

Bura, E. and Cook, R.D. (2001) Estimating the structural dimension of regressions via parametric inverse regression, *Journal of the Royal Statistical Society Series B*, 63-2, 393-410.

Cleveland, W.S. (1979) Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association*, 74, 829-836.

Cleveland, W.S., Devlin, S.J. (1988) Locally weighted regression: an approach to regression analysis by local fitting, *Journal of the American Statistical Association*, 83, 596-610.

Cook, R.D. and Weisberg, S. (1991) Discussion of "Sliced inverse regression for dimension reduction," by Li [1991], *Journal of the American Statistical Association*, 86, 328-332.

Cook, R.D. (1994) On the interpretation of regression plots, *Journal of the American Statistical Association*, 89, 177-189.

Cook, R.D. (1996) Graphics for regressions with a binary response, *Journal of the American Statistical Association*, 91, 983-992.

Cook, R.D. (1998) Regression graphics: ideas for studying regression through graphics, *New York: Wiley*.

Cook, R.D. (2004) Testing predictor contributions in sufficient dimension reduction, *Annals of Statistics*, 32, 1061-1092.

Cook, R.D. and Ni, L. (2005) Sufficient dimension reduction via inverse regression: a minimum discrepancy approach, *Journal of the American Statistical Association*, 100, 410-428.

Crump, R.K., Hotz, V.J., Imbens, G.W., Mitnik, O.A. (2006) Nonparametric tests for treatment effect heterogeneity, *Technical Working Paper*, 324, National Bureau of Economic Research.

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

Fan, J., Gijbels, I. (1996) *Local polynomial modelling and its applications*. 1st Edition. Chapman & Hall.

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

Fan, J., Liu, H. (2013) Statistical analysis of big data on pharmacogenomics, *Advanced Drug Delivery Reviews*, 65, 987-1000.

Foster, J.C., Taylor, J.M.G., Ruberg, S.J. (2011) Subgroup identification from randomized clinical trial data, *Statistics in Medicine*, 30, 2867-2880.

Golub, G.H., Heath, M., and Wahba, G. (1979) Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics*, 21, 215-223.

Hall, P. and Li, K.C. (1993) On almost linearity of low dimensional projections from high dimensional data, *Annals of Statistics*, 21, 867-889.

Hastie, T., Loader, C. (1993) Local regression: automatic kernel carpentry, *Statistical Science*, 8(2), 120-129.

Hurvich, C.M., Tsai, C. (1989) Regression and time series model selection in small samples, *Biometrika*, 76, 297-307.

Hurvich, C.M., Simonoff, J.S., Tsai, C. (1998) Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, *Journal of the Royal Statistical Society Series B*, 60, 271-293.

Johnstone, I.M. (2001) On the distribution of the largest eigenvalue in principal component analysis, *Annals of Statistics*, 29(2), 295-327.

Johnstone, I.M. and Lu, A. Y. (2004) Sparse principal components analysis, Unpublished Manuscript.

Johnstone, I.M. (2006) High dimensional statistical inference and random matrices, Proceedings of the International Congress of Mathematics, Madrid, Spain.

Ledoit, O. and Wolf, M.(2004) A well-conditioned estimator for large-dimensional covariance matrics, *Journal of Multivariate Analysis*, 88(2), 365-411.

Li, K.C. (1991) Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association*, 86, 316-327.

Li, K.C. (1992) On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma, *Journal of the American Statistical Association*, 87, 1025-1034.

Li, L. and Yin, X. (2008) Sliced inverse regression with regularizations, *Biometrics*, 64, 124-131.

Lipkovich, I., Dmitrienko, A., Denne, J., Enas, G. (2011) Subgroup identification based on differential effect search - A recursive partitioning method for establishing response to treatment in patient subpopulations, *Statistics in Medicine*, 30, 2601-2621.

Ma, Y. and Zhu, L. (2013) Efficiency loss and the linearity condition in dimension reduction, *Biometrika*, 100(2), 371-383.

Mallows, C.L. (1973) Some comments on C_p, *Technometrics*, 15, 661-676.

Marĉenko, V.A., and Pastur, L.A. (1967) Distributions of eigenvalues of some sets of random matrices, *Math. USSR-Sb*, 1:507-536.

Moineddin, R. et al. (2008) Identifying subpopulations for subgroup analysis in a longitudinal clinical trial, *Contemporary Clinical Trials*, 29, 817-822.

Moore, J.H. et al. (2010) Bioinformatics challenges for genome-wide association studies, *Bioinformatics*, 26, 445-455.

Mulligan, G. et al. (2007) Gene expression profiling and correlation with outcome in clinical trials of the proteasome inhibitor bortezomib, *Blood*, 109, 3177-3188.

Rothman, A.J., Levina, E., and Zhu, J. (2010) Generalized thresholding of large covariance matrices, *Journal of Americal Statistical Association*, 104(485), 177-186.

Rothman, A.J. (2012) Positive definite estimators of large covariance matrices, *Biometrika*, 99(3), 733-740.

Ruberg, S.J., Chen, L., Wang, Y. (2010) The mean does not mean as much anymore: finding sub-groups for tailored therapeutics, *Clinical Trials*, 7, 574-583.

Ruppert, D., Wand, M.P. (1994) Multivariate locally weighted least squares regression, *The Annals of Statistics*, 22, 1346-1370.

Seber, G.A.F and Lee, A.J. (2003) Linear regression analysis, *Wiley*, 2nd edition.

Schwarz, G. (1978) Estimating the dimension of a model, *The Annals of Statistics*, 6, 461-464.

Storlie, C.B., Helton, J.C. (2007) Multiple predictor smoothing methods for sensitivity analysis: desciption of techniques, *Reliability Engineering and System Safety*, 93, 28-54.

Tibshirani, R. (1996) Regression shrinkage and selection via the LASSO, *Journal of the Royal Statistical Society Series B*, 58, 267-288.

Wright, G., Tan, B., Rosenwald, A., Hurt, E.H., Wiestner, A., Staudt, L.M.(2003) A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma, *Proceedings of the National Academy of Sciences*, 100, 9991-9996.

Wu, Q., Liang, F., and Mukherjee, S. (2008) Regularized sliced inverse regression for kernel models, *Technical report, Duke University, Durham, NC.*

Zhang, B., Tsiatis, A.A., Laber, E.B., Davidian, M. (2012) A robust method for estimating optimal treatment regimes, *Biometrics*, 68, 1010-1018.

Zhong, W., Zeng, P., Ma, P., Liu, J.S. and Zhu, Y. (2005) RSIR: Regularized sliced inverse regression for motif discovery, *Bioinformatics*, 21, 4169-4175.

Zhu, L., Miao, B., and Peng, H. (2006) On sliced inverse regression with large dimensional covariates, *Journal of Americal Statistical Association*, 101, 630-643.

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

Zou, H. (2006) The adaptive lasso and its oracle properties, *Journal of Americal Statistical Association*, 101, 1418-1429.

VITA

VITA

Jingyi Zhu received her MS degree in statistics from University of Illinois at Urbana-Champaign in 2009. Since Fall 2009, she has been pursuing a Ph. D. degree in the Department of Statistics at Purdue University, West Lafayette, Indiana. She has received the Bilsland Fellowship from May 2013 to May 2014. Her research interests include statistical modeling, variable selection, dimension reduction, and machine learning.

Jingyi Zhu's publications for her research work at Purdue include:

Journal Papers

- J. Zhu and J. Xie. Nonparametric variable selection for predictive models and subpopulations in clinical trials. Journal of Biopharmaceutical Statistics, DOI: 10.1080/10543406.2014.920861.
- J. Zhu and J. Xie. Extension of sliced inverse regression for high dimensional but low sample size data. In preparation.

Conference Papers

- J. Zhu and J.Xie. Extension of sliced inverse regression for high dimensional but low sample size data. Presented in SRC2013 Conference.