SUFFICIENT DIMENSION REDUCTION IN COMPLEX DATASETS

A Dissertation
Submitted to
the Temple University Graduate Board

in Partial Fulfillment
of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

by
Chaozheng Yang
July, 2016

Examining Committee Members:

Dr. Yuexiao Dong, Advisory Chair, Statistics
Dr. William Wei, Statistics
Dr. Cheng Yong Tang, Statistics
Dr. Shanshan Ding, External Member, University of Delaware

ABSTRACT

SUFFICIENT DIMENSION REDUCTION IN COMPLEX DATASETS

Chaozheng Yang

DOCTOR OF PHILOSOPHY

Temple University, July, 2016

Dr. Yuexiao Dong, Chair

This dissertation focus on two problems in dimension reduction: one is using permutation approach to test predictor contribution; the other one is through combining clustering method with robust regression to estimate dimension reduction subspace.

Aiming to test predictor contribution in a model-free fashion, marginal coordinate tests based on sliced inverse regression (SIR) and sliced average variance estimation (SAVE) have been studied. Estimating the null distributions of the test statistics is a critical step for such tests. We propose a novel permutation test approach to facilitate the marginal coordinate tests, which applies to existing tests such as SIR and SAVE, and can be readily extended to a new marginal coordinate test based on directional regression.

Least absolute deviation (LAD) regression is an important alternative to ordinary least squares (OLS) regression in linear models. A surprising result

in Li and Duan (1989) showed that OLS can be used for dimension reduction in single-index models as long as the predictor distribution satisfies a global linear conditional mean assumption. The proposal in Li and Duan (1989) has two limitations. First, it is well-known that OLS is sensitive to outliers and fails in the case of heavy-tailed error distribution. Second, the global linearity assumption for the predictor distribution can be violated when there is nonlinear relationship among the predictors. To address these limitations, cluster-based LAD for dimension reduction is proposed. By inheriting the benefit of LAD over OLS in linear models, our proposal becomes more robust to outliers or heavy-tailed error distribution. We also replace the global linearity assumption with the more flexible local linearity assumption through k-means clustering.

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Dr. Yuexiao Dong, who has been a tremendous mentor for me. His guidance and inspiration have benefited me enormously in all the time of research and writing of this thesis.

I would also like to thank the rest of my thesis committee memebers: Dr. William Wei, Dr. Cheng Yong Tang and Dr. Shanshan Ding for their valuable comments and suggestions.

Last but not least, a special thanks to my parents, my wife and my son for their unconditional support, love, and encouragement during this memorable journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Motivation

Regression analysis is widely used for examining relationship among variables. Given a univariate response and its corresponding explanatory predictors, when the link function is known, traditional parametric approaches (e.g. the least squares method or maximum likelihood method) can be used to estimate the regression model.

However, it is also not uncommon to have unknown parametric model. To find relationship among variables for this situation, nonparametric regression techniques such as local smoothing may be applied. However, smoothing techniques may not be sufficient as many of them can be used only to one-dimensional problem and methods can be failed along with increasing of predictors dimensions. Actually, along with the dimension of predictors increasing, the total number of observations needed for local smoothing escalates exponentially and the sparseness of data points make it harder to continue to meet $\sqrt{n}$ convergence rate. Therefore, simply smoothing over high-dimensional space can restrain the accuracy of statistical inference and this phenomenon is called the curse of dimensionality (Bellman, 1961).

To avoid the curse of dimensionality, one remedy is to project high-dimensional data onto a low-dimensional space. In another word, it is aimed at reducing the number of variables while still maintaining the information for the regression. Methods in this area is called *dimension reduction*.

In this proposal, we will study dimension reduction approaches. Following this introduction, we will show basic dimension reduction model along with several important concepts. After that, there will be a review of classic approaches in dimension reduction.

1.2 Central Space

Dimension reduction is efficient to address curse of dimensionality and is an effective tool in high-dimensional data analysis.

Consider a univariate response Y and a p-dimensional predictor $\boldsymbol{X} = (X_1, \ldots, X_p)^T$. The goal for dimension reduction is to find linear combinations of $\boldsymbol{X}$, such that Y is independent of $\boldsymbol{X}$ given these linear combinations.

Mathematically, it seeks a matrix $\boldsymbol{\beta} \in \mathbb{R}^{p \times d}$ with $d<p$ that can meet the condition

$$Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{\beta}^T \boldsymbol{X}$$

, where $\perp\!\!\!\perp$ denotes conditional independence. Note that for any $d \times d$ nonsingular matrix $\boldsymbol{A}$, its multiplication to $\boldsymbol{\beta}$ from right will not change the conditional independence, which means $Y \perp\!\!\!\perp \boldsymbol{X} | (\boldsymbol{\beta}\boldsymbol{A})^T \boldsymbol{X}$ can be met if and only if $Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{\beta}^T \boldsymbol{X}$. Therefore, instead of matrix $\boldsymbol{\beta}$ itself, the column space of $\boldsymbol{\beta}$ is the one that really drives this conditional independence relationship. We call the column space of $\boldsymbol{\beta}$ as a dimension reduction subspace (DRS).

Also note that the column space of $\boldsymbol{\beta}$ can be contained by another matrix's column space. For example, assuming $\boldsymbol{\gamma} \in \mathbb{R}^{p \times q}$ also satisfying $Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{\gamma}^T \boldsymbol{X}$, then it is possible that $\mathrm{Span}(\boldsymbol{\beta}) \subseteq \mathrm{Span}(\boldsymbol{\gamma})$ with $d \leq q$. Here, $\mathrm{Span}(\boldsymbol{A})$ means the space spanned by the columns of any matrix $\boldsymbol{A}$. Because of this, conditional independence relation and its relative dimension reduction subspace are not necessary to be unique. Aiming at finding a minimum dimension reduction subspace, Cook (1998) introduced the idea of central space and recovering such space is one of major goals of dimension reduction. And the central space definition is illustrated as below.

Definition 1.2.1. *The central space (CS) for* $(\boldsymbol{X}, Y)$ *is defined as the intersection of all dimension reduction subspaces for* $(\boldsymbol{X}, Y)$*. This space is denoted as* $\mathcal{S}_{Y|\boldsymbol{X}}$*.*

The central space dimension $d = \dim(\mathcal{S}_{Y|\boldsymbol{X}})$ is called the *structural dimension* of the regression and we assume the structure dimension d is known in our work.

Let $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X})$ and $\boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{X})$, then $\boldsymbol{Z} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$ is the standardized predictor. The relationship between the $\boldsymbol{Z}$-scale central space $\mathcal{S}_{Y|\boldsymbol{Z}}$ and the $\boldsymbol{X}$-scale central space $\mathcal{S}_{Y|\boldsymbol{X}}$ is stated next.

Proposition 1.2.1. *Suppose $\mathcal{S}_{Y|\boldsymbol{Z}}$ has basis $\boldsymbol{\eta}$ such that $\mathcal{S}_{Y|\boldsymbol{Z}} = \mathrm{Span}(\boldsymbol{\eta})$. Then $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta}$ is a basis for $\mathcal{S}_{Y|\boldsymbol{X}}$ and satisfies $\mathcal{S}_{Y|\boldsymbol{X}} = \mathrm{Span}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta})$.*

The relationship in proposition 1.2.1 is known as the invariance property of the central space. Due to this invariance property, we can first estimate of the $\boldsymbol{Z}$-scale central space, and then transform it back to the $\boldsymbol{X}$-scale.

1.3 Some Popular Estimators of the Central Space

In this section, we are going to review several classic dimension reduction methods. For each method, prerequisites and estimating algorithm will be stated. Besides, we will discuss the advantages and disadvantages of these methods.

1.3.1 Sliced Inverse Regression

Sliced Inverse Regression (SIR) was proposed by Li (1991) and this is one of the most common dimension reduction techniques. Stimulated by SIR, various Inverse regression based methods have been developed.

Many inverse regression methods require a key assumption which is called as the linear conditional mean assumption (LCM). This assumption is illustrated as below.

Assumption 1.3.1. *Suppose* $\boldsymbol{\beta} \in \mathbb{R}^{p \times d}$ *is a basis for the central space* $\mathcal{S}_{Y|\boldsymbol{X}}$*, then the linear conditional mean assumption assumes* $\mathrm{E}(X|\boldsymbol{\beta}^T\boldsymbol{X})$ *is a linear function of* $\boldsymbol{\beta}^T\boldsymbol{X}$*.*

The main idea for inverse regression based method is fulfilled through reversing standardized predictors $\boldsymbol{Z}$ and response Y in regression. Since response Y is univariate, the high-dimensional problem is simplified to a low dimensional problem and the issue of curse of dimensionality for high-dimensional data can be avoided.

Motivated by this idea, Li (1991) proved the following theorem which is the key result for sliced inverse regression.

Theorem 1.3.1. *Assume assumption 1.3.1 hold, then the inverse regression curve* $\mathrm{E}(\boldsymbol{Z}|Y)$ *belongs to the central space* $\mathcal{S}_{Y|\boldsymbol{Z}}$*. Also, the column space of the matrix* $\boldsymbol{M}_{SIR} = \mathrm{Var}(\mathrm{E}(\boldsymbol{Z}|Y))$ *is a subspace of* $\mathcal{S}_{Y|\boldsymbol{Z}}$*.*

We use $\boldsymbol{M}$ to represent kernel matrix for each method; and in above theorem, $\boldsymbol{M}_{\text{SIR}}$ represents for SIR's kernel matrix.

For discrete response Y, assume without loss of generality that Y has support $\Pi = \{1, 2, \ldots, H\}$. Then for $h \in \Pi$, $\mathrm{E}(\boldsymbol{Z}|Y = h)$ denotes the within group mean for the hth category of Y. The meaning of $\mathrm{E}(\boldsymbol{Z}|Y)$ for continuous response Y is explained next. Let $\{J_1, \ldots, J_H\}$ be a partition of the support of Y. Then $\mathrm{E}(\boldsymbol{Z}|Y)$ is discretized as $\mathrm{E}(\boldsymbol{Z}|Y \in J_h)$, $h = 1, \ldots, H$. Therefore, regardless response Y's variable type, the discretized version of the theorem 1.3.1 is applied in practice.

Let f_h be the probability of $Y \in J_h$ and denote $\boldsymbol{\xi}_h = \mathrm{E}(\boldsymbol{Z}|Y \in J_h)$. A discretized version of $\boldsymbol{M}_{\text{SIR}}$ thus becomes $\boldsymbol{M}_{\text{SIR}} = \sum_{h=1}^{H} f_h \boldsymbol{\xi}_h \boldsymbol{\xi}_h^T$.

One of nice properties for SIR is it can achieve $\sqrt{n}$ convergence rate. Besides, SIR can be used to determine structural dimension d of the central space $\mathcal{S}_{Y|\boldsymbol{Z}}$ through a sequential hypothesis test (Li, 1991).

On the other side, though SIR has many advantages and can be implemented to solve curse of dimensionality for a wide scope of high-dimensional data, this method has its restrictions: symmetric on $\boldsymbol{Z}$ may cause the conditional expectation to be zero and this can fail SIR (e.g. $\boldsymbol{Z} = (Z_1, \ldots, Z_p) \sim N(0, \boldsymbol{I}_p)$ and $Y = Z_1^2$). Another limitation for SIR is assuming there are m distinct values for predictor Y, then this method can estimate at most $m - 1$ directions in the central space. For example, if Y is a binary variable, then

SIR can estimate at most one direction.

1.3.2 Sliced Average Variance Estimator

As reviewed, SIR will be failed if there is a U-shaped curve, or in another way of saying, this method doesn't work well for the case of $E(\boldsymbol{Z} \mid Y) = 0$. To overcome this difficulty, instead of calculating first-moment, Cook and Weisberg (1991) proposed a method to calculate the within slice variance to estimate central space and this approach is well known as Sliced Average Variance Estimates (SAVE).

Compared with first-moment based method, second-moment based method requires more assumptions. In addition to the LCM assumption 1.3.1, it needs to meet the following constant conditional variance (CCV) assumption.

Assumption 1.3.2. *Suppose $\boldsymbol{\beta} \in \mathbb{R}^{p \times d}$ is a basis for the $\mathcal{S}_{Y|Z}$. The constant conditional variance assumption assumes that*

$$\mathrm{Var}(\boldsymbol{Z}|\boldsymbol{\beta}^T\boldsymbol{Z})$$

is a non-random matrix.

The basis to support this method is illustrated as the follows and $\boldsymbol{I}_p$ indicates the identity matrix of dimension p.

Theorem 1.3.2. *Under Assumptions 1.3.1 and 1.3.2, the column of the random* *matrix*

$$\boldsymbol{I}_p - \mathrm{Var}(\boldsymbol{Z} \mid Y)$$

belongs to the central space $\mathcal{S}_{Y|\boldsymbol{Z}}$. *Consequently, the column of the random matrix*

$$\boldsymbol{M}_{SAVE} = \mathrm{E}(\boldsymbol{I}_p - \mathrm{Var}(\boldsymbol{Z} \mid Y))^2$$

belongs to the central space $\mathcal{S}_{Y|\boldsymbol{Z}}$.

SAVE's has $\sqrt{n}$ convergence rate. Also, this method can accurately estimate U-shape curve and provide exhaustively estimate for that case. However, since SAVE is a second-order method, as a trade off, it requires both LCM and CCV assumptions. Compared with SIR, which only requires LCM condition, the extra condition from SAVE increases complexity of computation and actually, it is not very efficient for estimating some simple cases such as monotone trend for small to moderate sample sizes.

1.3.3 Directional Regression

As reviewed, SIR and SAVE are constructed through the first two conditional moments $\mathrm{E}(\boldsymbol{Z} \mid Y)$ and $\mathrm{E}(\boldsymbol{Z}\boldsymbol{Z}^T \mid Y)$, respectively. SIR fails when the response surface is symmetric about the origin. Unlike SIR, SAVE can successfully identify U-shape for the response, however, it is not very efficient in estimating monotone trend for small to moderate sample sizes.

To overcome these shortcomings, Li and Wang proposed Directional Regression (DR; 2007). This approach synthesize the dimension reduction methods based upon first two conditional moments, such as SIR and SAVE, so it combines the advantages of both of these methods. DR has $\sqrt{n}$ convergence rate and under mild condition, it provides exhaustive estimator of the central space.

Let $(\widetilde{\boldsymbol{Z}}, \widetilde{Y})$ be an independent copy of $(\boldsymbol{Z}, Y)$, and define

$$A(Y, \widetilde{Y}) = \mathrm{E}((\boldsymbol{Z} - \widetilde{\boldsymbol{Z}})(\boldsymbol{Z} - \widetilde{\boldsymbol{Z}})^T \mid Y, \widetilde{Y}).$$

Then the following theorem provides the theoretical background for DR.

Theorem 1.3.3. *Under Assumptions 1.3.1 and 1.3.2, the column space of the random matrix*

$$2\boldsymbol{I}_p - A(Y, \widetilde{Y})$$

is contained in $\mathcal{S}_{Y|\boldsymbol{Z}}$.

Theorem 1.3.3 suggests the DR's kernel matrix

$$\boldsymbol{M}_{\mathrm{DR}} = \mathrm{E}(2\boldsymbol{I}_p - A(Y, \widetilde{Y}))^2$$

as the population version of the estimate for $\mathcal{S}_{Y|\boldsymbol{Z}}$.

Also, the estimator $\boldsymbol{M}_{\mathrm{DR}}$ above can be re-expressed as a nonlinear functional of conditional moments of $\boldsymbol{Z}$ given Y, whose estimation requires only $O(n)$ operations. The updated kernel matrix is provided in the following theorem.

Theorem 1.3.4. *DR's kernel matrix can be re-expressed as*

$$\boldsymbol{M}_{DR} = 2\mathrm{E}[\mathrm{E}^2(\boldsymbol{Z}\boldsymbol{Z}^T \mid Y)] + 2\mathrm{E}^2[\mathrm{E}(\boldsymbol{Z} \mid Y)\mathrm{E}(\boldsymbol{Z}^T \mid Y)]$$

$$+2\mathrm{E}[\mathrm{E}(\boldsymbol{Z}^T \mid Y)\mathrm{E}(\boldsymbol{Z} \mid Y)]\mathrm{E}[\mathrm{E}(\boldsymbol{Z} \mid Y)\mathrm{E}(\boldsymbol{Z}^T \mid Y)] - 2\boldsymbol{I}_p.$$

1.3.4 Sample Estimators

As reviewed, SIR, SAVE and DR are efficient dimension reduction methods and we have discussed how these methods can effectively find central space through theorem and next, we are going to take a look of their relative sample algorithms.

According to proposition 1.2.1, the central space has an important invariance property and because of that, we can always estimate of the $\boldsymbol{Z}$-scale central space, and then transform it back to the $\boldsymbol{X}$-scale. To facilitate our discussions, same as introduction for theorems, the sample level algorithms are also illustrated based on $\boldsymbol{Z}$-scale.

First, let's take a look of SIR. The sample version of its kernel matrix $\mathrm{Var}[\mathrm{E}(\boldsymbol{Z}|Y \in J_h)]$ will be used to estimate the central space for sliced inverse regression. And its sample level algorithm is described through the following steps.

1. Divide range of Y into H slices, $J_1, \ldots, J_H$, and compute the average of $\boldsymbol{Z}$ within each slice; that is

$$\hat{\boldsymbol{\xi}}_h = \frac{1}{n_h} \sum_{j \in J_h} \boldsymbol{Z}_j,$$

where n_h is the number of Y that are fall into the hth slice J_h.

2. Construct kernel matrix for SIR:

$$\widehat{\boldsymbol{M}}_{\text{SIR}} = \sum_{h=1}^{H} \frac{n_h}{n} \hat{\boldsymbol{\xi}}_h \hat{\boldsymbol{\xi}}_h^T.$$

3. Assume structural dimension d is known. Conducting a principal component analysis for the sample version of SIR's kernel matrix $\widehat{\boldsymbol{M}}_{\text{SIR}}$, then the first d eigenvectors $\hat{\gamma}_1, \ldots, \hat{\gamma}_d$ corresponding to the largest d eigenvalues are used to estimate $\mathcal{S}_{Y|\boldsymbol{Z}}$.

Similar as SIR's sample algorithm, the last step for SAVE and DR also needs to apply eigenvalue decomposition on kernel matrix to estimate central space. The major difference is each method has its unique kernel matrix and next, we are going to show how to estimate these kernel matrices at sample level.

To estimate central space through SAVE, SIR's mean calculation $\hat{\xi}_h$ needs to be updated as

$$\hat{\boldsymbol{\xi}}_h = \boldsymbol{I}_p - \widehat{\text{Var}}(\boldsymbol{Z} \mid Y \in J_h).$$

The above adjustment leads the sample version of SIR's kernel matrix is replaced by sample version of SAVE's kernel matrix and this estimate is provided as below:

$$\widehat{\boldsymbol{M}}_{\text{SAVE}} = \sum_{h=1}^{H} \frac{n_h}{n} (\boldsymbol{I}_p - \widehat{\text{Var}}(\boldsymbol{Z} \mid Y \in J_h))^2.$$

Following SIR and SAVE, let's take a look of DR. Its sample level algorithm is similar as other inverse regression methods, such as SIR or SAVE, except the sample version of DR's kernel matrix is deployed to estimate $S_{Y|\boldsymbol{Z}}$. Still, let use $J_1, \ldots, J_H$ to partition in support of Y and assume $f_h = P(Y \in J_h)$, then DR's kernel matrix is discretized as

$$\widehat{\boldsymbol{M}}_{\mathrm{DR}} = 2\sum \mathrm{E}_n^2(\boldsymbol{Z}\boldsymbol{Z}^T - \boldsymbol{I}_p \mid Y \in J_h) f_h + 2(\sum \mathrm{E}_n(\boldsymbol{Z} \mid Y \in J_h)\mathrm{E}_n(\boldsymbol{Z}^T \mid Y \in J_h) f_h)^2$$
$$+2\sum \mathrm{E}_n(\boldsymbol{Z}^T \mid Y \in J_h)\mathrm{E}_n(\boldsymbol{Z} \mid Y \in J_h) f_h \sum \mathrm{E}_n(\boldsymbol{Z} \mid Y \in J_h)\mathrm{E}_n(\boldsymbol{Z}^T \mid Y \in J_h) f_h,$$

where the summation is over $h = 1, \ldots, H$ and the notation such as $\mathrm{E}_n(\boldsymbol{Z} \mid Y \in J_h)$ stands for sample conditional moments, defined by

$$\mathrm{E}_n(\boldsymbol{Z} \mid Y \in J_h) = \frac{\mathrm{E}_n[\boldsymbol{Z} I(Y \in J_h)]}{\mathrm{E}_n I(Y \in J_h)} = \frac{\sum_{i=1}^n \boldsymbol{Z}_i I(Y_i \in J_h)}{\sum_{i=1}^n I(Y_i \in J_h)} = \frac{1}{n_h} \sum_{j \in J_h} \boldsymbol{Z}_j.$$

CHAPTER 2

ON PERMUTATION TESTS FOR PREDICTOR CONTRIBUTION IN SUFFICIENT DIMENSION REDUCTION

2.1 Introduction

Besides estimating relationship among predictors, another important topic in dimension reduction is to determine predictor's contribution via hypothesis test.

For $\boldsymbol{X} \in \mathbb{R}^p$ and subscript $i \in \{1, 2, \ldots, p\}$, denote $\boldsymbol{X}_{-i} \in \mathbb{R}^{p-1}$ as

$$(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p)^T.$$

One can test the contribution of X_i through the following hypotheses,

$$H_0 : Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{X}_{-i} \text{ versus } H_a : Y \text{ is not independent of } \boldsymbol{X} \text{ given } \boldsymbol{X}_{-i}, \quad (2.1)$$

where $\perp\!\!\!\perp$ denotes statistical independence. In (2.1), the null hypothesis implies that Y is independent of $\boldsymbol{X}$ given $\boldsymbol{X}_{-i}$. If we fail to reject H_0, we conclude that predictor X_i does not have significant contribution to the regression between Y and $\boldsymbol{X}$. Hypotheses (2.1) are first proposed as the marginal coordinate hypotheses in the seminal work of Cook (2004).

Cook (2004) also proposed the corresponding marginal coordinate test based on SIR. A similar test based on SAVE has been discussed in Shao, Cook and Weisberg (2007). More recently, the marginal coordinate test based on DR is developed in Yu and Dong (2015). For marginal coordinate tests based on SIR, SAVE and DR, a technical difficulty is to develop the distributions of their corresponding test statistics under H_0. The asymptotic distribution for each method-specific test statistic turns out to be a sum of weighted $\chi^2(1)$ distributions, where the exact weights for each method can be found in Cook (2004), Shao, Cook and Weisberg (2007) and Yu and Dong (2015) respectively.

As an alternative to derive the null distribution of the aforementioned marginal coordinate tests, we propose a unified permutation test approach. Our proposal is easy to implement, as it only involves random permutations of the observed predictors while fixing the responses before recalculating the sample test statistics with the permuted samples. It applies to marginal coordinate tests based on SIR, SAVE and DR, and no longer requires calculation of the method-specific weights to determine the asymptotic null distribution.

The rest of the chapter is organized as follows. The permutation approach with SIR is developed in Section 2.2, and the analogous approaches for SAVE and DR are proposed in Section 2.3. Extensive numerical studies are carried out in Section 2.4. For ease of presentation, all the proofs are delegated to the last section 2.5. We assume normality for the predictor $\boldsymbol{X}$ throughout this chapter.

2.2 Permutation test for predictor contribution with SIR

2.2.1 Test statistic construction

Still, we use $\boldsymbol{M}_{\mathrm{SIR}}$, $\boldsymbol{M}_{\mathrm{SAVE}}$ and $\boldsymbol{M}_{\mathrm{DR}}$ to represent kernel matrix for each method. The we have

Proposition 2.2.1. *Suppose $\boldsymbol{Z}$ is normal. Then* $\mathrm{Span}(\boldsymbol{M}) \subseteq \mathcal{S}_{Y|\boldsymbol{Z}}$*, where* $\boldsymbol{M}$

can be $\boldsymbol{M}_{SIR}$, $\boldsymbol{M}_{SAVE}$ *or* $\boldsymbol{M}_{DR}$.

We remark that the normality assumption can be relaxed to weaker assumptions. See, for example, the discussions in Li and Wang (2007). Proposition 2.2.1 suggests that the eigenvectors corresponding to the nonzero eigenvalues of kernel matrices $\boldsymbol{M}_{\text{SIR}}$, $\boldsymbol{M}_{\text{SAVE}}$ and $\boldsymbol{M}_{\text{DR}}$ can be used to recover the $\boldsymbol{Z}$-scale central space.

Also, recall that the null hypothesis $H_0 : Y \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{X}_{-i}$ in (2.1) implies that X_i has no additional contribution to Y given the other predictors. For $i = 1, \ldots, p$, define $\boldsymbol{e_i} \in \mathbb{R}^p$, where the ith element of $\boldsymbol{e_i}$ is 1 and all the other elements are zero. Alternatively, we can define $\boldsymbol{e_i}$ as the ith column of the identity matrix $\boldsymbol{I}_p$.

The next observation is key to develop the test statistics for (2.1).

Proposition 2.2.2. *Suppose* $\mathcal{S}_{Y|\boldsymbol{Z}} = \text{Span}(\boldsymbol{\eta})$ *for* $\boldsymbol{\eta} \in \mathbb{R}^{p\times d}$. *Then* $\boldsymbol{e_i}^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta} = 0$ *if and only if* $Y \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{X}_{-i}$.

Note that due to Proposition 1.2.1, $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta}$ is the basis of the $\boldsymbol{X}$-scale central space. Proposition 2.2.2 thus implies $\boldsymbol{\eta}$, or the basis of the $\boldsymbol{Z}$-scale central space, can be used to test the conditional independence $Y \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{X}_{-i}$ at the $\boldsymbol{X}$-scale.

In the case of SIR, we have $\boldsymbol{M}_{\text{SIR}} = \sum_{h=1}^{H} f_h \boldsymbol{\xi}_h \boldsymbol{\xi}_h^T$ with $\boldsymbol{\xi}_h = \text{E}(\boldsymbol{Z}|Y \in J_h)$.

To test $H_0 : Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{X}_{-i}$, Proposition 2.2.2 suggests us to consider

$$T_{\mathrm{SIR}}(\boldsymbol{e_i}) = \boldsymbol{e_i}^T \boldsymbol{\Sigma}^{-1/2} \boldsymbol{M}_{\mathrm{SIR}} \boldsymbol{\Sigma}^{-1/2} \boldsymbol{e_i}. \tag{2.2}$$

The next result is due to Cook (2004).

Proposition 2.2.3. *Suppose* $\mathrm{Span}(\boldsymbol{M}_{SIR}) = \mathcal{S}_{Y|\boldsymbol{Z}}$. *Then* $T_{SIR}(\boldsymbol{e_i}) = 0$ *if and only if* $Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{X}_{-i}$.

We have seen in Proposition 2.2.1 that the normality of the predictor will guarantee $\mathrm{Span}(\boldsymbol{M}_{\mathrm{SIR}}) \subseteq \mathcal{S}_{Y|\boldsymbol{Z}}$. The assumption that $\mathrm{Span}(\boldsymbol{M}_{\mathrm{SIR}}) = \mathcal{S}_{Y|\boldsymbol{Z}}$ is a stronger assumption, and is known as the exhaustive recovery assumption in the sufficient dimension reduction literature. A similar exhaustive recovery assumption has been made in Cook (2004).

As a direct result of Proposition 2.2.3, the sample estimator of $nT_{\mathrm{SIR}}(\boldsymbol{e_i})$, denoted by $\hat{T}_{n,\mathrm{SIR}}(\boldsymbol{e_i})$, can be constructed to test the marginal hypotheses (2.1). Let $I(\cdot)$ be the indicator function and define $\phi_h(\boldsymbol{e_i}) = E[\boldsymbol{e_i}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) I(Y \in J_h)]$. Then $T_{\mathrm{SIR}}(\boldsymbol{e_i})$ can be reexpressed as $\sum_{h=1}^{H} f_h^{-1} \phi_h(\boldsymbol{e_i}) \phi_h^T(\boldsymbol{e_i})$. At the sample level, let $\{(\boldsymbol{X}_j, Y_j) : j = 1, \ldots, n\}$ be a random sample of $(\boldsymbol{X}, Y)$. Set $\hat{\boldsymbol{\mu}} = n^{-1} \sum_{j=1}^{n} \boldsymbol{X}_j$ and $\hat{\boldsymbol{\Sigma}} = n^{-1} \sum_{j=1}^{n} (\boldsymbol{X}_j - \hat{\boldsymbol{\mu}})(\boldsymbol{X}_j - \hat{\boldsymbol{\mu}})^T$. The sample estimator of $nT_{\mathrm{SIR}}(\boldsymbol{e_i})$ becomes

$$\hat{T}_{n,\mathrm{SIR}}(\boldsymbol{e_i}) = n \sum_{h=1}^{H} \hat{f}_h^{-1} \hat{\phi}_h(\boldsymbol{e_i}) \hat{\phi}_h^T(\boldsymbol{e_i}), \tag{2.3}$$

where $\hat{f}_h = n^{-1} \sum_{j=1}^{n} I(Y_j \in J_h)$ and $\hat{\phi}_h(\boldsymbol{e_i}) = n^{-1} \sum_{j=1}^{n} \boldsymbol{e_i}^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{X}_j - \hat{\boldsymbol{\mu}}) I(Y_j \in J_h)$. As a special case of Theorem 1 in Cook (2004), we know that $\hat{T}_{n,\mathrm{SIR}}(\boldsymbol{e_i}) \xrightarrow{D}$

$\sum_{h=1}^{H} \omega_{ih}^{\text{SIR}} \chi_h^2(1)$ as $n \to \infty$ under $H_0 : Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{X}_{-i}$. Here "$\xrightarrow{D}$" means converge in distribution, $\chi_h^2(1)$ are independent chi-square with one degree of freedom for $h = 1, \ldots, H$, and $\omega_{i1}^{\text{SIR}} \geq \omega_{i2}^{\text{SIR}} \geq \cdots \geq \omega_{iH}^{\text{SIR}}$ are eigenvalues of some $\boldsymbol{\Omega}_{\boldsymbol{e_i}}^{\text{SIR}} \in \mathbb{R}^{H \times H}$. Please refer to equation (12) of Cook (2004) for the detailed form of $\boldsymbol{\Omega}_{\boldsymbol{e_i}}^{\text{SIR}}$. For the ease of reference, we denote $\boldsymbol{D}_{\boldsymbol{e_i}}^{\text{SIR}} \sim \sum_{h=1}^{H} \omega_{ih}^{\text{SIR}} \chi_h^2(1)$ as the asymptotic null distribution of $\hat{T}_{n,\text{SIR}}(\boldsymbol{e_i})$.

2.2.2 A permutation test algorithm based on SIR

Permutation test is a useful tool to test independence. In the sufficient dimension reduction literature, permutation test has been widely used to determine the dimensionality of the central space (Cook and Yin, 2001). To use permutation test for predictor contribution, we have to transform the conditional independence $Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{X}_{-i}$ into the independence test. From the definition of $T_{\text{SIR}}(\boldsymbol{e_i})$ in (2.2), we can rewrite $T_{\text{SIR}}(\boldsymbol{e_i})$ as

$$T_{\text{SIR}}(\boldsymbol{e_i}) = \sum_{h=1}^{H} f_h \mathrm{E}(\boldsymbol{e_i}^T \boldsymbol{\Sigma}^{-1/2} \boldsymbol{Z} | Y \in J_h) \mathrm{E}^T(\boldsymbol{e_i}^T \boldsymbol{\Sigma}^{-1/2} \boldsymbol{Z} | Y \in J_h).$$

The expression above suggests that a permutation test algorithm could be naturally designed under the independence between $\boldsymbol{e_i}^T \boldsymbol{\Sigma}^{-1/2} \boldsymbol{Z}$ and Y. The next observation is key to the permutation test algorithm.

Proposition 2.2.4. *Suppose $\boldsymbol{X}$ is normal. Then under $H_0 : Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{X}_{-i}$, we have $Y \perp\!\!\!\perp \boldsymbol{e_i}^T \boldsymbol{\Sigma}^{-1/2} \boldsymbol{Z}$.*

Given $\{(\boldsymbol{X}_j, Y_j) : j = 1, \ldots, n\}$ as a random sample of $(\boldsymbol{X}, Y)$, the SIR-based permutation test algorithm to get the null distribution of $\hat{T}_{n,\mathrm{SIR}}(\boldsymbol{e_i})$ is formally described next.

A1. Based on the original sample, calculate $\hat{T}_{n,\mathrm{SIR}}(\boldsymbol{e_i})$ as defined in (2.3).

A2. Fix $\{Y_j : j = 1, \ldots, n\}$. For $b = 1, \ldots, B$, denote $\{\boldsymbol{X}_j^{\{b\}} : j = 1, \ldots, n\}$ as the bth random permutation of $\{\boldsymbol{X}_j : j = 1, \ldots, n\}$. Then calculate $\hat{T}^{\{b\}}_{n,\mathrm{SIR}}(\boldsymbol{e_i})$ based on the permuted sample $\{(\boldsymbol{X}_j^{\{b\}}, Y_j) : j = 1, \ldots, n\}$.

A3. Calculate the p-value $p_{\mathrm{SIR}} = B^{-1} \sum_{b=1}^{B} I(\hat{T}^{\{b\}}_{n,\mathrm{SIR}}(\boldsymbol{e_i}) > \hat{T}_{n,\mathrm{SIR}}(\boldsymbol{e_i}))$. For given significance level α, reject $H_0 : Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{X}_{-i}$ if $p_{\mathrm{SIR}} < \alpha$.

The permutation test procedure with SIR is valid due to the next result, which states that the test statistic $\hat{T}^{\{b\}}_{n,\mathrm{SIR}}(\boldsymbol{e_i})$ based on the permutated sample has the same asymptotic null distribution as $\hat{T}_{n,\mathrm{SIR}}(\boldsymbol{e_i})$. Note that the exhaustive recovery condition is not required for the next Theorem.

Theorem 2.2.1. *Suppose $\boldsymbol{X}$ is normal. Then under $H_0 : Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{X}_{-i}$, we have $\hat{T}^{\{b\}}_{n,SIR}(\boldsymbol{e_i}) \xrightarrow{D} D^{SIR}_{\boldsymbol{e_i}}$.*

We remark that step A2 in the SIR-based algorithm can be replaced by an equivalent step as follows

A2*. Fix $\boldsymbol{X}_j$, $j = 1, \ldots, n$. For $b = 1, \ldots, B$, denote $\{Y_j^{\{b\}} : j = 1, \ldots, n\}$ as the bth random permutation of $\{Y_j : j = 1, \ldots, n\}$. Then calculate $\check{T}^{\{b\}}_{n,\mathrm{SIR}}(\boldsymbol{e_i})$ based on the permuted sample $\{(\boldsymbol{X}_j, Y_j^{\{b\}}) : j = 1, \ldots, n\}$.

It follows in step A3* that the p-value is calculated as the proportion of $\check{T}^{\{b\}}_{n,\mathrm{SIR}}(\boldsymbol{e_i})$ exceeding $\hat{T}_{n,\mathrm{SIR}}(\boldsymbol{e_i})$. One can prove $\check{T}^{\{b\}}_{n,\mathrm{SIR}}(\boldsymbol{e_i}) \xrightarrow{D} D^{\mathrm{SIR}}_{\boldsymbol{e_i}}$ under $H_0 : Y \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{X}_{-i}$. The proof is similar to the proof of Theorem 2.2.1, and is thus omitted.

2.3 Permutation tests with SAVE, PHD and DR

2.3.1 The permutation test with SAVE

As we have seen, the discretized kernel matrix for SAVE is $\boldsymbol{M}_{\mathrm{SAVE}} = \sum_{h=1}^{H} \boldsymbol{A}_h^2$ with $\boldsymbol{A}_h = f_h^{1/2}[\boldsymbol{I}_p - \mathrm{Var}(\boldsymbol{Z} \mid Y \in J_h)]$. Parallel to the development of $T_{\mathrm{SIR}}(\boldsymbol{e_i})$ in Section 2.2.1, we define $T_{\mathrm{SAVE}}(\boldsymbol{e_i}) = \sum_{h=1}^{H}(\boldsymbol{e_i}^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{A}_h\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e_i})^2$. Recall that $f_h = E[I(Y \in J_h)]$ and $\phi_h(\boldsymbol{e_i}) = E[\boldsymbol{e_i}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})I(Y \in J_h)]$. Denote $\psi_h(\boldsymbol{e_i}) = E\{\boldsymbol{e_i}^T[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - (\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T)\boldsymbol{\Sigma}^{-1}]\boldsymbol{e_i}I(Y \in J_h)\}$. We can rewrite $T_{\mathrm{SAVE}}(\boldsymbol{e_i})$ as

$$T_{\mathrm{SAVE}}(\boldsymbol{e_i}) = \sum_{h=1}^{H} f_h^{-1}[\psi_h(\boldsymbol{e_i}) + f_h^{-1}\phi_h(\boldsymbol{e_i})\phi_h^T(\boldsymbol{e_i})]^2.$$

The next result is parallel to Proposition 2.2.3, and its proof is omitted.

Proposition 2.3.1. *Suppose* $\mathrm{Span}(\boldsymbol{M}_{SAVE}) = \mathcal{S}_{Y|\boldsymbol{Z}}$. *Then* $T_{SAVE}(\boldsymbol{e_i}) = 0$ *if and only if* $Y \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{X}_{-i}$.

The exhaustive coverage assumption $\text{Span}(\boldsymbol{M}_{\text{SAVE}}) = \mathcal{S}_{Y|\boldsymbol{Z}}$ is commonly used in the literature. See, for example, Shao, Cook and Weisberg (2007).

Given $\{(\boldsymbol{X}_j, Y_j) : j = 1, \ldots, n\}$ as a random sample of $(\boldsymbol{X}, Y)$, the SAVE marginal coordinate test statistic, denoted by $\hat{T}_{n,\text{SAVE}}(\boldsymbol{e_i})$, is constructed as the sample estimator of $nT_{\text{SAVE}}(\boldsymbol{e_i})$. More specifically,

$$\hat{T}_{n,\text{SAVE}}(\boldsymbol{e_i}) = n \sum_{h=1}^{H} \hat{f}_h^{-1}[\hat{\psi}_h(\boldsymbol{e_i}) + \hat{f}_h^{-1}\hat{\phi}_h(\boldsymbol{e_i})\hat{\phi}_h^T(\boldsymbol{e_i})]^2, \tag{2.4}$$

where $\hat{\psi}_h(\boldsymbol{e_i}) = n^{-1}\sum_{j=1}^{n} \boldsymbol{e_i}^T[\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\Sigma}} - (\boldsymbol{X}_j - \hat{\boldsymbol{\mu}})(\boldsymbol{X}_j - \hat{\boldsymbol{\mu}})^T)\hat{\boldsymbol{\Sigma}}^{-1}]\boldsymbol{e_i} I(Y_j \in J_h)$, $\hat{f}_h$ and $\hat{\phi}_h(\boldsymbol{e_i})$ are defined in Section 2.2.1. As discussed in Shao, Cook and Weisberg (2007), we know that $\hat{T}_{n,\text{SAVE}}(\boldsymbol{e_i}) \xrightarrow{D} \sum_{h=1}^{H} \omega_{ih}^{\text{SAVE}} \chi_h^2(1)$ as $n \to \infty$ under $H_0 : Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{X}_{-i}$. Here $\omega_{i1}^{\text{SAVE}} \geq \omega_{i2}^{\text{SAVE}} \geq \cdots \geq \omega_{iH}^{\text{SAVE}}$ are eigenvalues of some $\boldsymbol{\Omega}_{\boldsymbol{e_i}}^{\text{SAVE}} \in \mathbb{R}^{H \times H}$. See Theorem 1 of Shao, Cook and Weisberg (2007) for the specific form of $\boldsymbol{\Omega}_{\boldsymbol{e_i}}^{\text{SAVE}}$. For the ease of reference, we denote the asymptotic null distribution of $\hat{T}_{n,\text{SAVE}}(\boldsymbol{e_i})$ by $D_{\boldsymbol{e_i}}^{\text{SAVE}}$.

The SAVE-based permutation test algorithm is presented as follows.

B1. Based on the original sample, calculate $\hat{T}_{n,\text{SAVE}}(\boldsymbol{e_i})$ as defined in (2.4).

B2. Fix $\{Y_j : j = 1, \ldots, n\}$. For $b = 1, \ldots, B$, denote $\{\boldsymbol{X}_j^{\{b\}} : j = 1, \ldots, n\}$ as the bth random permutation of $\{\boldsymbol{X}_j : j = 1, \ldots, n\}$. Then calculate $\hat{T}_{n,\text{SAVE}}^{\{b\}}(\boldsymbol{e_i})$ based on the permuted sample $\{(\boldsymbol{X}_j^{\{b\}}, Y_j) : j = 1, \ldots, n\}$.

B3. Calculate the p-value $p_{\text{SAVE}} = B^{-1}\sum_{b=1}^{B} I(\hat{T}_{n,\text{SAVE}}^{\{b\}}(\boldsymbol{e_i}) > \hat{T}_{n,\text{SAVE}}(\boldsymbol{e_i}))$. For given significance level α, reject $H_0 : Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{X}_{-i}$ if $p_{\text{SAVE}} < \alpha$.

The permutation test procedure with SAVE is justified by the next result.

Theorem 2.3.1. *Suppose $\boldsymbol{X}$ is normal. Then under $H_0 : Y \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{X}_{-i}$, we have $\hat{T}^{\{b\}}_{n,SAVE}(\boldsymbol{e_i}) \xrightarrow{D} D^{SAVE}_{\boldsymbol{e_i}}$.*

Similar to the SIR-based algorithm in Section 2.2.2, step B2 in the SAVE-based algorithm can be replaced by an equivalent step as follows

B2*. Fix $\boldsymbol{X}_j$, $j = 1, \ldots, n$. For $b = 1, \ldots, B$, denote $\{Y_j^{\{b\}} : j = 1, \ldots, n\}$ as the bth random permutation of $\{Y_j : j = 1, \ldots, n\}$. Then calculate $\check{T}^{\{b\}}_{n,\mathrm{SAVE}}(\boldsymbol{e_i})$ based on the permuted sample $\{(\boldsymbol{X}_j, Y_j^{\{b\}}) : j = 1, \ldots, n\}$.

The p-value in step B3* is then calculated as the proportion of $\check{T}^{\{b\}}_{n,\mathrm{SAVE}}(\boldsymbol{e_i})$ exceeding $\hat{T}_{n,\mathrm{SAVE}}(\boldsymbol{e_i})$.

2.3.2 The permutation test with DR

As discussed, the discretized kernel matrix for DR is $\boldsymbol{M}_{\mathrm{DR}} = \sum_{h=1}^{H}\sum_{k=1}^{H}\boldsymbol{B}_{hk}^2$ with $\boldsymbol{B}_{hk} = f_h^{1/2}f_k^{1/2}\{2\boldsymbol{I}_p - \mathrm{E}[(\boldsymbol{Z}-\tilde{\boldsymbol{Z}})(\boldsymbol{Z}-\tilde{\boldsymbol{Z}})^T|Y \in J_h, \tilde{Y} \in J_k]\}$, where $(\tilde{Y}, \tilde{\boldsymbol{Z}})$ is an independent copy of $(Y, \boldsymbol{Z})$. Note that $\boldsymbol{M}_{\mathrm{DR}}$ has the similar form as $\boldsymbol{M}_{\mathrm{SAVE}}$. One can thus follow $T_{\mathrm{SAVE}}(\boldsymbol{e_i})$ and define $T^0_{\mathrm{DR}}(\boldsymbol{e_i}) = \sum_{h=1}^{H}\sum_{k=1}^{H}(\boldsymbol{e_i}^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}_{hk}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}$ After some algebra, $T^0_{\mathrm{DR}}(\boldsymbol{e_i})$ can be rewritten as $T^0_{\mathrm{DR}}(\boldsymbol{e_i}) = 2\sum_{h=1}^{H} f_h^{-1}(\psi_h(\boldsymbol{e_i}))^2 + 4[\sum_{h=1}^{H} f_h^{-1}\phi_h(\boldsymbol{e_i})\phi_h^T(\boldsymbol{e_i})]^2$. We restate Proposition 1 of Yu and Dong (2015) as follows.

Proposition 2.3.2. *Suppose* $\text{Span}(\boldsymbol{M}_{DR}) = \mathcal{S}_{Y|\boldsymbol{Z}}$. *Then* $T^0_{DR}(\boldsymbol{e_i}) = 0$ *if and only if* $Y \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{X}_{-i}$.

The above result is parallel to Proposition 2.2.3 and Proposition 2.3.1. Please refer to Li and Wang (2007) for detailed discussions about the exhaustive coverage condition $\text{Span}(\boldsymbol{M}_{\text{DR}}) = \mathcal{S}_{Y|\boldsymbol{Z}}$.

At the sample level, we define the sample estimator of $nT^0_{\text{DR}}(\boldsymbol{e_i})$ as

$$\hat{T}^0_{n,\text{DR}}(\boldsymbol{e_i}) = 2n \sum_{h=1}^{H} \hat{f}_h^{-1}(\hat{\psi}_h(\boldsymbol{e_i}))^2 + 4n \left[\sum_{h=1}^{H} \hat{f}_h^{-1} \hat{\phi}_h(\boldsymbol{e_i}) \hat{\phi}_h^T(\boldsymbol{e_i}) \right]^2 .$$

We have seen in the proof of Theorem 2.3.1 that under $H_0 : Y \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{X}_{-i}$, $\sum_{h=1}^H \hat{f}_h^{-1}(\hat{\psi}_h(\boldsymbol{e_i}))^2 = O_p(n^{-1})$, $\sum_{h=1}^H \hat{f}_h^{-1}\hat{\phi}_h(\boldsymbol{e_i})\hat{\phi}_h^T(\boldsymbol{e_i}) = O_p(n^{-1})$ and $\hat{T}_{n,\text{SAVE}}(\boldsymbol{e_i}) \overset{D}{=} n\sum_{h=1}^H f_h^{-1}(\hat{\psi}_h(\boldsymbol{e_i}))^2$. Thus the first term in $\hat{T}^0_{n,\text{DR}}(\boldsymbol{e_i})$ is of order $O_p(1)$ while its second term has order $O_p(n^{-1})$. The effect of the second term vanishes and we have $\hat{T}^0_{n,\text{DR}}(\boldsymbol{e_i}) \overset{D}{=} 2\hat{T}_{n,\text{SAVE}}(\boldsymbol{e_i})$. Recall that the second term involves $\hat{T}_{n,\text{SIR}}(\boldsymbol{e_i}) = n\sum_{h=1}^H \hat{f}_h^{-1}\hat{\phi}_h(\boldsymbol{e_i})\hat{\phi}_h^T(\boldsymbol{e_i})$. Thus the effect of $\hat{T}_{n,\text{SIR}}(\boldsymbol{e_i})$ is ignored in the formulation of $\hat{T}^0_{n,\text{DR}}(\boldsymbol{e_i})$. To keep the balance between the term due to SAVE and the term due to SIR, we define the modified DR test statistic

$$\hat{T}_{n,\text{DR}}(\boldsymbol{e_i}) = 2n \sum_{h=1}^{H} \hat{f}_h^{-1}(\hat{\psi}_h(\boldsymbol{e_i}))^2 + 4n \sum_{h=1}^{H} \hat{f}_h^{-1} \hat{\phi}_h(\boldsymbol{e_i}) \hat{\phi}_h^T(\boldsymbol{e_i}). \tag{2.5}$$

As a sufficient dimension reduction method, an attractive property of DR is that it combines the strength of SIR and SAVE. Our formulation in (2.5) aims to inherit this desirable property. According to Theorem 1 of Yu and

Dong (2015), $\hat{T}_{n,\mathrm{DR}}(\boldsymbol{e_i}) = O_p(1)$ and converges to a sum of weighted $\chi^2(1)$ distributions under H_0, which we denote as $D^{\mathrm{DR}}_{\boldsymbol{e_i}}$.

Next we describe the DR-based permutation test algorithm.

C1. Based on the original sample, calculate $\hat{T}_{n,\mathrm{DR}}(\boldsymbol{e_i})$ as defined in (2.5).

C2. Fix $\{Y_j : j = 1, \ldots, n\}$. For $b = 1, \ldots, B$, denote $\{\boldsymbol{X}_j^{\{b\}} : j = 1, \ldots, n\}$ as the bth random permutation of $\{\boldsymbol{X}_j : j = 1, \ldots, n\}$. Then calculate $\hat{T}^{\{b\}}_{n,\mathrm{DR}}(\boldsymbol{e_i})$ based on the permuted sample $\{(\boldsymbol{X}_j^{\{b\}}, Y_j) : j = 1, \ldots, n\}$.

C3. Calculate the p-value $p_{\mathrm{DR}} = B^{-1} \sum_{b=1}^{B} I(\hat{T}^{\{b\}}_{n,\mathrm{DR}}(\boldsymbol{e_i}) > \hat{T}_{n,\mathrm{DR}}(\boldsymbol{e_i}))$. For given significance level α, reject $H_0 : Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{X}_{-i}$ if $p_{\mathrm{DR}} < \alpha$.

The permutation test procedure is justified by the next result. Its proof is similar to the proofs of Theorem 2.2.1 and Theorem 2.3.1, and is thus omitted.

Theorem 2.3.2. *Suppose $\boldsymbol{X}$ is normal. Then under $H_0 : Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{X}_{-i}$, we have $\hat{T}^{\{b\}}_{n,DR}(\boldsymbol{e_i}) \xrightarrow{D} D^{DR}_{\boldsymbol{e_i}}$.*

Similar to the SIR-based algorithm in Section 2.2.2, step C2 in the DR-based algorithm can be replaced by an equivalent step as follows

C2*. Fix $\boldsymbol{X}_j$, $j = 1, \ldots, n$. For $b = 1, \ldots, B$, denote $\{Y_j^{\{b\}} : j = 1, \ldots, n\}$ as the bth random permutation of $\{Y_j : j = 1, \ldots, n\}$. Then calculate $\check{T}^{\{b\}}_{n,\mathrm{DR}}(\boldsymbol{e_i})$ based on the permuted sample $\{(\boldsymbol{X}_j, Y_j^{\{b\}}) : j = 1, \ldots, n\}$.

The p-value in step C3* is then calculated as the proportion of $\check{T}^{\{b\}}_{n,\mathrm{DR}}(\boldsymbol{e_i})$ exceeding $\hat{T}_{n,\mathrm{DR}}(\boldsymbol{e_i})$.

2.4 Numerical studies

2.4.1 Comparisons with asymptotic tests

To demonstrate the performance of permutation test through SIR, SAVE and DR, we generate synthetic data from the following models.

$$\text{Model I: } Y = 3\sin(X_1) + 3\sin(X_p) + .1\epsilon.$$

$$\text{Model II: } Y = \text{sgn}(X_1 + X_p)\exp(X_2 + X_{p-1}) + .1\epsilon.$$

$$\text{Model III: } Y = (X_1 + X_p)^2 + X_2 + X_{p-1} + (X_3 + 1)^2\epsilon.$$

Here, $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ is multivariate normal with mean $\mu = 0$. The covariance between X_i and X_j is $.5^{|i-j|}$ for $1 \leq i, j \leq p$. The error ϵ is independent of $\boldsymbol{X}$, and follows the standard normal distribution. In Model II, $\text{sgn}(\cdot)$ denotes the sign function. The number of slices is $H = 5$, the predictor dimension is set as $p = 10$, and we consider sample size $n = 100$, 400, and 800. The number of permutations is fixed at $B = 500$.

Based on 500 repetitions, the proportions of p-values being smaller than the nominal level $\alpha = 0.05$ are reported. In each repetition, we test $H_0 : Y \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{X}_{-i}$ for $i = 1, 2, \ldots, p$. For predictors that are not predictive of Y, we are not expected to reject H_0, and the proportions should be close the nominal level. For predictors that are predictive of Y, we are expected to reject H_0 with a large probability, and the proportions are the estimated powers of the test. The estimated powers are boldfaced for easy reference. All the permutation

Table 2.1: *Model I results. Frequencies of rejecting H_0 with nominal* 5% *tests are reported.*

Model I	Method	Test	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
$n = 100$	SIR	Asymptotic	**1**	.092	.080	.098	.094	.078	.060	.084	.094	**1**
		Permutation	**1**	.024	.030	.030	.024	.026	.026	.038	.024	**1**
	SAVE	Asymptotic	**.310**	.024	.026	.012	.020	.032	.036	.022	.020	**.306**
		Permutation	**.438**	.048	.050	.040	.042	.068	.062	.052	.044	**.402**
	DR	Asymptotic	**.962**	.024	.030	.026	.032	.040	.024	.026	.026	**.950**
		Permutation	**.990**	.034	.044	.040	.042	.058	.044	.034	.040	**.972**
$n = 400$	SIR	Asymptotic	**1**	.056	.056	.062	.062	.064	.062	.074	.064	**1**
		Permutation	**1**	.024	.024	.022	.026	.028	.024	.042	.034	**1**
	SAVE	Asymptotic	**.996**	.048	.044	.036	.034	.044	.040	.032	.044	**1**
		Permutation	**.998**	.056	.064	.042	.046	.056	.048	.046	.054	**1**
	DR	Asymptotic	**1**	.032	.046	.024	.040	.046	.034	.042	.054	**1**
		Permutation	**1**	.036	.036	.028	.032	.050	.036	.044	.052	**1**

test results are based on permutation of the predictors as described in A2, B2 and C2. The results based on permutation of the responses (described in A2*, B2* and C2*) are very similar, and are thus omitted. The permutation tests with SIR, SAVE and DR are compared with their corresponding asymptotic tests, which have been introduced in Cook (2004), Shao, Cook and Weisberg (2007), and Yu and Dong (2015).

We summarize the simulation results of Model I in Table 2.1. For the active predictors X_1 and X_{10}, the estimated powers of the SIR-based tests and the DR-based tests are one or close to one. The SAVE-based tests have less desirable powers with $n = 100$. Although the powers improve as sample size increases, the SAVE-based tests are still not as powerful as the SIR and DR counterparts. It is well-known in the sufficient dimension reduction literature that SAVE is suboptimal when the link function between the response and the predictor is monotone or close to linear. We see from Table 2.1 that the SAVE-based tests inherit this limitation. For the inactive predictors X_2 through X_9, the estimated Type-I errors are close to the nominal level. When the sample size increases, the approximation of the estimated Type-I errors to the nominal level improves. For the SAVE-based and the DR-based tests in Model I, the overall performances of the proposed permutation tests and the existing asymptotic tests are very similar in terms of both the estimated powers and the estimated Type-I errors. In terms of computational efficiency, the permutation test is computationally intensive when the sample size is large, and the asymptotic test is preferable with large n. On the other hand, the asymptotic test seems to be too liberal for the SIR-based tests, especially when $n = 100$. Thus the permutation test provides a safe alternative to the asymptotic test when the sample size is small.

Table 2.2 summarizes the results of Model II. The permutation tests and the asymptotic tests again have very similar performances. For inactive predictors X_3 through X_8, the estimated Type-I errors are generally close to the nominal level. Note that X_1 and X_{10} are active due to the sign link function, while X_2 and X_9 are active due to the exponential link function. The SIR-based tests and the DR-based tests have large powers for all the active predictors. Because both the sign function and the exponential function are monotone, the SAVE-based tests suffer from low powers when $n = 100$, and

Table 2.2: *Model II results. Frequencies of rejecting H_0 with nominal* 5% *tests are reported.*

Model II	Method	Test	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
$n = 100$	SIR	Asymptotic	**1**	**1**	.084	.104	.068	.090	.084	.084	**1**	**1**
		Permutation	**.996**	**1**	.020	.018	.008	.018	.014	.024	**1**	**.998**
	SAVE	Asymptotic	**.098**	**.080**	.036	.040	.026	.026	.014	.020	**.088**	**.104**
		Permutation	**.150**	**.148**	.060	.082	.042	.048	.046	.052	**.164**	**.144**
	DR	Asymptotic	**.882**	**.906**	.034	.042	.030	.028	.020	.024	**.910**	**.862**
		Permutation	**.910**	**.936**	.038	.050	.028	.036	.026	.036	**.934**	**.892**
$n = 400$	SIR	Asymptotic	**1**	**1**	.046	.054	.048	.052	.044	.054	**1**	**1**
		Permutation	**1**	**1**	.016	.016	.012	.014	.018	.018	**1**	**1**
	SAVE	Asymptotic	**.926**	**.980**	.048	.034	.030	.036	.052	.044	**.988**	**.912**
		Permutation	**.946**	**.986**	.062	.042	.042	.040	.062	.060	**.994**	**.932**
	DR	Asymptotic	**1**	**1**	.046	.038	.036	.032	.042	.054	**1**	**1**
		Permutation	**1**	**1**	.042	.030	.024	.028	.034	.038	**1**	**1**

the powers of the SAVE-based tests improve when $n = 400$.

The results of Model III are reported in Table 2.3. Similar to what we have seen in Models I and II, the performances of the permutation tests and the asymptotic tests are close to each other. Note that X_1 and X_{10} are active due to the square link function in the regression mean, X_2 and X_9 are active due to the linear link function in the regression mean, and X_3 is active due to the heteroscedasticity in the regression variance. For inactive predictors X_4 through X_8, the estimated Type-I errors are close to the nominal level. Due to the complexity of the link functions, all three tests have low powers with $n = 100$. The powers increase as n increases to 400, and the improvement varies across different methods and different active predictors. For X_1 and X_{10} that appear in the square term, the SIR-based tests suffer from low powers even with $n = 400$. As a dimension reduction method, it is known that SIR is not

Table 2.3: *Model III results. Frequencies of rejecting H_0 with nominal 5% tests are reported.*

Model III	Method	Test	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
$n = 100$	SIR	Asymptotic	**.178**	**.600**	**.480**	.086	.090	.092	.070	.094	**.638**	**.162**
		Permutation	**.108**	**.486**	**.364**	.036	.036	.050	.034	.038	**.526**	**.082**
	SAVE	Asymptotic	**.232**	**.022**	**.034**	.020	.032	.024	.020	.020	**.014**	**.222**
		Permutation	**.318**	**.054**	**.078**	.040	.066	.052	.060	.048	**.040**	**.328**
	DR	Asymptotic	**.258**	**.226**	**.178**	.022	.038	.032	.022	.028	**.236**	**.236**
		Permutation	**.330**	**.286**	**.222**	.032	.064	.058	.034	.044	**.284**	**.314**
$n = 400$	SIR	Asymptotic	**.408**	**.998**	**.978**	.068	.058	.070	.064	.064	**.998**	**.436**
		Permutation	**.382**	**.998**	**.974**	.048	.032	.050	.052	.040	**.998**	**.418**
	SAVE	Asymptotic	**.966**	**.110**	**.150**	.038	.034	.040	.036	.038	**.102**	**.974**
		Permutation	**.976**	**.128**	**.186**	.044	.048	.058	.040	.046	**.124**	**.978**
	DR	Asymptotic	**.978**	**.972**	**.872**	.042	.022	.042	.058	.042	**.968**	**.976**
		Permutation	**.972**	**.976**	**.880**	.044	.026	.052	.050	.050	**.974**	**.982**

exhaustive when the link function is symmetric. Although the Type-I errors are not affected, we observe that the SIR-based tests will have large Type-II errors for those active predictors that appear in the symmetric terms. The SAVE-based tests have large powers for X_1 and X_{10} with $n = 400$. However, for X_2 and X_9 that appear in the linear term and for X_3 that appears in the variance term, the SAVE-based tests suffer from low powers. The DR-based tests enjoy the best overall performances, and have large powers for all active predictors when $n = 400$. As we have discussed in Section 2.3.2, the DR-based tests combine the strength of the SIR-based tests and the SAVE-based tests. The simulation results in Model III confirm that the DR-based tests enjoy the best performance across a wide range of link functions.

2.4.2 Permutation tests with transformations for non-normal predictors

Since the proposed permutation tests rely on the normality assumption of the predictors, it is interesting to examine their performances when the normality assumption is violated. In the presence of non-normal predictors, marginal predictor transformation has been suggested in the sufficient dimension reduction literature to facilitate estimation of the central space. See, for example, Wang, Guo and Zhu (2014), and Mai and Zou (2015). In this section, we study the effect of marginal predictor transformation on the permutation test for non-normal predictors. We focus on the setting when the original predictor $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ becomes normal after some suitable marginal transformations. Consider the following model

$$\text{Model IV: } Y = 3\sin(w(X_1)) + 3\sin(w(X_p)) + .1\epsilon.$$

Here the error ϵ is independent of $\boldsymbol{X}$ and follows the standard normal distribution. Denote the transformed predictors as $\boldsymbol{W} = (w(X_1), \ldots, w(X_p))^T$, where $w : \mathbb{R} \mapsto \mathbb{R}$ is a marginal transformation of the original predictor. Suppose $\boldsymbol{W}$ is multivariate normal with mean 0, and the covariance between $w(X_i)$ and $w(X_j)$ is $.5^{|i-j|}$ for $1 \leq i, j \leq p$. Consider three cases for the marginal distribution of X_i, $i = 1, \ldots, p$. In case (i), $X_i \sim \text{Uniform}(0, 1)$ has the standard uniform distribution, and the corresponding transformation is

$w(\cdot) = \Phi^{-1}\{F_u(\cdot)\}$. Here Φ^{-1} is the inverse of the standard normal distribution function and F_u is the standard uniform distribution function, which is the identity mapping. In case (ii), $X_i \sim \chi^2(2)$ follows the chi-squared distribution with 2 degrees of freedom, and the transformation is $w(\cdot) = \Phi^{-1}\{F_{\chi^2}(\cdot)\}$, where F_{χ^2} is the $\chi^2(2)$ distribution function. In case (iii), $X_i \sim \text{Cauchy}(0, 1)$ has the standard Cauchy distribution, and the corresponding transformation is $w(\cdot) = \Phi^{-1}\{F_c(\cdot)\}$, where F_c is the standard Cauchy distribution function. At the sample level, the marginal predictor transformation described above is known as the Yeo-Johnson transformation (Yeo & Johnson, 2000).

Suppose we observe $\{(\boldsymbol{X}_{(j)}, Y_{(j)}) : j = 1, \ldots, n\}$ from Model IV, where $\boldsymbol{X}_{(j)} = (X_{(j)1}, \ldots, X_{(j)p})^T$. Based on these observations, we want to test $H_0 : Y \perp\!\!\!\perp X | \boldsymbol{X}_{-i}$ for $i = 1, 2, \ldots, p$. Two approaches are considered for the permutation tests. The first approach is to carry out the proposed permutation test directly as in Section 2.4.1. The second approach is to address the non-normality of $\boldsymbol{X}$ through marginal predictor transformations. Since the actual transformation $w(\cdot)$ is unknown in practice, we estimate $w(\cdot)$ through the empirical distribution function. Specifically, let $r_{(j)i} = \sum_{k=1}^{n} I(X_{(k)i} \leq X_{(j)i})$ be the rank of $X_{(j)i}$ among $\{X_{(1)i}, \ldots, X_{(n)i}\}$. Then $F_u(X_{(j)i})$, $F_{\chi^2}(X_{(j)i})$ and $F_c(X_{(j)i})$ can all be estimated by $\hat{F}(X_{(j)i}) = r_{(j)i}/(n + 1)$. The marginally transformed predictors become $\hat{\boldsymbol{W}}_{(j)} = (\hat{W}_{(j)1}, \ldots, \hat{W}_{(j)p})^T$, where $\hat{W}_{(j)i} = \Phi^{-1}\{\hat{F}(X_{(j)i})\}$ for $j = 1, \ldots, n$ and $i = 1, \ldots, p$. This transformation is known as the Yeo-Johnson transformation (Yeo and Johnson, 2000) in the literature. The permutation test is then carried out based on these transformed predictors. We fix $n = 100$ and $p = 10$. The number of slices is set as $H = 5$ and the number of permutations is fixed at $B = 500$.

Based on 500 repetitions, we report the results of the SIR-based permutation test in Table 2.4. In cases (i) and (ii), the permutation tests with or without predictor transformation have similarly good performances, where the estimated powers are large and the estimated Type-I errors are close to the nominal level. When the predictors have the uniform distribution or the chi-squared distribution in this model, the proposed permutation test can still work well and is not sensitive to the normality assumption. For predictors with the Cauchy distribution in case (iii), the permutation test without predictor transformation no longer works well. We see that the estimated powers are low for active predictors X_1 and X_{10}, and the estimated Type-I errors for inactive predictors X_2 and X_9 are inflated. On the other hand, the permutation test with marginal predictor transformation performs very good, implying that predictor transformation can be beneficial when the normality assumption is violated.

Table 2.4: *Model IV with SIR-based permutation test. Frequencies of rejecting H_0 with nominal 5% tests are reported.*

Case	Transformation	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
(i)	No	**1**	.026	.028	.036	.020	.022	.026	.038	.024	**1**
	Yes	**1**	.026	.034	.036	.026	.028	.026	.038	.024	**1**
(ii)	No	**.988**	.032	.020	.030	.018	.032	.018	.026	.030	**.990**
	Yes	**1**	.022	.028	.034	.026	.028	.028	.030	.026	**1**
(iii)	No	**.476**	.140	.046	.054	.042	.038	.046	.038	.132	**.462**
	Yes	**1**	.028	.028	.032	.030	.032	.030	.032	.024	**1**

2.4.3 A real data analysis

In this section, we consider a real data analysis with the horse mussel data. This data set contains $n = 82$ observations collected in an ecological study in New Zealand. The response is a mussel's muscle mass in g. The four predictors are the shell height in mm, the shell length in mm, the shell mass in g, and the shell width in mm. This data set is available in R under the `dr` package. It has been studied before in Cook & Weisberg (1994) and Wang et al. (2014). Due to the potential non-normality in the predictors, we apply the Yeo-Johnson transformation to all four predictors. In the discussions below, the response variable muscle mass is denoted as M. For the marginally transformed predictors height, length, mass and width, we denote them as Ht, L, S and W, respectively.

Without fitting a model, we want to decide which of the four predictors are predictive of the response. Consider the backward elimination procedure with the DR-based permutation test. This procedure can be viewed as a model-free version of the classical backward elimination procedure in linear models. As we have seen in section 2.4.1, the DR-based test has the best performance across a wide range of link functions. Fix the number of permutations to be $B = 2000$ and set the nominal level at .05. In the first iteration, the p-values are .118, .011, .084 and .041 for Ht, L, S and W. Thus we delete the least significant predictor Ht. In the second iteration, the p-values for L, S and W are .054, .025 and .032, and the least significant predictor L is deleted. In the third iteration, the p-values for S and W are .010 and .011. Using .05 as the nominal level, no more predictors can be deleted. The conclusion is thus $M \perp\!\!\!\perp (Ht, L, S, W)|(S, W)$, or the mussel's muscle mass can be fully predicted by the shell mass and the shell width.

Next we consider a modified version of the problem. We keep the first two predictors shell height Ht and shell length L, and replace the remaining two predictors with two independent standard normal predictors Z_1 and Z_2. We refer to (Ht, L, Z_1, Z_2) as the modified predictors. We rerun the backward elimination procedure to decide which of the modified predictors are predictive of the response. Based on the same DR-based permutation test with $B = 2000$ repetitions, the p-values for Ht, L, Z_1 and Z_2 are .051, .065, .624, and 0.966. The least significant predictor Z_2 is thus deleted in the first iteration. In the second iteration, the p-values for Ht, L and Z_1 are .038, .047 and .603, and Z_1 is deleted. In the third iteration, the p-values for Ht and L are .032 and .040.

With both p-values smaller than .05, no more predictors can be deleted. The conclusion becomes $M \perp\!\!\!\perp (Ht, L, Z_1, Z_2)|(Ht, L)$, or the mussel's muscle mass can be fully predicted by the shell height and the shell length.

To understand the seemingly contradicting results in the two settings, the pairwise scatterplot matrix of Ht, L, S, W and M is provided in the left panel of Figure 2.1. We see that the four variables in the original data are highly correlated, and they are all highly correlated with the response M. This scatterplot matrix provides additional insight to our conclusion in the first setting. Although Ht and L are marginally correlated with the response M, due to their large correlations with S and W, we do not have enough evidence to reject the hypothesis that Ht and L are independent of M conditioning on S and W. In the right panel of Figure 2.1, we provide the pairwise scatterplot matrix for the modified predictors Ht, L, Z_1, Z_2 together with the response M. It is obvious in the second setting that M is independent of Z_1 and Z_2, and we can reach the conclusion that $M \perp\!\!\!\perp (Ht, L, Z_1, Z_2)|(Ht, L)$, or M depends on (Ht, L, Z_1, Z_2) only through Ht and L.

2.5 Proofs

Proof of Proposition 2.2.2. Let $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta}$. Proposition 1.2.1 states that $\mathrm{Span}(\boldsymbol{\beta}) = \mathcal{S}_{Y|\boldsymbol{X}}$ and it follows that $Y \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{\beta}^T\boldsymbol{X}$. Note that e_i is the ith column of I_p. Denote $\boldsymbol{I}_{-i} \in \mathbb{R}^{p\times(p-1)}$ as the $p-1$ columns of $\boldsymbol{I}_p$ other than the

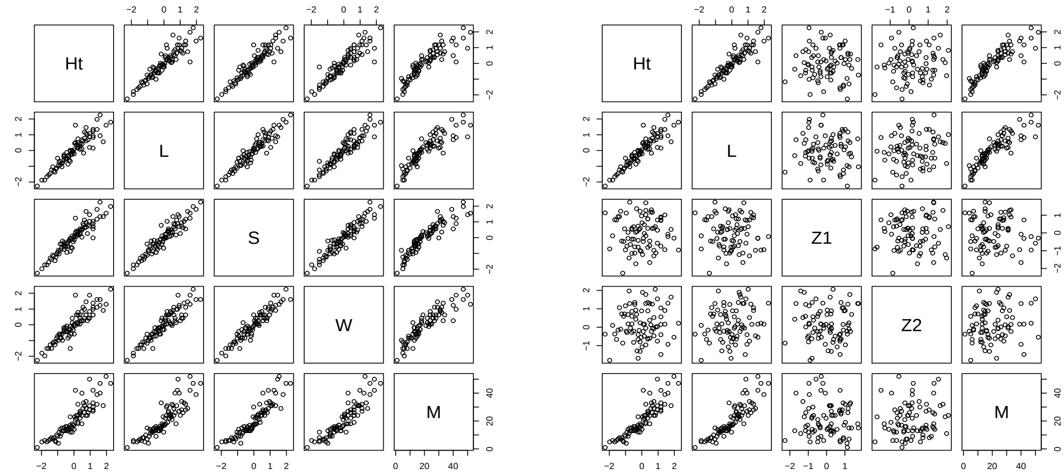

Figure 2.1: Left panel: the scatterplot matrix of the response M together with the transformed predictors Ht, L, S and W. Right panel: the scatterplot matrix after replacing predictors S and W with independent standard normal predictors Z_1 and Z_2.

ith column. Then $e_i^T \boldsymbol{X} = X_i$ and $\boldsymbol{I}_{-i}^T \boldsymbol{X} = \boldsymbol{X}_{-i}$. For the "if" part, $Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{X}_{-i}$ implies $Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{I}_{-i}^T \boldsymbol{X}$. Thus $\text{Span}(\boldsymbol{\beta}) \subseteq \text{Span}(\boldsymbol{I}_{-i})$, and $\boldsymbol{\beta} = \boldsymbol{I}_{-i}\boldsymbol{C}$ for some $\boldsymbol{C} \in \mathbb{R}^{(p-1)\times d}$. Together with the fact that $e_i^T \boldsymbol{I}_{-i} = 0$ and $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta}$, we have $e_i^T \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta} = e_i^T \boldsymbol{I}_{-i}\boldsymbol{C} = 0$. For the "only if" part, $e_i^T \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta} = 0$ implies that $\boldsymbol{\beta}^T e_i = 0$. Since e_i is the ith column of I_p, the ith column of $\boldsymbol{\beta}^T = \boldsymbol{\beta}^T I_p$ is 0. We denote $\boldsymbol{\beta}_{-i} \in \mathbb{R}^{(p-1)\times d}$ as the $p-1$ rows of $\boldsymbol{\beta}$ other than the ith row. Now that the ith row of $\boldsymbol{\beta}$ is 0, it is easy to see that $\boldsymbol{\beta}^T \boldsymbol{X} = \boldsymbol{\beta}_{-i}^T \boldsymbol{X}_{-i}$. Together with the fact that $Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{\beta}^T \boldsymbol{X}$, we have $Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{\beta}_{-i}^T \boldsymbol{X}_{-i}$, which then leads to $Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{X}_{-i}$. □

Proof of Proposition 2.2.3. From Proposition 2.2.2, all we need to show is that $T_{\text{SIR}}(e_i) = 0$ if and only if $e_i^T \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta} = 0$, where $\text{Span}(\boldsymbol{\eta}) = \mathcal{S}_{Y|\boldsymbol{Z}}$. For the "if" part, note that $\boldsymbol{\xi}_h \in \text{Span}(\boldsymbol{\eta})$. Thus $e_i^T \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta} = 0$ implies $e_i^T \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\xi}_h = 0$ for $h = 1, \ldots, H$, and $T_{\text{SIR}}(e_i) = \sum_{h=1}^{H} f_h (e_i^T \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\xi}_h)(e_i^T \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\xi}_h)^T = 0$ as a result. For the "only if" part, $T_{\text{SIR}}(e_i) = 0$ implies that $e_i^T \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\xi}_h = 0$ for $h = 1, \ldots, H$. Because $\text{Span}(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_H) = \text{Span}(\boldsymbol{M}_{\text{SIR}}) = \mathcal{S}_{Y|\boldsymbol{Z}} = \text{Span}(\boldsymbol{\eta})$, we have $e_i^T \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta} = 0$. □

Proof of Proposition 2.2.4. Denote $\boldsymbol{I}_{-i} \in \mathbb{R}^{p\times(p-1)}$ as the $p-1$ columns of $\boldsymbol{I}_p$ other than the ith column e_i. Then we have $\boldsymbol{I}_{-i}^T \boldsymbol{X} = \boldsymbol{X}_{-i}$. Hence $Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{X}_{-i}$ leads to $Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{I}_{-i}^T \boldsymbol{X}$, or equivalently $Y \perp\!\!\!\perp \boldsymbol{Z} | \boldsymbol{I}_{-i}^T \boldsymbol{\Sigma}^{1/2} \boldsymbol{Z}$. The last conditional independency guarantees that $Y \perp\!\!\!\perp e_i^T \boldsymbol{\Sigma}^{-1/2}\boldsymbol{Z}|\boldsymbol{I}_{-i}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}$. On the other hand, we have $\mathrm{Cov}(e_i^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{Z}, \boldsymbol{I}_{-i}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}) = e_i^T\boldsymbol{\Sigma}^{-1/2}\mathrm{Var}(\boldsymbol{Z})\boldsymbol{\Sigma}^{1/2}\boldsymbol{I}_{-i} = e_i^T\boldsymbol{I}_{-i} = 0$. The normality assumption thus implies $e_i^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{I}_{-i}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}$. Together with $Y \perp\!\!\!\perp e_i^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{Z}|\boldsymbol{I}_{-i}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}$, we have $Y \perp\!\!\!\perp (e_i^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{Z}, \boldsymbol{I}_{-i}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z})$. It follows immediately that $Y \perp\!\!\!\perp e_i^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{Z}$. □

Proof of Theorem 2.2.1. Note that $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are the same before and after the permutation. Recall that $\hat{\phi}_h(e_i) = n^{-1}\sum_{j=1}^n e_i^T\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X}_j - \hat{\boldsymbol{\mu}})I(Y_j \in J_h)$ and denote $\hat{\phi}_h^*(e_i) = n^{-1}\sum_{j=1}^n e_i^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_j - \boldsymbol{\mu})I(Y_j \in J_h)$. Then $\hat{\phi}_h(e_i) = O_p(n^{-1/2})$, $\hat{\phi}_h^*(e_i) = O_p(n^{-1/2})$ and $\hat{\phi}_h(e_i) = \hat{\phi}_h^*(e_i) + O_p(n^{-1})$. We also have $\hat{f}_h^{-1} - f_h^{-1} = O_p(n^{-1/2})$. It follows that

$$\hat{T}_{n,\mathrm{SIR}}(e_i) = n\sum_{h=1}^H \hat{f}_h^{-1}\hat{\phi}_h(e_i)\hat{\phi}_h^T(e_i) \stackrel{D}{=} n\sum_{h=1}^H f_h^{-1}\hat{\phi}_h^*(e_i)(\hat{\phi}_h^*(e_i))^T, \tag{2.6}$$

where "$\stackrel{D}{=}$" means having the same asymptotic distribution. Let $\hat{\phi}_h^{\{b\}}(e_i) = n^{-1}\sum_{j=1}^n e_i^T\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{X}_j^{\{b\}} - \hat{\boldsymbol{\mu}})I(Y_j \in J_h)$ and $\hat{\phi}_h^{\{b\}*}(e_i) = n^{-1}\sum_{j=1}^n e_i^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_j^{\{b\}} - \boldsymbol{\mu})I(Y_j \in J_h)$. Then $\hat{\phi}_h^{\{b\}}(e_i) = O_p(n^{-1/2})$, $\hat{\phi}_h^{\{b\}*}(e_i) = O_p(n^{-1/2})$ and $\hat{\phi}_h^{\{b\}}(e_i) = \hat{\phi}_h^{\{b\}*}(e_i) + O_p(n^{-1})$. Thus we have

$$\hat{T}_{n,\mathrm{SIR}}^{\{b\}}(e_i) = n\sum_{h=1}^H \hat{f}_h^{-1}\hat{\phi}_h^{\{b\}}(e_i)(\hat{\phi}_h^{\{b\}}(e_i))^T \stackrel{D}{=} n\sum_{h=1}^H f_h^{-1}\hat{\phi}_h^{\{b\}*}(e_i)(\hat{\phi}_h^{\{b\}*}(e_i))^T \tag{2.7}$$

To prove $\hat{T}_{n,\mathrm{SIR}}(e_i) \stackrel{D}{=} \hat{T}_{n,\mathrm{SIR}}^{\{b\}}(e_i)$, we see from (2.6) and (2.7) that it suffices to show that $\{\hat{\phi}_1^{\{b\}*}(e_i), \ldots, \hat{\phi}_H^{\{b\}*}(e_i)\} \overset{D}{=} \{\hat{\phi}_1^*(e_i), \ldots, \hat{\phi}_H^*(e_i)\}$. We express

$$\hat{\phi}_h^*(e_i) = n^{-1} \sum_{j=1}^{n} u_{jh} \text{ and } \hat{\phi}_h^{\{b\}*}(e_i) = n^{-1} \sum_{j=1}^{n} u_{jh}^{\{b\}}, h = 1, \ldots, H, \tag{2.8}$$

where $u_{jh} = e_i^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_j - \boldsymbol{\mu}) I(Y_j \in J_h)$ and $u_{jh}^{\{b\}} = e_i^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_j^{\{b\}} - \boldsymbol{\mu}) I(Y_j \in J_h)$. Denote $\boldsymbol{U}_j = (u_{j1}, \ldots, u_{jH})^T \in \mathbb{R}^H$ and $\boldsymbol{U}_j^{\{b\}} = (u_{j1}^{\{b\}}, \ldots, u_{jH}^{\{b\}})^T \in \mathbb{R}^H$. From (2.8) and the multivariate continuous mapping theorem, we know $\{\hat{\phi}_1^{\{b\}*}(e_i), \ldots, \hat{\phi}_H^{\{b\}*}(e_i)\} \overset{D}{=} \{\hat{\phi}_1^*(e_i), \ldots, \hat{\phi}_H^*(e_i)\}$ as long as $\{\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n\} \overset{D}{=} \{\boldsymbol{U}_1^{\{b\}}, \ldots, \boldsymbol{U}_n^{\{b\}}\}$.

From the independency of the random sample and the result of Proposition 2.2.4, $\{e_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X}_1, \ldots, e_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X}_n, Y_1, \ldots, Y_n\}$ are mutually independent. By noticing $\boldsymbol{X}_j \overset{D}{=} \boldsymbol{X}_j^{\{b\}}$, we have the following facts: a) $\boldsymbol{U}_j \overset{D}{=} \boldsymbol{U}_j^{\{b\}}$ for any fixed j; b) $U_1, \ldots, \boldsymbol{U}_n$ is i.i.d.; c) $\boldsymbol{U}_1^{\{b\}}, \ldots, \boldsymbol{U}_n^{\{b\}}$ is i.i.d. Combine a), b) and c) and we have $\{\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n\} \overset{D}{=} \{\boldsymbol{U}_1^{\{b\}}, \ldots, \boldsymbol{U}_n^{\{b\}}\}$. □

Proof of Theorem 2.3.1. Define $\boldsymbol{K} = (k_1, \ldots, k_H)^T \in \mathbb{R}^H$ with the hth element as $k_h = f_h^{-1/2}[\psi_h(e_i) + f_h^{-1} \phi_h(e_i) \phi_h^T(e_i)]$. Then $T_{\text{SAVE}}(e_i) = \boldsymbol{K}^T \boldsymbol{K}$. The sample version $\hat{\boldsymbol{K}} = (\hat{k}_1, \ldots, \hat{k}_H)^T$ has the hth element as $\hat{k}_h = \hat{f}_h^{-1/2}[\hat{\psi}_h(e_i) + \hat{f}_h^{-1} \hat{\phi}_h(e_i) \hat{\phi}_h^T(e_i)]$. Thus we have $\hat{T}_{n,\text{SAVE}}(e_i) = n \hat{\boldsymbol{K}}^T \hat{\boldsymbol{K}}$. Following the proof of Theorem 1 in Shao, Cook and Weisberg (2007), we know $\hat{k}_h = O_p(n^{-1/2})$ under $H_0 : Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{X}_{-i}$. By noting that $\hat{\psi}_h(e_i) = O_p(n^{-1/2})$ and $\hat{\phi}_h(e_i) = O_p(n^{-1/2})$

under H_0, we have

$$\hat{T}_{n,\mathrm{SAVE}}(e_i) = n\sum_{h=1}^{H}\hat{f}_h^{-1}[\hat{\psi}_h(e_i)+\hat{f}_h^{-1}\hat{\phi}_h(e_i)\hat{\phi}_h^T(e_i)]^2 \overset{D}{=} n\sum_{h=1}^{H}\hat{f}_h^{-1}(\hat{\psi}_h(e_i))^2.$$

Recall that $\hat{\psi}_h(e_i) = n^{-1}\sum_{j=1}^{n} e_i^T[\hat{\mathbf{\Sigma}}^{-1}(\hat{\mathbf{\Sigma}}-(\boldsymbol{X}_j-\hat{\boldsymbol{\mu}})(\boldsymbol{X}_j-\hat{\boldsymbol{\mu}})^T)\hat{\mathbf{\Sigma}}^{-1}]e_i I(Y_j \in J_h)$. Define $\hat{\psi}_h^*(e_i) = n^{-1}\sum_{j=1}^{n} e_i^T[\mathbf{\Sigma}^{-1}(\mathbf{\Sigma}-(\boldsymbol{X}_j-\boldsymbol{\mu})(\boldsymbol{X}_j-\boldsymbol{\mu})^T)\mathbf{\Sigma}^{-1}]e_i I(Y_j \in J_h)$. It can be shown that under H_0, $\hat{\psi}_h^*(e_i) = O_p(n^{-1/2})$ and $\hat{\psi}_h(e_i) = \hat{\psi}_h^*(e_i) + O_p(n^{-1})$. Together with $\hat{f}_h^{-1} - f_h^{-1} = O_p(n^{-1/2})$, we have under H_0,

$$\hat{T}_{n,\mathrm{SAVE}}(e_i) \overset{D}{=} n\sum_{h=1}^{H} f_h^{-1}(\hat{\psi}_h^*(e_i))^2. \tag{2.9}$$

After the random permutation, use similar argument and we have under H_0

$$\hat{T}_{n,\mathrm{SAVE}}^{\{b\}}(e_i) \overset{D}{=} n\sum_{h=1}^{H} f_h^{-1}(\hat{\psi}_h^{\{b\}*}(e_i))^2, \tag{2.10}$$

where $\hat{\psi}_h^{\{b\}*}(e_i) = n^{-1}\sum_{j=1}^{n} e_i^T\ [\mathbf{\Sigma}^{-1}(\mathbf{\Sigma}-(\boldsymbol{X}_j^{\{b\}}-\boldsymbol{\mu})(\boldsymbol{X}_j^{\{b\}}-\boldsymbol{\mu})^T)\mathbf{\Sigma}^{-1}]e_i\ I(Y_j \in J_h)$. To prove $\hat{T}_{n,\mathrm{SAVE}}(e_i) \overset{D}{=} \hat{T}_{n,\mathrm{SAVE}}^{\{b\}}(e_i)$, we see from (2.9) and (2.10) that it suffices to show that $\{\hat{\psi}_1^{\{b\}*}(e_i),\ldots,\hat{\psi}_H^{\{b\}*}(e_i)\} \overset{D}{=} \{\hat{\psi}_1^*(e_i),\ldots,\hat{\psi}_H^*(e_i)\}$. We express

$$\hat{\psi}_h^*(e_i) = n^{-1}\sum_{j=1}^{n} v_{jh} \text{ and } \hat{\psi}_h^{\{b\}*}(e_i) = n^{-1}\sum_{j=1}^{n} v_{jh}^{\{b\}}, h = 1,\ldots,H, \tag{2.11}$$

where $v_{jh} = e_i^T[\mathbf{\Sigma}^{-1}(\mathbf{\Sigma}-(\boldsymbol{X}_j-\boldsymbol{\mu})(\boldsymbol{X}_j-\boldsymbol{\mu})^T)\mathbf{\Sigma}^{-1}]e_i I(Y_j \in J_h)$ and $v_{jh}^{\{b\}} = e_i^T[\mathbf{\Sigma}^{-1}(\mathbf{\Sigma}-(\boldsymbol{X}_j^{\{b\}}-\boldsymbol{\mu})(\boldsymbol{X}_j^{\{b\}}-\boldsymbol{\mu})^T)\mathbf{\Sigma}^{-1}]e_i I(Y_j \in J_h)$. Denote $\boldsymbol{V}_j = (v_{j1},\ldots,v_{jH})^T \in \mathbb{R}^H$ and $\boldsymbol{V}_j^{\{b\}} = (v_{j1}^{\{b\}},\ldots,v_{jH}^{\{b\}})^T \in \mathbb{R}^H$. From (2.11) and the multivariate continuous mapping theorem, we know $\{\hat{\psi}_1^{\{b\}*}(e_i),\ldots,\hat{\psi}_H^{\{b\}*}(e_i)\} \overset{D}{=} \{\hat{\psi}_1^*(e_i),\ldots,\hat{\psi}_H^*(e_i)\}$

if we have $\{\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n\} \stackrel{D}{=} \{\boldsymbol{V}_1^{\{b\}}, \ldots, \boldsymbol{V}_n^{\{b\}}\}$. The latter is true due to the following facts: a) $\boldsymbol{V}_j \stackrel{D}{=} \boldsymbol{V}_j^{\{b\}}$ for any fixed j; b) $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n$ is i.i.d.; c) $\boldsymbol{V}_1^{\{b\}}, \ldots, \boldsymbol{V}_n^{\{b\}}$ is i.i.d. $\square$

CHAPTER 3

CLUSTER-BASED LEAST ABSOLUTE DEVIATION REGRESSION FOR DIMENSION REDUCTION

3.1 Introduction

Least absolute deviation (LAD) regression is an important tool for regression analysis. In the case of heavy-tailed error distribution, asymmetric error distribution, and in the presence of outliers in the response variable, it is well-known that LAD is preferable to the ordinary least squares (OLS) estimation in linear regression models. Compared to the widespread applications of OLS estimation, the use of LAD is limited historically due to its computational challenges. With the availability of high-speed modern computation and the advancement of ℓ_1-type regression algorithms, LAD for linear regression has seen much development in recent years. Among others, Koenker and Bassett (1978) investigated the asymptotic analysis of LAD regression as a special case of quantile regression, a comparison between the computational aspects of LAD and OLS was discussed in Portnoy and Koenker (1997), and a survey of LAD regression was provided in Narula and Wellington (1982).

As a popular semiparametric method, single-index model (Brillinger, 1983) keeps the index structure of the classical linear regression, while the identity link function in linear regression is generalized. Namely, the response depends on a linear combination of the predictors with some unknown link function. The asymptotic properties of fitting single-index models through nonlinear least squares have been systematically treated in Ichimura (1993) and Härdle et al. (1993). More recently, the single-index quantile regression was investigated in Wu et al. (2010).

Li and Duan (1989) revealed an interesting fact about direction estimation in single-index models. Instead of estimating both the unknown link function and the unknown index for dimension reduction, it was shown that one can directly recover the linear combination through OLS without estimating the unknown link function. Although the underlying link function in single-index models is not the identity link, one can use OLS for the purpose of dimension reduction as if the identity link assumption still held. Li and Duan (1989) referred to this phenomenon as "*regression analysis under link violation*". A global linear conditional mean assumption was introduced in Li and Duan (1989) to facilitate their analysis. Li et al. (2004) proposed a cluster-based OLS procedure, which relaxed the global linear conditional mean assumption to the more flexible local linearity assumptions at the cluster level.

Both the procedures in Li and Duan (1989) and Li et al. (2004) will fail when the error distribution is heavy-tailed or contaminated by outliers. To address this limitation, we propose a cluster-based LAD procedure for dimension reduction in single-index models. Our proposal naturally inherits the benefits of LAD over OLS in linear regression, and is insensitive to the error distribution or the outliers in the response.

The rest of the chapter is organized as follows. LAD and cluster-based LAD for dimension reduction are developed in Section 3.2 and Section 3.3 respectively. Numerical studies are presented in Section 3.4.

3.2 LAD for dimension reduction

3.2.1 LCM assumption and LAD for dimension reduction

For univariate continuous response Y and p-dimensional predictor $\boldsymbol{X}$, suppose Y and $\boldsymbol{X}$ follow the single-index model

$$Y = g(\alpha + \boldsymbol{\beta}^T \boldsymbol{X}) + \varepsilon. \tag{3.1}$$

Here $\alpha \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^p$, the error ε is independent of $\boldsymbol{X}$, and $g : \mathbb{R} \mapsto \mathbb{R}$ is an unknown link function. When $g(\cdot)$ is the identity link, model (3.1) becomes the linear regression model.

With identity link function in (3.1), we estimate $\boldsymbol{\beta}$ through the classical least squares optimization problem

$$\arg\min E\{(Y - a - \boldsymbol{b}^T \boldsymbol{X})^2\} \text{ over } a \in \mathbb{R} \text{ and } \boldsymbol{b} \in \mathbb{R}^p.$$

The minimizer of $\boldsymbol{b}$ is known as the OLS estimator, and we denote it as $\boldsymbol{\beta}_{\text{OLS}}$. It turns out that $\boldsymbol{\beta}_{\text{OLS}} = \boldsymbol{\Sigma}^{-1}\text{cov}(\boldsymbol{X}, Y)$, where $\boldsymbol{\Sigma} = \text{Var}(\boldsymbol{X})$. When $g(\cdot)$ is the identity link, $\text{cov}(\boldsymbol{X}, Y) = \boldsymbol{\Sigma}\boldsymbol{\beta}$ and $\boldsymbol{\beta}_{\text{OLS}}$ becomes exactly $\boldsymbol{\beta}$ in (3.1). Under the linear conditional mean (LCM) assumption that

$$E(\boldsymbol{X}|\boldsymbol{\beta}^T \boldsymbol{X}) \text{ is a linear function of } \boldsymbol{\beta}^T \boldsymbol{X}, \tag{3.2}$$

Li and Duan (1989) showed that $\boldsymbol{\beta}_{\text{OLS}}$ is proportional to $\boldsymbol{\beta}$ for unknown $g(\cdot)$

in (3.1). Hence OLS can be used to recover $\boldsymbol{\beta}$ without estimating the link function.

Let $\{(\boldsymbol{X}_{(c)}, Y_{(c)}) : c = 1, 2, \ldots, k\}$ be a partition of $(\boldsymbol{X}, Y)$ according to $\boldsymbol{X}$. Li et al. (2004) considered

$$\arg\min E\{(Y_{(c)} - a_{(c)} - \boldsymbol{b}_{(c)}^T \boldsymbol{X}_{(c)})^2\} \text{ over } a_{(c)} \in \mathbb{R} \text{ and } \boldsymbol{b}_{(c)} \in \mathbb{R}^p.$$

Denote the minimizer of $\boldsymbol{b}_{(c)}$ in the cth cluster as $\boldsymbol{\beta}_{\text{OLS},(c)}$, $c = 1, 2, \ldots, k$. Assuming that the LCM assumption holds for each cluster $\boldsymbol{X}_{(c)}$, then $\boldsymbol{\beta}_{\text{OLS},(c)}$ will be proportional to $\boldsymbol{\beta}$. At the sample level, the sample estimators $\hat{\boldsymbol{\beta}}_{\text{OLS},(c)}$ across different clusters are synthesized to get the final cluster-based OLS estimator of $\boldsymbol{\beta}$. Please refer to Li et al. (2004) for the details of the sample cluster-based OLS algorithm.

Consider sample level linear regression model $Y_i = \alpha + \boldsymbol{\beta}^T \boldsymbol{X}_i + \varepsilon_i$, $i = 1, \ldots, n$. The LAD regression then minimizes the mean absolute deviation $n^{-1} \sum_{i=1}^n |\varepsilon_i| = n^{-1} \sum_{i=1}^n |Y_i - \alpha - \boldsymbol{\beta}^T \boldsymbol{X}_i|$. The LAD estimators are maximum likelihood estimates when the error ε_i's are i.i.d. double exponential distribution. The main result in this section is that LAD can be used to recover $\boldsymbol{\beta}$ in the single-index model (3.1).

Proposition 3.2.1. *Under model (3.1) and the LCM assumption (1.3.1), consider minimization problem*

$$\min L(a, \boldsymbol{b}) \text{ over } a \in \mathbb{R} \text{ and } \boldsymbol{b} \in \mathbb{R}^p, \text{ where } L(a, \boldsymbol{b}) = E\{|Y - a - \boldsymbol{b}^T \boldsymbol{X}|\}.$$

Suppose the minimizer of $\boldsymbol{b}$ *is unique and denote it as* $\boldsymbol{\beta}_{LAD}$*. Then* $\boldsymbol{\beta}_{LAD}$ *is proportional to* $\boldsymbol{\beta}$ *in (3.1).*

PROOF. By the law of iterative expectation and Jensen's inequality, we have

$$L(a, \boldsymbol{b}) = E\left[E\left\{|Y - a - \boldsymbol{b}^T\boldsymbol{X}| \,\middle|\, Y, \boldsymbol{\beta}^T\boldsymbol{X}\right\}\right] \geq E\left\{\left|E(Y - a - \boldsymbol{b}^T\boldsymbol{X}|Y, \boldsymbol{\beta}^T\boldsymbol{X})\right|\right\}.$$

Under model (3.1), Y is independent of $\boldsymbol{b}^T\boldsymbol{X}$ given $\boldsymbol{\beta}^T\boldsymbol{X}$, and $E(\boldsymbol{b}^T\boldsymbol{X}|Y, \boldsymbol{\beta}^T\boldsymbol{X})$ becomes $E(\boldsymbol{b}^T\boldsymbol{X}|\boldsymbol{\beta}^T\boldsymbol{X})$. The LCM assumption (1.3.1) implies that $E(b^T\boldsymbol{X}|\boldsymbol{\beta}^T\boldsymbol{X}) = \tau\boldsymbol{\beta}^T\boldsymbol{X}$ for some constant τ. Together, we have

$$L(a, \boldsymbol{b}) = E\{|Y - a - \boldsymbol{b}^T\boldsymbol{X}|\} \geq E\{|Y - a - \tau\boldsymbol{\beta}^T\boldsymbol{X}|\}.$$

Thus if $\boldsymbol{\beta}_{\text{LAD}}$ minimizes $L(a, \boldsymbol{b})$, it has to be proportional to $\boldsymbol{\beta}$ given that the minimizer is unique. □

3.2.2 An illustration: the role of LCM assumption

We consider a toy example to fix the ideas. Suppose there are two predictors X_1 and X_2. Let $X_1 \sim \text{Uniform}(0, 1)$, $X_2 = X_1 + e$ with $e \sim \text{Uniform}(-0.3, 0.3)$, and $Y = \log(X_1) + \varepsilon$. Then $\boldsymbol{X} = (X_1, X_2)^T$, $\boldsymbol{\beta} = (1, 0)^T$ and $\boldsymbol{\beta}^T\boldsymbol{X} = X_1$ in model (3.1). The population correlation between X_1 and X_2 is calculated as 0.857. We want to compare the performance of the OLS estimator $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ and the LAD estimator $\hat{\boldsymbol{\beta}}_{\text{LAD}}$. Since $\boldsymbol{\beta}$ has unit length, we normalize $\hat{\boldsymbol{\beta}}_{\text{OLS}}$

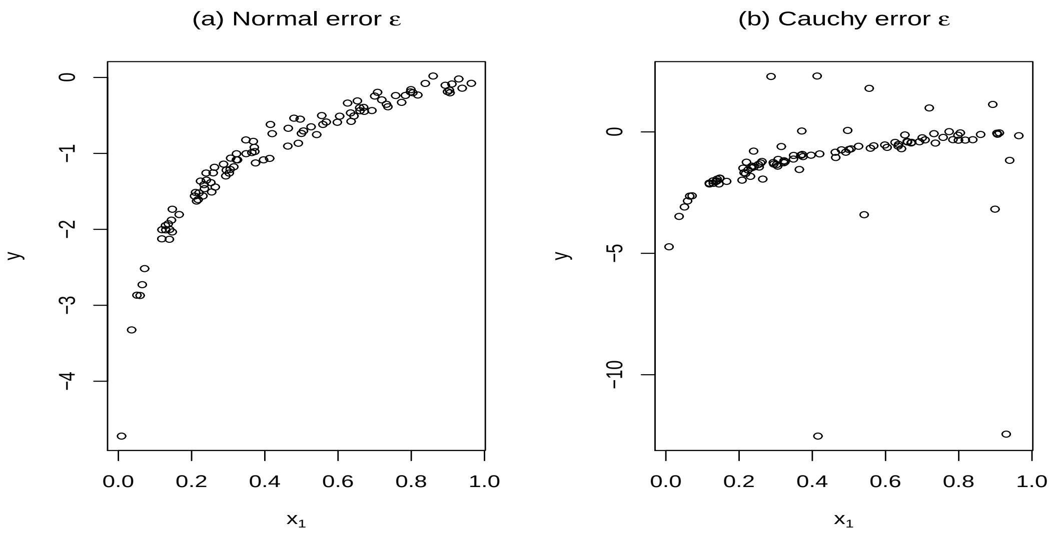

Figure 3.1: Scatterplot of Y versus X_1 for model $Y = \log(X_1) + \varepsilon$.

and $\hat{\boldsymbol{\beta}}_{\text{LAD}}$ such that they also have unit length. The sample size is fixed at $n = 100$.

First we consider the case with $\varepsilon \sim \text{Normal}(0, 0.1^2)$. This is exactly the case considered in Section 2.4 of Li et al. (2004). In a typical run as depicted in panel (a) of Figure 3.1, we get $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (0.999, 0.053)^T$ and $\hat{\boldsymbol{\beta}}_{\text{LAD}} = (0.992, 0.126)^T$. Although we have a nonlinear logarithm link function and there is high correlation between X_1 and X_2, both OLS and LAD successfully recover the true direction $\boldsymbol{\beta} = (1, 0)^T$. Next we consider the case when ε follows a Cauchy distribution with location parameter 0 and scale parameter 0.1. In a typical run as depicted in panel (b) of Figure 3.1, we get $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (0.953, 0.302)^T$ and $\hat{\boldsymbol{\beta}}_{\text{LAD}} = (0.999, 0.025)^T$. Note that Y has a much wider range in this case due to the heavy-tailed distribution of ε. As a result,

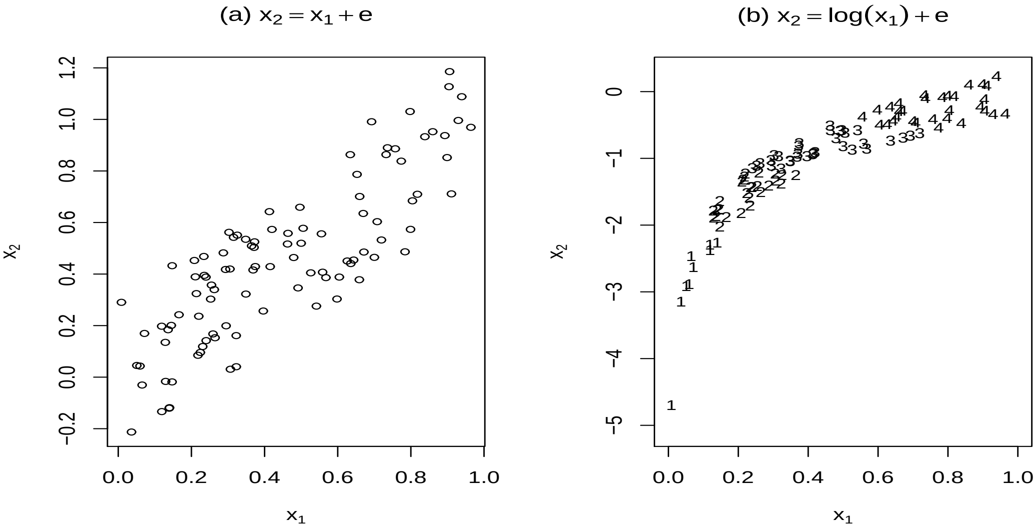

Figure 3.2: Scatterplot of X_2 versus X_1 with $e \sim$ Uniform$(-0.3, 0.3)$. The labels for the points in panel (b) denote the clustering result from k-means with $k = 4$.

OLS for dimension reduction does not work as well as the normal error case, while LAD keeps up the good performance. This is as expected, and we have demonstrated that LAD inherits its advantage over OLS in linear regression.

Now we explore the limitation of LAD for dimension reduction. Keep all the other settings the same except that we now set $X_2 = \log(X_1) + e$ with $e \sim$ Uniform$(-0.3, 0.3)$. Consider the case as in panel (b) of Figure 3.1, where ε has the Cauchy distribution. Without ambiguity, we will refer to this setting as the modified setting. In a typical run of this modified setting, we get $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (-0.768, 0.639)^T$ and $\hat{\boldsymbol{\beta}}_{\text{LAD}} = (0.693, 0.721)^T$. We see that LAD no longer performs well. As discussed in Section 3.2, the LCM assumption (1.3.1) plays a key role in recovering $\boldsymbol{\beta}$. From panel (a) of Figure 3.2, we see that the LCM assumption holds when $X_2 = X_1 + e$. On the other hand, $E(X_2|X_1)$ is nonlinear in panel (b) of Figure 3.2 and the LCM assumption does not hold when $X_2 = \log(X_1) + e$. From the proof of Proposition 3.2.1, it is not a surprise that $\hat{\boldsymbol{\beta}}_{\text{LAD}}$ can no longer recover $\boldsymbol{\beta}$.

3.3 Cluster-based LAD for dimension reduction

3.3.1 A sample level algorithm

Motivated from the discussions in Section 3.2.2, we propose cluster-based LAD for dimension reduction. Clustering provides a natural solution to the violation of the LCM assumption (1.3.1). We implement k-means clustering (Hartigan, 1975) with $k = 4$, and present the clustering result in panel (b) of Figure 3.2. Here all the points with the same label (1 through 4) belong to the same cluster. We choose k-means as it is one the most popular clustering algorithms and it is easy to implement.

Although $E(X_2|X_1)$ is nonlinear over the entire support of X_1, we see that $E(X_2|X_1)$ is approximately linear within each cluster. Recall that when the response Y has Cauchy error, the normalized LAD estimator in the modified setting is $\hat{\boldsymbol{\beta}}_{\text{LAD}} = (0.693, 0.721)^T$ without clustering. Denote $\hat{\boldsymbol{\beta}}_{\text{LAD},(c)}$ as the normalized LAD estimator of $\boldsymbol{\beta}$ within each cluster for $c = 1, 2, 3, 4$. It turns out that $\hat{\boldsymbol{\beta}}_{\text{LAD},(1)} = (0.995, 0.099)^T$, $\hat{\boldsymbol{\beta}}_{\text{LAD},(2)} = (0.944, 0.331)^T$, $\hat{\boldsymbol{\beta}}_{\text{LAD},(3)} = (0.999, 0.016)^T$, and $\hat{\boldsymbol{\beta}}_{\text{LAD},(4)} = (0.982, 0.188)^T$. Obviously, the LAD estimator within each cluster is more accurate than the original LAD estimator. After synthesizing across different clusters, we denote the final cluster-based LAD through k-means as $\hat{\boldsymbol{\beta}}_{[k]-\text{LAD}}$, which turns out to be very accurate as $\hat{\boldsymbol{\beta}}_{[4]-\text{LAD}} = (0.999, 0.033)^T$.

Now we present the step-by-step cluster-based LAD algorithm, and we will see exactly how to synthesize estimators across different clusters. Let $\{(\boldsymbol{X}_i, Y_i) : i = 1, 2, \ldots, n\}$ be a random sample of $(\boldsymbol{X}, Y)$ generated from the single-index model (3.1).

1. Perform k-means clustering according to $\{\boldsymbol{X}_i : i = 1, 2, \ldots, n\}$. After clustering, denote the observations in the cth cluster as $\{(\boldsymbol{X}_{i,(c)}, Y_{i,(c)}) : i = 1, 2, \ldots, n_c\}$ for $c = 1, 2, \ldots, k$. Here n_c is the sample size of the cth cluster and $\sum_{c=1}^{k} n_c = n$.

2. For $c = 1, 2, \ldots, k$, solve minimization problem
$$\arg\min n_c^{-1} \sum_{i=1}^{n_c} \left| Y_{i,(c)} - a_{(c)} - \boldsymbol{b}_{(c)}^T \boldsymbol{X}_{i,(c)} \right| \text{ over } a_{(c)} \in \mathbb{R} \text{ and } \boldsymbol{b}_{(c)} \in \mathbb{R}^p.$$
Denote the minimizer of $\boldsymbol{b}_{(c)}$ as $\hat{\boldsymbol{\beta}}^*_{\text{LAD},(c)}$.

3. Let $\hat{\ell}^*_{(c)} = \sqrt{(\hat{\boldsymbol{\beta}}^*_{\text{LAD},(c)})^T \hat{\boldsymbol{\beta}}^*_{\text{LAD},(c)}}$ be the length of $\hat{\boldsymbol{\beta}}^*_{\text{LAD},(c)}$. Calculate the normalized within cluster estimator as $\hat{\boldsymbol{\beta}}_{\text{LAD},(c)} = \hat{\boldsymbol{\beta}}^*_{\text{LAD},(c)} / \hat{\ell}^*_{(c)}$. Denote

$\hat{\boldsymbol{B}} \in \mathbb{R}^{p\times k}$ as $\hat{\boldsymbol{B}} = \{\hat{\boldsymbol{\beta}}_{\text{LAD},(1)}, \ldots, \hat{\boldsymbol{\beta}}_{\text{LAD},(k)}\}$, where the cth column of $\hat{\boldsymbol{B}}$ is $\hat{\boldsymbol{\beta}}_{\text{LAD},(c)}$.

4. Denote diagonal matrix $\hat{\boldsymbol{D}} \in \mathbb{R}^{k\times k}$ as $\hat{\boldsymbol{D}} = \text{diag}(n_1/n, \ldots, n_k/n)$, where the cth diagonal element of $\hat{\boldsymbol{D}}$ is n_c/n. Calculate $\hat{\boldsymbol{\Omega}} = \hat{\boldsymbol{B}}\hat{\boldsymbol{D}}\hat{\boldsymbol{B}}^T$.

5. Perform eigen-value decomposition of $\hat{\boldsymbol{\Omega}}$. The eigenvector corresponding to the largest eigenvalue of $\hat{\boldsymbol{\Omega}}$ is the final cluster-based LAD estimator, and is denoted as $\hat{\boldsymbol{\beta}}_{[k]-\text{LAD}}$.

The algorithm above synthesizes the within cluster estimators through eigen-value decomposition. Since it is the direction, not the scale, of the LAD estimator $\hat{\boldsymbol{\beta}}^*_{\text{LAD},(c)}$ in Step 2 that we want to synthesize, the normalized estimator $\hat{\boldsymbol{\beta}}_{\text{LAD},(c)}$ is calculated in Step 3. Step 4 is a reweighting step where the weight of each within cluster estimator is proportional to the correponding cluster size.

3.3.2 Population level justification

We provide the justification for the sample level cluster-based LAD algorithm in this section. Suppose $(\boldsymbol{X}, Y)$ are generated from model (3.1). Let $\{H_1, H_2, \ldots, H_k\}$ be a partition of the support of $\boldsymbol{X}$. For $c = 1, \ldots, k$, define $\boldsymbol{X}_{(c)} = \boldsymbol{X}I(\boldsymbol{X} \in H_c)$ and $Y_{(c)} = YI(\boldsymbol{X} \in H_c)$, where $I(\boldsymbol{X} \in H_c)$ is the indicator function of $\boldsymbol{X}$ belonging to H_c. Let $D = \text{diag}(f_1, \ldots, f_k)$, where

$f_c = E\{I(X \in H_c)\}$ is the probability of $\boldsymbol{X}$ belonging to H_c. For $c = 1, \ldots, k$, consider optimization problem

$$\arg\min E\{|Y_{(c)} - a_{(c)} - \boldsymbol{b}_{(c)}^T \boldsymbol{X}_{(c)}|\} \text{ over } a_{(c)} \in \mathbb{R} \text{ and } \boldsymbol{b}_{(c)} \in \mathbb{R}^p.$$

Suppose the minimizer of $\boldsymbol{b}_{(c)}$ is unique and denote it as $\boldsymbol{\beta}^*_{\mathrm{LAD},(c)}$. Normalize $\boldsymbol{\beta}^*_{\mathrm{LAD},(c)}$ to get $\boldsymbol{\beta}_{\mathrm{LAD},(c)} = \boldsymbol{\beta}^*_{\mathrm{LAD},(c)}/\ell^*_{(c)}$, where $\ell^*_{(c)}$ is the length of $\boldsymbol{\beta}^*_{\mathrm{LAD},(c)}$. Denote $\boldsymbol{B} = \{\boldsymbol{\beta}_{\mathrm{LAD},(1)}, \ldots, \boldsymbol{\beta}_{\mathrm{LAD},(k)}\}$ and $\boldsymbol{\Omega} = \boldsymbol{BDB}^T$. Let $\boldsymbol{\beta}_{[k]-\mathrm{LAD}}$ be the eigenvector corresponding to the nonzero eigenvalue of $\boldsymbol{\Omega}$. We have

Proposition 3.3.1. *Suppose that $E(\boldsymbol{X}_{(c)}|\boldsymbol{\beta}^T\boldsymbol{X}_{(c)})$ is a linear function of $\boldsymbol{\beta}^T\boldsymbol{X}_{(c)}$ for $c = 1, \ldots, k$. Then $\boldsymbol{\beta}_{[k]-LAD}$ is proportional to $\boldsymbol{\beta}$ in (3.1).*

PROOF. From Proposition 3.3.1, we know that $\boldsymbol{\beta}^*_{\mathrm{LAD},(c)}$ is proportional to $\boldsymbol{\beta}$. Thus $\boldsymbol{\beta}_{\mathrm{LAD},(c)}$ is proportional to $\boldsymbol{\beta}$. It follows that the column space of $\boldsymbol{B}$ and the column space of $\boldsymbol{\Omega}$ are both spanned by $\boldsymbol{\beta}$. As the eigenvector corresponding to the only nonzero eigenvalue of $\boldsymbol{\Omega}$, $\boldsymbol{\beta}_{[k]-\mathrm{LAD}}$ is proportional to $\boldsymbol{\beta}$. □

We refer to the condition in Proposition 3.3.1 as the local LCM assumption, which is the key assumption to guarantee the proportionality between $\boldsymbol{\beta}^*_{\mathrm{LAD},(c)}$ and $\boldsymbol{\beta}$. The normalization step and the reweighting step of the algorithm are not essential to the proof of Proposition 3.3.1, and different weighting schemes can be used in applications. Generally speaking, let $w_1, \ldots, w_k$ be positive constants and define $\boldsymbol{\Omega}_w = \sum_{c=1}^k w_c \boldsymbol{\beta}^*_{\mathrm{LAD},(c)}(\boldsymbol{\beta}^*_{\mathrm{LAD},(c)})^T$. Then we can recover

$\boldsymbol{\beta}$ through eigenvalue decomposition of $\boldsymbol{\Omega}_w$.

3.3.3 Determination of cluster number k

An important issue in application is to determine the number of clusters in the k-means algorithm. We propose to implement a data-driven procedure based on local linear median regression, which is a special case of the local quantile linear regression in Yu and Jones (1998). Compared with the local linear least squares regression, local linear median regression is less sensitive to outliers in the response and potential heavy-tailed error distribution.

For random sample $\{(\boldsymbol{X}_i, Y_i) : i = 1, 2, \ldots, n\}$ and candidate cluster number k, $\hat{\boldsymbol{\beta}}_{[k]-\mathrm{LAD}}$ is the cluster-based LAD estimator. Let $\hat{z}_i^{[k]} = \hat{\boldsymbol{\beta}}_{[k]-\mathrm{LAD}}^T \boldsymbol{X}_i$ and we perform local smoothing based on $\{(\hat{z}_i^{[k]}, Y_i) : i = 1, 2, \ldots, n\}$. Specifically, for fixed j, consider

$$\arg\min \sum_{i=1, i\neq j}^{n} \left| Y_i - a_j - b_j \left(\hat{z}_i^{[k]} - \hat{z}_j^{[k]} \right) \right| \phi_h \left(\hat{z}_i^{[k]} - \hat{z}_j^{[k]} \right) \text{ over } a_j, b_j \in \mathbb{R}.$$

Here $\phi_h(\cdot) = \phi(\cdot/h)/h$ with $\phi(\cdot)$ being the standard normal density, and h is the tuning parameter commonly known as the window size in the local smoothing literature. Denote the minimizer of a_j as $\hat{Y}_j^{[k]}$. Then $\hat{Y}_j^{[k]}$ is the jth estimated response and $\hat{\varepsilon}_j^{[k]} = Y_j - \hat{Y}_j^{[k]}$ is the the jth residual through local linear median regression. Repeat this procedure for $j = 1, \ldots, n$. The median

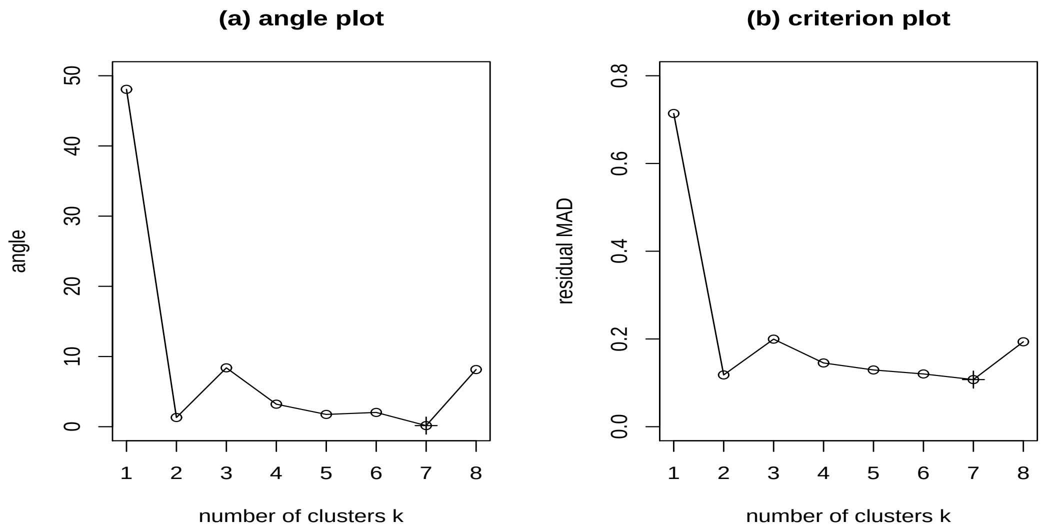

Figure 3.3: Angle plot (panel (a)) and criterion plot (panel (b)) with $\hat{\sigma}_{[k]}$ for Example 1. The "+" sign in each plot denotes the respective minimum.

absolution deviation (MAD) of $\{\hat{\varepsilon}_j^{[k]} : j = 1, \ldots, n\}$ is then calculated as

$$\hat{\sigma}_{[k]} = \text{Median}\left(\left|\hat{\varepsilon}_j^{[k]} - \text{Median}\left(\hat{\varepsilon}_j^{[k]}\right)\right|\right). \tag{3.3}$$

The residual MAD is a function of cluster number k, and we can choose the optimal cluster number such that $\hat{\sigma}_{[k]}$ is minimized. Since $\hat{\sigma}_{[k]}$ also depends on the choice of window size h, we may try a set of different h values and choose the one with the minimum MAD.

3.4 Numerical studies

In this section, we carry out numerical studies to evaluate the performances of LAD and cluster-based LAD for dimension reduction. To evaluate the performance of $\hat{\boldsymbol{\beta}}$ as an estimator of the true direction $\boldsymbol{\beta}$, we report the acute angle between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$.

Example 1. This example focuses on determining the number of clusters through residual MAD in (3.3). Let $\boldsymbol{X} = (X_1, X_2)^T$ with X_1 and X_2 i.i.d. $N(0,1)$, $Y = X_1^2 + \varepsilon$, where ε follows a Cauchy distribution with location parameter 0 and scale parameter 0.1. We implement cluster-based LAD with cluster numbers $k = 1, 2, \ldots, 8$, where $k = 1$ corresponds to the regular LAD without clustering. The results based on one typical run with sample size $n = 100$ are summarized in Figure 3.3. In panel (a), we plot the angle between $\hat{\boldsymbol{\beta}}_{[k]-\mathrm{LAD}}$ and the true direction $\boldsymbol{\beta} = (1,0)^T$ versus the number of clusters k. In panel (b) of Figure 3.3, we plot $\hat{\sigma}_{[k]}$, the residual MAD in (3.3), versus k.

We clearly see that the two plots have very similar patterns. A larger residual MAD value means a poorer fit between the response Y_i and $\hat{\boldsymbol{\beta}}^T_{[k]-\mathrm{LAD}}\boldsymbol{X}_i$, which corresponds to a larger angle between $\hat{\boldsymbol{\beta}}_{[k]-\mathrm{LAD}}$ and the true $\boldsymbol{\beta}$. On the other hand, a smaller residual MAD value corresponds to a better fit, and a smaller angle consequently. In this example, cluster-based LAD with $k > 1$ can significantly outperform the classical LAD with $k = 1$. The minimum value in both plots are achieved with cluster number $k = 7$. Compared to the 48.06° angle from the LAD estimator, cluster LAD with $k = 7$ leads to the 0.16° angle between $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}_{[k]-\mathrm{LAD}}$. In Section 3.2.2, we have seen that cluster-based LAD can outperform LAD when the global LCM assumption

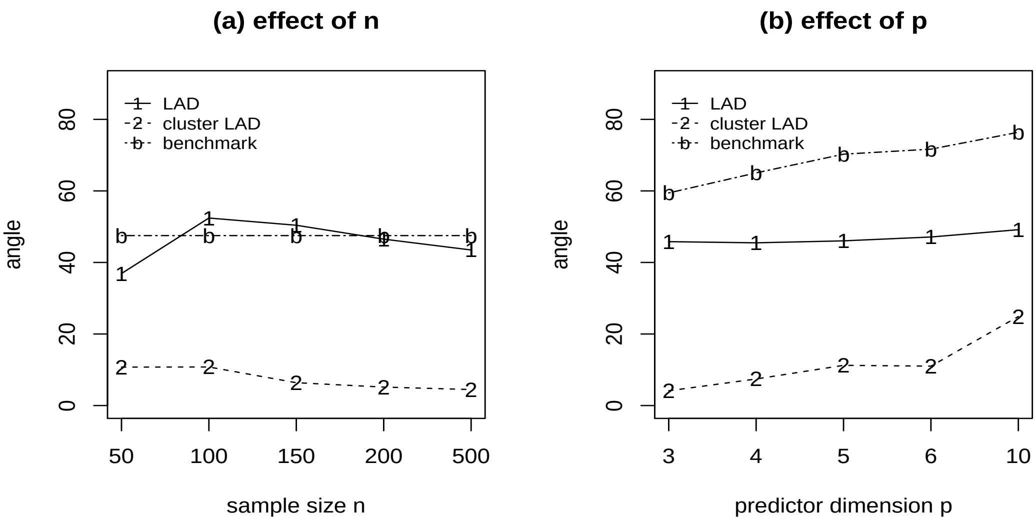

Figure 3.4: Effect of n (panel (a)) and effect of p (panel (b)) for Example 2.

is violated. In this example, $\boldsymbol{X}$ is normal and the global LCM assumption is satisfied. Thus we provide another instance when cluster-based LAD is preferred over LAD.

Example 2. This example examines the effects of sample size n and predictor dimension p on estimating $\boldsymbol{\beta}$. Let $X_1, X_3, X_4, \ldots, X_p$ be i.i.d. Uniform$(0,3)$, $X_2 = \sin(X_1) + e$ with $e \sim$ Uniform$(-0.3, 0.3)$, and $\boldsymbol{X} = (X_1, \ldots, X_p)^T$. Let $Y = \sin(X_1 - X_2) + .2\varepsilon$, where $\varepsilon \sim t_2$ has t distribution with 2 degrees of freedom. Then the normalized direction is $\boldsymbol{\beta} = (1, -1, 0, \ldots, 0)/\sqrt{2}$. Based on 100 repetitions, we report the median of the angles between $\boldsymbol{\beta}$ and its estimators in Figure 3.4. Suppose U follows a uniform distribution on the surface of of $\mathbb{R}^p$ dimensional sphere. Let η be a random draw from U, which can be seen as a random guess of $\boldsymbol{\beta}$. The benchmark is then set as the median

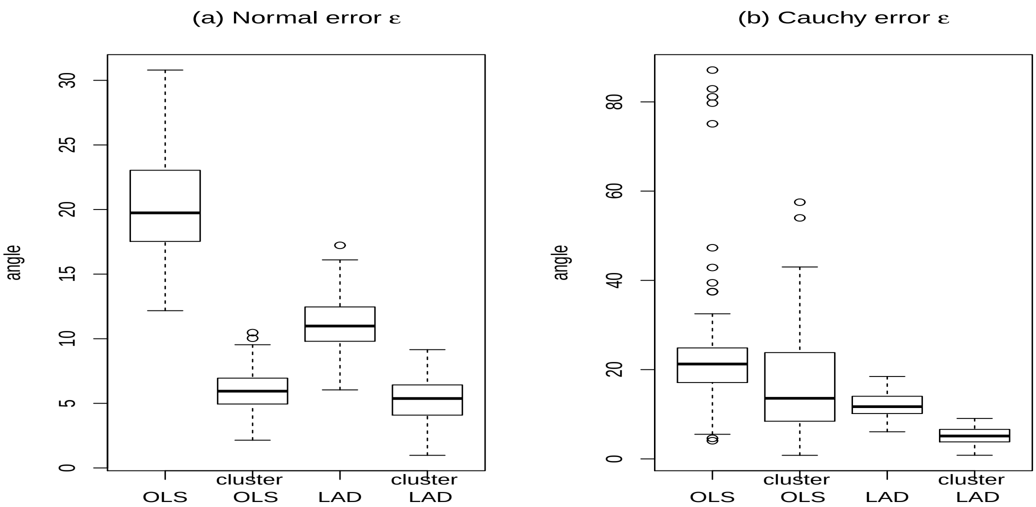

Figure 3.5: Boxplots for OLS, cluster-based OLS, LAD, and cluster-based LAD for Example 3. Normal error and Cauchy error of ε are plotted in panel (a) and panel (b) respectively.

angle between η and $\boldsymbol{\beta}$ based on 1000 random guesses. The benchmark value is a function of p, and does not change as n varies.

In panel (a) of Figure 3.4, we fix $p = 2$ and consider $n = 50, 100, 150, 200, 500$. In panel (b), we fix $n = 500$ and consider $p = 3, 4, 5, 6, 10$. We see that cluster-based LAD is consistently better than LAD. While cluster-based LAD improves as n increases and deteriorates as p increases, LAD seems to be consistently bad across all settings. LAD is not much better than the benchmark, or the median angle of random guesses, in panel (b), and LAD is not significantly different from the benchmark in panel (a).

Example 3. Let X_1, X_2 and X_3 be i.i.d. Uniform$(0, 1)$, $X_4 = -\log(X_1) + e$ with

$e \sim \text{Uniform}(-0.3, 0.3)$, $\boldsymbol{X} = (X_1, X_2, X_3, X_4)^T$, and $Y = \exp\{0.5(0.6X_1 + 0.8X_4) + 1\} + 0.1\varepsilon$. Then we have $\boldsymbol{\beta} = (0.6, 0, 0, 0.8)^T$. Consider two types of error distribution, where ε is either $N(0, 1)$ or a Cauchy distribution with location parameter 0 and scale parameter 0.1. Fix sample size $n = 200$, and we get the estimator $\hat{\boldsymbol{\beta}}$ based on OLS, cluster-based OLS, LAD, and cluster-based LAD. The boxplots for the angle between $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ based on 100 repetitions are summarized in Figure 3.5. In panel (a) with normal error, we see that cluster-based OLS and cluster-based LAD have similar performances, and they improve over OLS and LAD. In panel (b) with Cauchy error, cluster-based OLS is outperformed by LAD and cluster-based LAD, with cluster-based LAD being the best. Recall from Section 3.2.2 that the global LCM assumption does not hold in this setting. As a result, the cluster-based methods are better than their counterparts without clustering in both panels.

Example 4. Consider the New Zealand horse mussel data with $n = 82$ observations, which has been studied in Cook (1998) and is available from the R package "dr". The response Y is the muscle mass in grams, the edible portion of the mussel. We consider predictor $\boldsymbol{X} = (X_1, X_2)^T$, where X_1 is the shell length in millimeters and X_2 is the shell mass in grams. The shell length and shell mass are highly correlated with sample Pearson correlation 0.899. Furthermore, the scatterplot between the predictors (not reported here) demonstrates some nonlinear relationship. The OLS estimator is $\hat{\boldsymbol{\beta}}_{\text{OLS}} =$

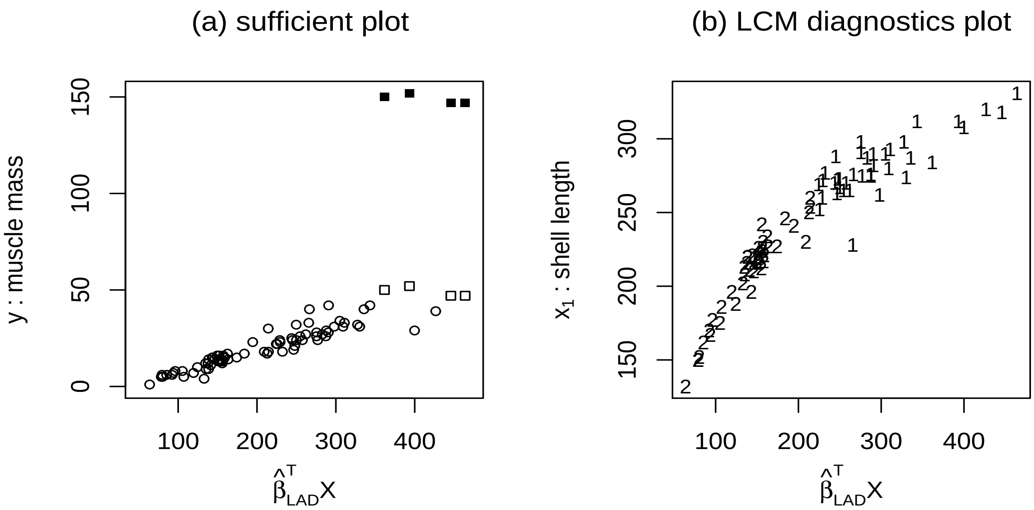

Figure 3.6: New Zealand horse mussel data in Example 4. In panel (a), the hollow circles and the hollow squares denote the original data, and the hollow squares are replaced by the solid squares for the contaminated data. In panel (b), the labels for the points denote the clustering result from k-means with $k = 2$.

$(0.463, 0.886)^T$, and the LAD estimator is $\hat{\boldsymbol{\beta}}_{\text{LAD}} = (0.415, 0.910)^T$. We refer to $\hat{\boldsymbol{\beta}}_{\text{LAD}}^T \boldsymbol{X}$ as the oracle predictor, which is essentially the same as $\hat{\boldsymbol{\beta}}_{\text{OLS}}^T \boldsymbol{X}$ because their sample Pearson correlation is 0.999. From the uncontaminated data in panel (a) of Figure 3.6, we observe a strong linear relationship between the response and the oracle predictor, which confirms the validity of the oracle predictor.

Next we argue that LAD is more robust to OLS in the presence of data contamination. Suppose the largest four response values (represented by the hollow squares in panel (a)) are mistakenly recorded due to data entry errors, such that the recorded values are their respective original values plus 100 (represented by the solid squares in panel (a)). Denote the OLS estimator and the LAD estimator based on this contaminated data as $\hat{\boldsymbol{\beta}}_{\text{OLS}}^*$ and $\hat{\boldsymbol{\beta}}_{\text{LAD}}^*$. Not surprisingly, the sample Pearson correlation between $(\hat{\boldsymbol{\beta}}_{\text{OLS}}^*)^T \boldsymbol{X}$ and the oracle predictor drops to 0.949, while the sample Pearson correlation between $(\hat{\boldsymbol{\beta}}_{\text{LAD}}^*)^T \boldsymbol{X}$ and the oracle predictor is 0.999.

Now we consider the cluster-based LAD estimator for the contaminated data. From panel (b) of Figure 3.6, $E(X_1 | \hat{\boldsymbol{\beta}}_{\text{LAD}}^T \boldsymbol{X})$ seems to have some nonlinear trend, and the global LCM condition is violated. The k-means clustering result with $k = 2$, on the other hand, suggests that the local LCM condition is more suitable than the global LCM condition. Denote $\hat{\boldsymbol{\beta}}_{[2]-\text{LAD}}^*$ as the corresponding cluster-based LAD estimator, and the sample Pearson correlation between $(\hat{\boldsymbol{\beta}}^*_{[2]-\mathrm{LAD}})^T\boldsymbol{X}$ and $(\hat{\boldsymbol{\beta}}^*_{\mathrm{LAD}})^T\boldsymbol{X}$ turns out to be 0.999. We see that when there is a strong linear relationship between the uncontaminated response and the predictors, clustering becomes unnecessary even when the global LCM condition is violated. This is as expected, since LAD does not require the global LCM condition in linear models.

CHAPTER 4

CONCLUSION AND FUTURE WORK

4.1 Conclusion and Summary

Testing predictor contribution has been an important topic in sufficient dimension reduction since the seminal work of Cook (2004). Existing marginal coordinate test procedures for SIR, SAVE and DR all rely on approximating the asymptotic distribution of the test statistics through sum of weighted χ^2 distributions. Under the normality assumption of the predictor distribution, we demonstrate the effectiveness of permutation test for predictor contribution in Chapter 2 through both theory and numerical examples. Our proposed procedures complement the existing permutation test procedure in the sufficient dimension reduction literature, which is a popular tool for determining the dimensionality of central space. The research from this chapter has been published on the Journal of Multivariate Analysis.

In Chapter 3, LAD and cluster-based LAD are proposed for dimension reduction in single-index models. Compared with the popular dimension reduction method OLS, our proposal is less sensitive to heavy-tailed error distributions. Clustering is used to facilitate dimension reduction when there is nonlinearity among the predictors. The research from this chapter has been published on the Journal of Statistical Theory and Practice.

4.2 Future Work

As discussed in Chapter 3, cluster-based regression has been proposed for dimension reduction. This method is helping with identifying relationship among predictors, but the marginal coordinate test is not yet developed. Because of that, we are not able to identify whether a predictor is contributed to response through hypothesis test. In Chapter 2, the proposed permutation approach for testing predictor contribution is a general framework, and has the potential to be extended to moment-based estimators other than SIR, SAVE and DR.

Derived from the above discussion, the synthesis of cluster-based regression and permutation procedure on hypothesis test provide a potential solution for testing predictors contribution in cluster-based regression. To be more specific, assuming $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_H$ are the kernel matrices from each proportion of data (e.g. cluster), then it is worth trying if permutation approach can be used to test predictors contribution through the following test statistic:

$$T(e_i) = \sum_{h=1}^{H} \frac{n_h}{n} (e_i^T \boldsymbol{M}_h e_i).$$

Also, because LAD regression is a special case of quantile regression, it will also be interesting to further investigate dimension reduction in quantile regression models without estimating the unknown link function.

Last but not least, the permutation approach can be extended from testing vector-valued predictor $\boldsymbol{X} \in \mathbb{R}^p$ to matrix-valued predictor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p_1 \times p_2}$. To test matrix-valued predictors contribution, we can consider three types of hypotheses: test for significant columns, test for significant rows, and test for significant sub-matrices. This framework is currently under investigation.

BIBLIOGRAPHY

[1] Brillinger, D. R. (1983). *A generalized linear model with "Gaussian" regressor variables. In A festschrift for Erich L. Lehmann (P. J. Bickel, K. A. Doksum, and J. L. Hodges, eds.)* Woodsworth International Group, Belmont, CA, 97–114.

[2] Cook, R. D. (1998). *Regression graphics: Ideas for studying regressions through graphics.* New York: Wiley.

[3] Cook, R. D. (2004). *Testing predictor contributions in sufficient dimension reduction.* The Annals of Statistics, **32**, 1062–1092.

[4] Cook, R.D., Li, B. (2002). *Dimension reduction for the conditional mean in regression.* The Annals of Statistics, **30**, 455–474.

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

[6] Cook, R. D. and Weisberg, S. (1994). *An Introduction to regression graphics*, New York: Wiley.

[7] Cook, R. D. and Yin, X. (2001). *Dimension reduction and visualization in discriminant analysis (with discussion).* Australian & New Zealand Journal of Statistics , **43**, 147–199.

[8] Ding, S. and Cook, R. D. (2014). *Dimension folding PCA and PFC for matrix-valued predictors.* Statistica Sinica , **24**, 463–492.

[9] Ding, S. and Cook, R. D. (2015). *Tensor sliced inverse regression.* Journal of Mltivariate Analysis , **133**, 216–231.

[10] Dong, Y. and Yang, C. (2016). *Cluster-based least absolute deviation regression for dimension reduction.* Journal of Statistical Theory and Practice, **1**, 121–132.

[11] Dong, Y., Yang, C. and Yu, Z. (2016). *On permutation tests for predictor contribution in sufficient dimension reduction.* Journal of Multivariate Analysis, **149**, 81–91.

[12] Härdle, W., Hall, P. and Ichimura, H. (1993). *Optimal smoothing in single-index models.* The Annals of Statistics, **21**, 157–158.

[13] Hartigan, J. (1975). *Clustering algorithms.* Wiley, New York.

[14] Ichimura, H. (1993). *Semiparametric least squares (SLS) and weighted SLS estimation of single-index models.* Journal of Econometrics. **58**, 71–120.

[15] Koenker, R. and Bassett, G. (1978). *Regression quantiles.* Econometrica, **46**, 33–50.

[16] Li, B. and Wang, S. (2007). *On directional regression for dimension reduction.* Journal of the American Statistical Association, **102**, 997–1008.

[17] Li, K. C. (1991). *Sliced inverse regression for dimension reduction (with discussion).* Journal of the American Statistical Association, **86**, 316–342.

[18] Li, K. C. and Duan N. (1989). *Regression analysis under link violation.* The Annals of Statistics, **17**, 1009–1052.

[19] Li, L., Cook, R. D., and Nachtsheim C. (2004). *Cluster-based estimation for sufficient dimension reduction.* Computational Statistics & Data Analysis, **47**, 175–193.

[20] Mai, Q. and Zou H. (2015). *Nonparametric variable transformation in sufficient dimension reduction.* Technometrics, **57**, 1–10.

[21] Narula, S. C. and Wellington, J. F. (1982). *The minimum sum of absolute errors regression: A state-of-the-art survey.* International Statistical Review, **50**, 317–326.

[22] Portnoy, S. and Koenker, R. (1997). *The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators.* Statistical Science, **12**, 279–300.

[23] Shao, Y, Cook, R. D. and Weisberg, S. (2007). *Mariginal tests with sliced average variance estimation.* Biometrika, **94**, 285–296.

[24] Wang, T., Guo, X. and Zhu, L. X. (2014). *Transformed sufficient dimension reduction.* Biometrika, **101**, 815–829.

[25] Wu, T., Yu, K. and Yu, Y. (2010). *Single-index quantile regression.* Journal of Multivariate Analysis, **101**, 1607–1621.

[26] Yeo, I.-K. and Johnson, R. A. (2000). *A new family of power transformations to improve normality or symmetry.* Biometrika, **87**, 954–959.

[27] Yu, K. and Jones, M. C. (1998). *Local linear quantile regression.* Journal of the American Statistical Association, **93**, 228–237.

[28] Yu, Z. and Dong, Y. (2015). *Model-free coordinate test and variable selection via directional regression.* Statistica Sinica, Accepted.