Stable and Pseudo Sufficient Dimension Reduction

by

Wenbo Wu

(Under the direction of Professor Xiangrong Yin)

Abstract

Due to recent advances of modern technology, abundant data are collected in many scientific areas. The developments in theory and methodology for *sufficient dimension reduction* (SDR) have provided a powerful tool to study such high dimensional data. Most existing methods are aiming at estimating the basis matrix and structural dimension of the *central subspace* (CS). This dissertation is composed with three parts. In the first study, we introduce stable estimation procedures for several aspects of a sufficient dimension reduction matrix, including a stable method for estimating structural dimension, a Grassmann Manifold sparse estimate for the CS, a stable nonsparse estimate for the CS. In the second study, in order to obtain a reliable estimate for correlated predictors, we uncover the underlying relationship between ridge regression and measurement error regression. With such a connection, we propose a general SDR estimation procedure to obtain an estimate from a different subspace instead of the targeted population parameter space. In the third study, we combine the stable and pseudo approach together and tackle the small n large p problem for dimension reduction. Theoretical results are established for our methods and the efficacy of the proposed methods is demonstrated by simulation studies and real data analyses.

Index words: Grassmann manifold, measurement regression, penalized estimator, ridge regression, subsampling, sufficient dimension reduction.

Stable and Pseudo Sufficient Dimension Reduction

by

Wenbo Wu

B.S., University of Texas at Austin, 2007

M.S., University of Central Florida, 2010

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2015

Stable and Pseudo Sufficient Dimension Reduction

by

Wenbo Wu

Approved:

Major Professor:	Xiangrong Yin
Committee:	William McCormick Jaxk Reeves John Stufken Lily Wang

Electronic Version Approved:

Julie Coffield
Interim Dean of the Graduate School
The University of Georgia
May 2015

Stable and Pseudo Sufficient Dimension Reduction

Wenbo Wu

April 23, 2015

Dedication

To my dear parents, Minhao Wu and Boying Mu;

my beloved wife, Wenjing Yao;

and my precious daughter, Joanna Wu.

Acknowledgments

I would like to express my deepest gratitude to Professor Xiangrong Yin, my major professor, for his guidance, support and encouragement. Having him as my dissertation advisor was the best decision that I have made during my Ph.D study. His enthusiasm, optimism, and confidence have set a great example for me to be a successful scientific researcher. It is like the old Chinese saying "One day as a teacher, lifetime as a parent". He has been and will continue to be a great mentor for me in lifetime.

I would also like to thank Dr. Stufken, Dr. McCormick, Dr. Wang and Dr. Reeves for serving on my dissertation committee. Their thoughtful comments and encouragement have helped me to produce better research during my study. I would like to extend my gratitude to many faculty members of the Department of Statistics at the University of Georgia. I ought to thank the faculty committee who brought me into this wonderful department. Thanks to Dr. Sriram, Dr. McCormick and Dr. Love-Myers for writing recommendations in the past. I also want to thank the Statistical Consulting Center for giving me opportunities working as a graduate consultant.

Last but not least, I want to thank my family. Thank my parents, who always unconditionally love, trust and understand me. Thank my wife, Wenjing, for your love, support and encouragement. I am never alone with you by my side. Thank my newborn baby, Joanna, you are the angel that God brought to us. You make me brave and strong.

I always thank God for his mercy and love. We should give all the glory to God.

Contents

Acknowledgments **v**

List of Figures **viii**

List of Tables **ix**

1 Overview **1**

2 Stable sufficient dimension reduction **6**

2.1 Introduction 6

2.2 Stable estimation in dimension reduction 8

2.3 Simulations 21

2.4 Discussion 28

3 Pseudo sufficient dimension reduction and sufficient variable selection **48**

3.1 Introduction 48

3.2 Pseudo SDR and Pseudo SVS 54

3.3 Illustrative algorithms 61

3.4 Numerical study 63

3.5 Discussion 76

4 Extrinsic sufficient dimension reduction 79

4.1 Introduction 79

4.2 Extrinsic and intrinsic sampling 81

4.3 Penalized extrinsic sampling approaches 84

4.4 Data augmentation by ER 86

4.5 Numerical study 88

4.6 Discussion and future work 92

Bibliography 94

List of Figures

2.1 Accuracy of Stable SED for testing structural dimension 23

2.2 Sensitivity of sliced inverse methods to the choices of H 46

2.3 Stabilizing the choice of H for inverse methods 47

3.1 Box plots of optimal ridge tuning parameter using $\mathbf{X}$ 59

3.2 Estimation accuracy of linear models 65

3.3 Estimation accuracy of SIR models 68

3.4 Scatter plot of lpsa vs the 1^{st} SIR direction 74

3.5 Box plot of optimal ridge tuning parameter using $\tilde{\mathbf{X}}$ 77

List of Tables

2.1 Stable Grassmann manifold sparse estimation 25

2.2 Stable nonsparse estimates of the CS 26

2.3 Estimates for Hitters' data 27

2.4 Definitions of sets 30

2.5 Additional simulation results for Section 2.3.2.1 I 43

2.6 Additional simulation results for Section 2.3.2.1 II 45

3.1 Optimal ridge tuning parameters using existing methods 67

3.2 Improvement over existing ridge estimates 69

3.3 TPRs and FPRs of pseudo SVS for linear model 71

3.4 TPRs and FPRs of pseudo SVS for SIR model 72

3.5 Estimating structural dimension by pseudo SIR 73

3.6 PE for prostate cancer data and managerial role performance data 76

4.1 Stability selection with extrinsic sampling, subsampling, and bootstrap . . . 89

4.2 TPR and FPR of variable selection for linear model 91

4.3 TPR and FPR of variable selection for SIR model 92

Chapter 1

Overview

The goal of a regression analysis is to understand the conditional distribution of the univariate response Y given a $p \times 1$ predictor vector $\mathbf{X}$. Attention is often restricted to the mean function $\mathrm{E}(Y|\mathbf{X})$, and perhaps the variance function $\mathrm{Var}(Y|\mathbf{X})$. As the dimensionality of the data has been significantly increased in the recent decades, many traditional statistical tools fail due to the "curse of dimensionality" (Bellman, 1961). Hence, dimension reduction has become one of the most popular topics in the statistical fields. Dimension reduction is useful not only for computational efficiency, but also for improving the accuracy of the analyses.

Among many dimension reduction methods, principal component analysis (PCA) as a multivariate technique is probably the most well-known approach. It has been used in almost any area with large number of variables. PCA explains the covariance structure through a few linear combinations of the variables. It has interpretation purpose and should be considered as an intermediate step in much larger investigation. However, PCA as a dimension reduction method in the regression sense is a naive approach, or marginal dimension reduction approach because it reduces the dimension of $\mathbf{X}$ without considering any information from Y.

Sufficient dimension reduction (SDR) that also reduces the dimension of predictors through a few linear combinations of the predictors, but it incorporates the information from the response variable as well. The goal of SDR (Li, 1991; Cook, 1994, 1996) is to infer the dimension reduction subspace $S(\mathbf{B})$, which is spanned by the columns of a $p \times d$ matrix $\mathbf{B}$, such that: $Y \perp\!\!\!\perp \mathbf{X} | \mathbf{B}^T\mathbf{X}$, where $\perp\!\!\!\perp$ stands for independence. This expression states that for a given value of $\mathbf{B}^T\mathbf{X}$, the distribution of Y is independent of $\mathbf{X}$. When the intersection of all dimension reduction subspaces itself is a dimension-reduction subspace, it is called the *central subspace* (CS) and denoted by $S_{Y|\mathbf{X}}$ (Cook, 1996). The dimension d of the CS, which is the number of linearly independent columns of the basis matrix $\mathbf{B}$, is called the *structural dimension* of the CS (Cook, 1996). Conditions for the existence of the CS are developed by Cook (1998) and Yin et al. (2008). In this dissertation, we assume that the CS exists. Inferences in SDR usually contain two parts: estimating the structural dimension d, and estimating the basis matrix $\mathbf{B}$ of the CS. Typical estimators of the CS are usually obtained by using dimension reduction matrices. These methods include well-known sliced inverse regression (SIR; Li 1991), principal Hessian directions (PHD; Li 1992) and sliced inverse variance estimate (SAVE; Cook and Weisberg 1991). The structural dimension of the CS is typically estimated by the number of nonzero eigenvalues of the dimension reduction matrix, while a basis matrix of the CS is estimated by their corresponding eigenvectors. There are many existing methods to estimate d, including chi-square test (Li, 1991), modified BIC criterion (Zhu et al., 2006) and more recently, the sparse eigenvalue decomposition (SED) procedure (Zhu et al., 2010a). Cook (2004) reformulated a dimension reduction matrix as a least squares approach over a Grassmann Manifold. This reformulation makes many techniques in the least squares approach become available for the purpose of dimension reduction. For instance, Li (2007) developed sparse sufficient dimension reduction methods (SSDR) by reformulating equivalent least squares estimators for dimension reduction matrices and then adapted penalization approaches. The SED (Zhu et al., 2010a) method reformulates the eigen-decomposition method as a least squares approach, then adapts penalization approaches such as LASSO (Tibshirani, 1996) or adaptive LASSO (Zou, 2006).

Variable selection is another important problem in the literature of SDR. Similarly to that of Cook (2004), Yin and Hilafu (2015) formally defined central variable selection space, $S^V_{Y|\mathbf{X}}$, to be the column space of $\boldsymbol{\alpha}$ where the columns of $\boldsymbol{\alpha}$ consist of unit vectors of $\mathbf{e}_j$ with j^{th} element 1 and 0 otherwise, such that $Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\alpha}^T\mathbf{X}$. They discussed the differences between SDR and sufficient variable selection (SVS). One approach to achieve SVS is the use of penalized method with SDR procedures. Many existing methods (Ni et al., 2005; Li, 2007; Zhou and He, 2008; Wang and Yin, 2008; Chen et al., 2010) impose a penalization to the estimated $S_{Y|\mathbf{X}}$ in order to obtain $S^V_{Y|\mathbf{X}}$.

In this dissertation, we focus on developing methodologies in SDR and SVS. In particular, we propose stable and pseudo estimates for SDR and SVS. In Chapter 2, we introduce stable estimation procedures for several aspects of a sufficient dimension reduction matrix. We first propose a stable method for estimating structural dimension, which only selects the correct directions in the central subspace with no false positive selection. We then provide a Grassmann Manifold sparse estimate for the central subspace. By using subsampling, we develop an ensemble method to obtain a stable nonsparse estimate for the central subspace. This ensemble idea is also used to stabilize the choice of the number of slices in sliced inverse methods. Theoretical results are established, and the efficacy of the proposed stable methods is demonstrated by simulation studies and the analysis of Hitters' salary data.

In Chapter 3, we study the case of high correlations among predictors, which is a commonly encountered problem in many fields. For example, the prostate cancer data (Stamey et al., 1989) which examines the correlation between the level of prostate specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy is of this type. This phenomenon usually results in unstable estimates with large standard errors (Hoerl and Kennard, 1970). Among many existing methods, ridge regression (Hoerl and Kennard, 1970) is a popular approach to handle such a problem. However, selection criterion of a possible optimal ridge tuning parameter is always difficult and it is only available in linear models, hence, model-dependent. We propose a new concept of pseudo SDR. We uncover the underlying relationship between ridge regression and measurement error regression. With such a connection, we propose a general SDR estimation procedure to obtain an estimate from a different subspace instead of the targeted population parameter space. Using an ensemble idea, our proposed pseudo estimate is better than the traditional estimate or a ridge estimate for highly correlated predictors, and avoids the difficulties for choosing a particular ridge tuning parameter. In addition, we propose a variable selection procedure based on the a pseudo confidence interval. When the sample size is small, we propose to pool many extrinsic samples together to enlarge the sample size. Our approach is useful when the original sample size is small, especially for a small n large p problem. With an enlarged sample, a small n large p problem becomes to a classic large n small p problem. By using multiple extrinsic samples, a nonsparse estimate can be obtained through an ensemble idea over each extrinsic sample estimate and a sparse estimate can be obtained based on the empirical confidence interval of estimates from multiple extrinsic samples. Theoretical properties are studied for pseudo estimators. Our method requires no parametric assumptions on the underlying model. The effectiveness of the newly proposed methods are demonstrated by simulation studies and two real data analyses.

In Chapter 4, we study another challenging problem in statistics, the small n (sample size) large p (number of predictors) problem. We will tackle this problem by combining both stable and pseudo estimation methods. Meinshausen and Bühlmann (2010) uses subsampling in their method to achieve stable variable selection. However, when the original observed sample size is relatively small, traditional re-sampling methods, such as subsampling, reduce the effective sample size even further in the estimation procedure. In this study, we propose an extrinsic sampling approach to ensure sufficient effective sample size. We illustrate extrinsic sampling methods by using ordinary least squares and sliced inverse regression. Simulation results demonstrate the effectiveness of our methods.

The connections between our approach and Meinshausen and Bühlmann (2010) suggest that we shall be able to establish similar theory. On the other hand, the extrinsic sample also brings different aspects, and perhaps difficulties. Our future work will investigate and establish such theories and apply our algorithm in other areas with small n large p data.

Chapter 2

Stable sufficient dimension reduction[1]

2.1 Introduction

Dimension reduction that reduces the dimension of predictors is useful not only for computational efficiency, but also for improving the accuracy of analysis. Let Y be a univariate response and $\mathbf{X} = (x_1, x_2, ..., x_p)^T$ be a $p \times 1$ predictor vector. The goal of dimension reduction (Li, 1991; Cook, 1996, 1994) is to infer the dimension reduction subspace $S(\mathbf{B})$, which is spanned by the columns of a $p \times k$ matrix $\mathbf{B}$, such that: $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{B}^T\mathbf{X}$, where $\perp\!\!\!\perp$ stands for independence. This expression states that for a given value of $\mathbf{B}^T\mathbf{X}$, the distribution of Y is independent of $\mathbf{X}$. When the intersection of all dimension reduction subspaces itself is a dimension-reduction subspace, it is called the *central subspace* (CS) and denoted by $S_{Y|\mathbf{X}}$ (Cook, 1996). The dimension k of the CS, which is the number of linearly independent columns in $\mathbf{B}$, is called the *structural dimension* of the CS (Cook, 1996). Conditions for the existence of the CS are developed by Cook (1998) and Yin et al. (2008). In this article, we assume that the CS exists. Inferences in dimension reduction usually contain two parts: estimating the structural dimension k, and estimating $\mathbf{B}$, a basis matrix of the CS.

[1] Wu, W. and Yin, X. (2015). Stable estimation in dimension reduction. *Journal of Computational and Graphical Statistics*, 24(1):104-120. Reprinted here with permission of publisher.

Classical estimators of the CS are usually obtained by using dimension reduction matrices. These methods include well-known sliced inverse regression (SIR; Li 1991), principal Hessian directions (PHD; Li 1992) and sliced inverse variance estimate (SAVE; Cook and Weisberg 1991). The structural dimension of the CS is typically estimated by the number of nonzero eigenvalues of the dimension reduction matrix, while a basis matrix of the CS is estimated by their corresponding eigenvectors. There are many existing methods to estimate k, including chi-square test (Li, 1991), modified BIC criterion (Zhu et al., 2006) and more recently, the sparse eigenvalue decomposition (SED) procedure (Zhu et al., 2010a). Cook (2004) reformulated a dimension reduction matrix as least squares approach over a Grassmann Manifold. This reformulation makes many techniques in the least squares approach become available for the purpose of dimension reduction. For instance, Li (2007) developed sparse sufficient dimension reduction methods (SSDR) by reformulating equivalent least squares estimators for dimension reduction matrices and then adapted penalization approaches. The SED (Zhu et al., 2010a) method reformulates the eigen-decomposition method as a least squares approach, then adapts penalization approaches such as LASSO (Tibshirani, 1996) or adaptive LASSO (Zou, 2006). These penalized methods can achieve a quite accurate estimate, when the tuning parameter is chosen correctly. While methods like SED or SSDR have achieved great success, their results can be influenced greatly by the selection of tuning parameters in the penalization steps.

In this paper, we propose stable procedures for dimension reduction so that the influence of the selection of tuning parameters is reduced when the penalization approach is used. In addition, we propose an ensemble idea with a subsampling scheme to improve the accuracy of the estimates. The paper is organized as follows: In Section 2.2.1, we combine a subsampling method and random weights scheme to stabilize the estimate of the SED method; in Section 2.2.2, we propose a Grassmann Manifold sparse estimate, providing a new sparse sufficient dimension reduction estimate and combining it with a subsampling and random weights approach to achieve the stability of the estimate; in Section 2.2.3, we propose an ensemble approach to combine results from different subsamples with a threshold idea to obtain a stable non-sparse estimate for the CS. In addition, we aggregate dimension reduction matrices with different numbers of slices in the inverse methods to overcome the well-known issue for choosing the number of slices. In Section 2.3 we conduct simulation studies. We conclude our paper with a short discussion in Section 2.4. All proofs and additional materials are arranged in the supplementary file (Web Appendix A$-$H).

In this paper, for dimension reduction we will work with the standardized predictors $\mathbf{Z} = \mathbf{\Sigma}_{\mathbf{x}}^{-1/2}(\mathbf{X} - E(\mathbf{X}))$ such that $E(\mathbf{Z}) = 0$ and $\mathbf{\Sigma}_{\mathbf{z}} = \mathbf{I}_p$ where $\mathbf{\Sigma}_{\mathbf{x}}$ and $\mathbf{\Sigma}_{\mathbf{z}}$ are covariance matrices of $\mathbf{X}$ and $\mathbf{Z}$, respectively. This is because the CS found in the $\mathbf{Z}$-scale can be easily transformed back to the $\mathbf{X}$-scale (Cook, 1998). However, for variable selection in Section 2.2.2, our discussion will focus on the original $\mathbf{X}$-scale. This is because sparsity is not generally transformed from one scale to another scale. That is, if a model is sparse in the $\mathbf{X}$-scale, it does not mean it is sparse in the $\mathbf{Z}$-scale, or vice versa. We also assume that the sample covariance of $\mathbf{X}$ based on a random sample $(Y_i, \mathbf{X}_i)$ for $i = 1, \cdots, n$, is invertible. Remarks on a non-invertible sample covariance matrix of $\mathbf{X}$ can be found in Section 2.4.

2.2 Stable estimation in dimension reduction

To obtain a stable estimate, we mainly use three ideas: subsampling, random weights and ensemble. For subsampling, Theorem 1 of Meinshausen and Bühlmann (2010) established a general upper bound of the expected number of falsely selected variables by using subsampling for any procedure that is not worse than random guessing. This result holds for most of the existing dimension reduction methods that consistently estimate the CS. Thus we omit its result and theoretical discussion here, but only adopt the subsampling scheme in our algorithms for dimension reduction. However, for the use of random weights in dimension reduction, while the idea is similar, the technique and setup are different, and so are the results. Hence, we will establish respective theoretical results in Section 2.2.1 where the structural dimension is estimated, and in Section 2.2.2 where the sparse estimate of the CS is obtained. In Section 2.2.3, an ensemble idea is proposed to obtain a stable non-sparse estimate of the CS. Furthermore, we use the ensemble idea to overcome the issue of using a fixed number of slices in the inverse dimension reduction methods.

2.2.1 Estimating structural dimension

Suppose that a dimension reduction matrix $\mathbf{M}$ is constructed using the standardized predictors $\mathbf{Z}$, then a typical eigen-decomposition solution can be rewritten as

$$\mathbf{M} = \sum_{i=1}^{p} \tilde{\beta}_i \eta_i \eta_i', \tag{2.1}$$

where $\tilde{\beta}_i$ and η_i are the corresponding eigenvalues and eigenvectors of $\mathbf{M}$. The number of nonzero eigenvalues of $\tilde{\beta}_i$ is the estimated structural dimension $\hat{k}$, while their corresponding eigenvectors span the estimated CS. Having such an $\hat{\eta} = (\eta_1, \cdots, \eta_p)$, in order to obtain a more accurate $\hat{k}$, we adopt a penalized method, similar to that of Zhu et al. (2010a):

$$\hat{\tilde{\boldsymbol{\beta}}} = \arg\min_{\tilde{\boldsymbol{\beta}}} \left(\| \mathbf{M} - \sum_{i=1}^{p} \tilde{\beta}_i \eta_i \eta_i^T \|^2 + \lambda \sum_{i=1}^{p} \frac{|\tilde{\beta}_i|}{W_i} \right), \tag{2.2}$$

subject to $\hat{\eta}'\hat{\eta} = \mathbf{I}_p$, and $\mathbf{W} = (W_1, \cdots, W_p)'$ is a known weight vector based on the estimates of $\tilde{\beta}_i$ for $i = 1, \cdots, p$. The notation $\| \cdot \|$ refers to the Frobenius norm for a matrix. The L^1 penalization of adaptive LASSO (Zou, 2006) shrinks some eigenvalues exactly to zero, depending on the tuning parameter λ. The structural dimension is then estimated by the number of non-zero eigenvalues. Although the penalty term is the same, our formulation is slightly different from Zhu et al. (2010a), in which we treat $\hat{\eta}$ as fixed while they simultaneously estimated it. They showed that with a well-chosen tuning parameter λ, the SED procedure can provide a very accurate $\hat{k}$. They provided an information criterion to select the tuning parameter, but the choice of its range is critical. We want to develop a subsampling and random weights approach to stabilize the choice of the tuning parameter, as Meinshausen and Bühlmann (2010) did for the linear model. To do so, we first reformulate (2.1) as:

$$\tilde{Y} = \sum_{j=1}^{p} \tilde{\beta}_j x_j = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}, \tag{2.3}$$

where $\tilde{Y} = \text{vec}(\mathbf{M})$, $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \cdots, \tilde{\beta}_p)'$, and $\tilde{\mathbf{X}} = (\tilde{x}_1, \cdots, \tilde{x}_p) = (\eta_1 \otimes \eta_1, \cdots, \eta_p \otimes \eta_p)$. Model (2.3) can be viewed as a regression model with no error term. Note that $\tilde{\mathbf{X}}$ has a simple structure as $\tilde{\mathbf{X}}'\tilde{\mathbf{X}} = (\eta_1 \otimes \eta_1, \cdots, \eta_p \otimes \eta_p)'(\eta_1 \otimes \eta_1, \cdots, \eta_p \otimes \eta_p) = \mathbf{I}_p$ because η_i's are orthonormal eigenvectors. If A and B are the column indexes of $\tilde{\mathbf{X}}$, it is easy to verify that $\tilde{\mathbf{X}}'_A\tilde{\mathbf{X}}_B = \mathbf{0}$, if $A \cap B = \varnothing$.

Let $\mathcal{W}$ be a $p \times p$ diagonal matrix with diagonal entries $\mathcal{W}_{kk} = W_k$ for all $k = 1, \cdots, p$, where W_k are generated from Uniform$[u, 1]$ for some $u \in (0, 1)$. Meinshausen and Bühlmann (2010) suggested an effective range for $u \in (0.2, 0.8)$, and to sample W_k's independently as $W_k = u$ with probability $p_w \in (0, 1)$, and $W_k = 1$ otherwise. Let $\tilde{\mathbf{X}}^{\mathbf{w}} = \tilde{\mathbf{X}}\mathcal{W}$, then

$$\tilde{\mathbf{X}}^{\mathbf{w}} = (W_1\eta_1 \otimes \eta_1, \cdots, W_p\eta_p \otimes \eta_p) = \Big((\sqrt{W_1}\eta_1) \otimes (\sqrt{W_1}\eta_1), \cdots, (\sqrt{W_p}\eta_p) \otimes (\sqrt{W_p}\eta_p)\Big),$$

and model (2.3) becomes to

$$\tilde{Y} = \tilde{\mathbf{X}}^{\mathbf{w}}\boldsymbol{\beta}, \tag{2.4}$$

where $\mathcal{W}\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$. Moreover, we have

$$(\tilde{\mathbf{X}}^{\mathbf{w}})'\tilde{\mathbf{X}}^{\mathbf{w}} = \mathcal{W}'\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\mathcal{W} = \mathcal{W}'\mathcal{W} = \mathcal{D} = \text{diag}(W_i^2). \tag{2.5}$$

With the random weights, the optimization problem (2.2) is simplified as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\parallel \tilde{Y} - \tilde{\mathbf{X}}^{\mathbf{w}} \boldsymbol{\beta} \parallel^2 + \lambda \sum_{i=1}^{p} |\beta_i| \right), \tag{2.6}$$

which is a LASSO problem (Tibshirani, 1996). Hence, we can use the techniques from Meinshausen and Bühlmann (2010). Let $\mathcal{I}$ be a random subsample drawn from $\{1, ..., n\}$ without replacement, and $\hat{S}^{\lambda}(\mathcal{I}) = \{k : \hat{\beta}_k^{\lambda} \neq 0\}$ be the selected set by (2.6) with tuning parameter λ for the random subsample $\mathcal{I}$. We define the selection probability for a set $K \subseteq \{1, ..., p\}$ and selected tuning parameter λ to be

$$\hat{\Pi}_K^{\lambda} = P^* \{K \subseteq \hat{S}^{\lambda}(\mathcal{I})\}. \tag{2.7}$$

For a given cutoff probability π_{thr} and a tuning parameter range Λ_0, the set of stable predictors is

$$\hat{S}^{stable} = \{k : \max_{\lambda \in \Lambda_0} (\hat{\Pi}_k^{\lambda}) \geq \pi_{thr}\}. \tag{2.8}$$

Meinshausen and Bühlmann (2010) suggested that the subsample size is $[n/2]$ and a reasonable range for π_{thr} is $[1/2,\, 1)$ and the stability selection results are little sensitive to the choices of π_{thr}.

To establish the theoretical result for the estimator of (2.6) using random weights, the Sparse Riesz Condition (Zhang and Huang, 2008; Meinshausen and Bühlmann, 2010) needs to be satisfied by $\tilde{\mathbf{X}}^{\mathbf{w}}$ in (2.4). Let

$$\phi_{min}(m) = \min_{|A|=m, \mathbf{v} \in \mathbb{R}^m} \frac{\parallel \tilde{\mathbf{X}}_A^{\mathbf{w}} \mathbf{v} \parallel^2}{p^2 \parallel \mathbf{v} \parallel^2}, \qquad \phi_{max}(m) = \max_{|A|=m, \mathbf{v} \in \mathbb{R}^m} \frac{\parallel \tilde{\mathbf{X}}_A^{\mathbf{w}} \mathbf{v} \parallel^2}{p^2 \parallel \mathbf{v} \parallel^2} \tag{2.9}$$

for ranks $0 \leq m \leq p$. In (2.9), $\phi_{min}(m)$ and $\phi_{max}(m)$ are called minimum and maximum eigenvalues of the design $\tilde{\mathbf{X}}_A^{\mathbf{w}}$ of rank $m \leq p$. The Sparse Riesz Condition is satisfied if

$\phi_{min}(m)$ and $\phi_{max}(m)$ are bounded below from 0 and above from ∞ for all $m \leq p$:

$$0 < \phi_* \leq \phi_{min}(m) \leq \phi_{max}(m) \leq \phi^* < \infty \quad \forall m \leq p, \tag{2.10}$$

where ϕ_* and ϕ^* are some positive finite constants. Properties of $\phi_{min}(m)$ and $\phi_{max}(m)$ can be found in Zhang and Huang (2008) and Meinshausen and Bühlmann (2010). Lemma 1 in Web Appendix A indicates that for design $\tilde{\mathbf{X}}^{\mathbf{w}}_A$ of rank m in model (2.4), the Sparse Riesz Condition is satisfied.

Let $\phi_* = \min_{m \leq p} \phi_{min}(m)$, $\phi^* = \max_{m \leq p} \phi_{max}(m)$, and define:

$$C = \frac{\phi^*}{\phi_*}, \qquad r_1 = 1 + C, \qquad r_2 = \frac{3C}{2}. \tag{2.11}$$

THEOREM 1. *Let q be the number of nonzero coefficients in model (2.4), and $\hat{S}^\lambda = \{k : \hat{\beta}^\lambda_k \neq 0\}$ be the set of selected predictors using (2.6). For r_1 and r_2 defined in (2.11), suppose $q \geq 1$, the following assertions hold for all penalization tuning parameter $\lambda \geq inf\{\lambda : r_1 q + 1 \leq p\}$:*

$$\hat{q}(\lambda) \leq \tilde{q} = \#\{j : \hat{\beta}^\lambda_j \neq 0 \ or \ j \in S\} \leq r_1 q, \tag{2.12}$$

$$\tilde{B}^2(\lambda) = \| \ (\boldsymbol{I} - \hat{\boldsymbol{P}})\tilde{\mathbf{X}}^{\boldsymbol{w}}\boldsymbol{\beta} \ \|^2 \leq r_2 \frac{q\lambda^2}{\phi^* p^2}, \tag{2.13}$$

where $\hat{\boldsymbol{P}}$ is the projection to the span of the selected design vectors $\{\tilde{x}^{\boldsymbol{w}}_j, j \in \hat{S}^\lambda\}$,

$$\Delta^2_2(\lambda) = \sum_{j \in S} |\beta_j|^2 \mathbb{1}_{\{\hat{\beta}^\lambda_j = 0\}} \leq r_2 \frac{q\lambda^2}{\phi^* \phi_* p^4}, \tag{2.14}$$

and $\hat{q}(\lambda)$ is the cardinality of $\hat{S}^\lambda$; $\tilde{B}^2(\lambda)$ measures the bias of the estimated model, and $\Delta^2_2(\lambda)$ counts the sum of squared nonzero coefficients in the true model that are missed in the estimated model, respectively.

For a given weight $\mathcal{W}$, (2.14) provides an upper bound of the sum of the coefficients of the variables that are missed in the selection procedure but have large coefficients in the true model. An immediate consequence of Theorem 1 leads to the following result.

THEOREM 2. *Let W_k be uniformly generated from $[u, 1]$ for some $u \in (0, 1)$. Let S be the set of predictors having nonzero coefficients in model (2.4), and q be the cardinality of S. If $\hat{S}^\lambda$ is the set of selected predictors using (2.6) , for any $\lambda \geq inf\{\lambda : r_1 q + 1 \leq p\}$,*

$$\hat{S}^\lambda \cap S^c = \varnothing, \tag{2.15}$$

and

$$(S \setminus S_{small}) \subseteq \hat{S}^\lambda, \tag{2.16}$$

where $S_{small} = \{k : \beta_k \leq \sqrt{1.5q}\lambda/u^2\}$.

Theorem 1 and Theorem 2 are proved in Web Appendix B and Web Appendix C, respectively. Result (2.15) states that the stable estimate of structural dimension selects no noise directions in the CS, meaning that the false positive selection is avoided by the proposed method. Result (2.16) concludes that all directions with sufficiently large signals, having eigenvalues greater than $\sqrt{1.5q}\lambda/u^2$, will be selected, although directions with small signals will be ignored. In general, we don't want to miss directions with small positive eigenvalue coefficients. Practically, however, leaving out weak signals (directions with small eigenvalue coefficients) will have no major effects on later analysis, since all sufficiently strong signals are kept in the downstream analysis. In addition, the magnitude of the eigenvalues (smallness, hence, the weakness of the signals) can be controlled by λ and u. Therefore, any adverse effect, if possible, can be reduced and controlled by the analyst.

Comparing to the results of Meinshausen and Bühlmann (2010), Theorem 2 requires weaker assumptions. These advantages are gained because model (2.4) has no error term, and because of the special orthogonal structure of predictors $\tilde{\mathbf{X}}^{\mathbf{w}}$ in (2.5). To implement the stable algorithm, we incorporate subsampling and randomized weight schemes in the following steps.

1. Draw a random subsample of size $\lfloor n/2 \rfloor$ without replacement from the original sample, and obtain the dimension reduction kernel matrix $\mathbf{M}_{\frac{n}{2}}$ from the subsample.

2. Obtain an estimate for the subsample.

    - Generate random weights $\hat{w}_i$, $i = 1, 2, ..., p$, as: fix a $u \in (0.2, 0.8)$ and fix a threshold probability $p_w \in (0, 1)$, sample random variables u_i from $U(0, 1)$ independently. If $u_i \leq p_w$, let $\hat{w}_i = u$, otherwise let $\hat{w}_i = 1$.

    - Apply eigenvalue decomposition on $\mathbf{M}_{\frac{n}{2}}$ to obtain a set of orthonormal eigenvectors $\hat{\eta}_i$. Solve (2.6) for $\hat{\boldsymbol{\beta}}$ by fixing the values $\hat{\eta}_i$ with the random weights $\hat{w}_i$ generated above. Obtain the number of non-zero elements in $\hat{\boldsymbol{\beta}}$ as an estimate of structural dimension for these weights.

    - Repeat above steps N_1 times and obtain N_1 estimates of the structural dimensions for the subsample.

    - For a given cutoff probability π_{thr} described in (2.8), the sample π_{thr}^{th} quantile of these N_1 estimates is our estimated $\hat{k}$ for this subsample.

3. Repeat Steps 1 and 2 N_2 times to obtain N_2 estimates of the structural dimension. For a given cutoff probability π_{thr} described in (2.8), the sample π_{thr}^{th} quantile of these N_2 estimates is our estimated $\hat{k}$.

In the second sub-step in Step 2, we fix the values $\hat{\eta}_i$ as the eigenvectors of $\mathbf{M}_{\frac{n}{2}}$ and only estimate $\hat{\hat{\boldsymbol{\beta}}}$. Zhu et al. (2010a) used the iterative algorithm to simultaneously estimate both $\hat{\hat{\boldsymbol{\beta}}}$ and $\hat{\eta}_i$. Both algorithms give very similar results (not reported) except our algorithm is less computationally intensive. In Step 2, we use the same BIC criterion as in Zhu et al.

(2010a) to select the tuning parameter. Our algorithm implements $N_{12} = N_1 \times N_2$ total computations to obtain a stable result. We set $N_1 = N_2 = 500$, $u = 0.5$, $p_w = 0.8$, and $\pi^{th}_{thr} = 0.85$ in our simulation studies.

2.2.2 Sparse estimation of the central subspace

In this section, we propose a general method for obtaining a sparse solution when a matrix belongs to a Grassmann Manifold. Suppose that $V = (v_1, \cdots, v_k)$ is a $p \times k$ matrix in a Grassmann Manifold with rank k. Set $V^* = (V, V^{\perp})$ so that V^* is a $p \times p$ nonsingular matrix. Then one can always find a nonsingular symmetric matrix G such that $V^{*\prime}GV^* = \mathbf{I}_p$, and construct a symmetric and positive definite matrix $\mathbf{M}$ so that

$$\mathbf{M}V^* = GV^*D, \tag{2.17}$$

where D is a diagonal matrix with chosen $\rho_1 > \rho_2 > \cdots > \rho_p > 0$. Thus V can be regarded as the eigenvectors of $\mathbf{M}$ corresponding to the eigenvalues of $\rho_1, \cdots, \rho_k$, under the constraint $V'GV = \mathbf{I}_k$. Equation (2.47) is established in Web Appendix D. We now provide a Grassmann Manifold sparse estimate for V under aforementioned G, D (hence, $\mathbf{M}$) setup. Note that for given a V we can construct an eigen-decomposition problem as above. However, given such an $\mathbf{M}$ and G, one can find a unique solution for V^* (hence, V) and D. The following result provides an equivalent solution.

Corollary 1. *Let m_i, $i = 1, \cdots, p$ denote the columns of $\boldsymbol{M}^{1/2}$, which is the square-root of $\boldsymbol{M}$ as defined in (2.47), and let $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ be $p \times k$ matrices. Consider the following optimization problem*

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = \min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left\{ \sum_{i=1}^{p} \parallel G^{-1}m_i - \boldsymbol{\alpha}\boldsymbol{\beta}^T m_i \parallel_G^2 \right\} \tag{2.18}$$

subject to $\alpha^T G \alpha = \boldsymbol{I}_k$, where the norm is the inner product with respect to G. Then $\hat{\beta}_j \propto v_j$, *where $\hat{\beta}_j$ is the j^{th} column of $\hat{\boldsymbol{\beta}}$ and v_j is the eigenvector of $\boldsymbol{M}$ corresponding to the j^{th} largest eigenvalue in the eigenvalue problem mentioned above.*

To obtain a sparse estimate, we can simply add a penalty term

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = \min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left\{ \sum_{i=1}^{p} \parallel G^{-1}m_i - \boldsymbol{\alpha}\boldsymbol{\beta}^T m_i \parallel_G^2 + \lambda \sum_{h=1}^{p} \frac{|\beta_{jh}|}{W_h} \right\}. \tag{2.19}$$

Corollary 2. *Given $\boldsymbol{\alpha}$ in (2.19), for each $\hat{\boldsymbol{\beta}}_j$, $j = 1, \cdots, k$ let*

$$\hat{\boldsymbol{\theta}}_j = \min_{\boldsymbol{\theta}_j} \left\{ \parallel \tilde{Y} - \tilde{\mathbf{X}}\boldsymbol{\theta}_j \parallel^2 + \lambda \sum_{h=1}^{p} \frac{|\theta_{jh}|}{W_h} \right\}, \tag{2.20}$$

where $\tilde{\mathbf{X}} = \boldsymbol{M}^{1/2}$, $\tilde{Y} = \boldsymbol{M}^{1/2}\alpha_j$ and the weights W_h are uniformly generated from $[u, 1]$ for some $u \in (0, 1)$. Then $\hat{\boldsymbol{\theta}}_j = \hat{\boldsymbol{\beta}}_j$.

Corollary 1 and Corollary 2 can be directly proved following Proposition 1 to Proposition 3 of Li (2007). Thus we omit their proofs. We call this the Grassmann Manifold sparse estimate (GMSE) method. Suppose that in (2.47) the chosen $\rho_1 > \rho_2 > \cdots > \rho_p > 0$ are all bounded below by 0 and above by ∞, then the Sparse Riesz Condition is satisfied. Therefore, we have the following result.

THEOREM 3. *Let S be the set of predictors having nonzero coefficients in model (2.47) and q be the cardinality of S. Let $\hat{S}^\lambda$ denote the set of selected predictors using (2.20). The ϕ_{min} and ϕ_{max} are defined in the same way as in (2.9); and $u^2 = \nu\phi_{min}(m)/m$ for any $\nu \in ((3/\kappa)^2, 1/\sqrt{2})$, and $m = \bar{C}q^2$. Assume that*

$$\frac{\phi_{max}(\bar{C}q^2)}{\phi_{min}^{3/2}(\bar{C}q^2)} < \frac{\bar{C}}{\kappa} \tag{2.21}$$

for some $\bar{C} > 1$ and some $\kappa \geq 4$. If $q \geq 3$, for any $\lambda > inf\{\lambda : r_1 q + 1 \leq p\}$ with r_1 defined *in (2.11), there exists some* $\tilde{p} \in (0, 1)$ *such that for all* $\pi_{thr} \geq 1 - \tilde{p}$

$$\hat{S}^{\lambda} \cap S^{c} = \varnothing, \tag{2.22}$$

and

$$(S \setminus S_{small}) \subseteq \hat{S}^{\lambda}, \tag{2.23}$$

where $S_{small} = \{k : \beta_k \leq \sqrt{1.5}\bar{C}q^{3/2}\lambda\}$.

Theorem 3 is proved in Web Appendix E. The results of Theorem 3 are similar to the results of Theorem 2. However, Theorem 3 cannot be directly applied to $\mathbf{M}$, a dimension reduction matrix obtained based on the original predictors $\mathbf{X}$. This is because the dimension of $\mathbf{M}$ is $k < p$, which means that some eigenvalues of $\mathbf{M}$ are 0, hence, are not bounded by 0. Thus the Sparse Riesz Condition is violated. Nevertheless, we can fix this problem by working on $\mathbf{M}_{\delta} = \mathbf{M} + \delta G$ for some positive constant δ, where $G = \hat{\Sigma}_x$. Then results of Theorem 3 hold for $\mathbf{M}_{\delta}$ as long as $\delta > 0$ (See Web Appendix D for details). Empirical evidence in Web Appendix F shows little effect of varying δ on the estimates. Note that our approach is proposed due to a referee's question on the theoretical development for Li (2007)'s approach. Li (2007)'s method used an additional tuning parameter λ_2 for the purpose of uniqueness of the eigenvectors in theory, but added difficulties for us when proving our result, as this violates the Sparse Riesz Condition. In addition, the selection of λ_2 in Li's algorithm adds extra computation time, while our computation time is less but results are essentially the same. See Web Appendix G for a further detailed comparison of our method with Li (2007)'s approach.

A stable Grassmann Manifold Sparse Estimate (SGMSE) for the CS can then be obtained by the following steps:

1. Draw a random subsample of size $\lfloor n/2 \rfloor$ without replacement from the original sample, and form a sample dimension reduction matrix $\mathbf{M}_{\frac{n}{2}}$.

2. Test the structural dimension $\hat{k}$ of $\mathbf{M}_{\frac{n}{2}}$ based on the subsample by existing methods such as the stable algorithm mentioned in Section 2.2.1. If $\hat{k} = k$, proceed to next step. Otherwise, go back to step 1 and draw another random subsample.

3. For a small positive constant $\delta > 0$, let $\mathbf{M}_{\frac{n}{2},\delta} = \mathbf{M}_{\frac{n}{2}} + \delta G$ where G is the covariance matrix of the subsample. Find the usual estimate of $\boldsymbol{\beta}$ based on $\mathbf{M}_{\frac{n}{2},\delta}$ without the LASSO constraint as an initial value for $\boldsymbol{\alpha}$.

4. Find a solution for the subsample.

   - Generate random weights w_{jh}, $h = 1, 2, ..., p$, for fixed j, $j = 1, \cdots, k$ as: fix an $u \in (0.2, 0.8)$ and fix a threshold probability $p_w \in (0, 1)$, sample random variables u_h from $U(0, 1)$ independently. If $u_h \leq p_w$, let $w_{jh} = u$, otherwise let $w_{jh} = 1$.

   - Given a fixed $\boldsymbol{\alpha}$, solve k independent LASSO problems (2.20) to obtain an estimate $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, ..., \hat{\boldsymbol{\beta}}_k)$ with the generated weights.

   - For a fixed $\hat{\boldsymbol{\beta}}$, carry out singular value decomposition of $G^{-1/2}\mathbf{M}_{\frac{n}{2},\delta}\boldsymbol{\beta} = UDV^T$, and update $\boldsymbol{\alpha} = G^{-1/2}UV^T$

   - Repeat the two previous sub-steps until the procedure converges, and we have an estimate $\hat{\boldsymbol{\beta}}$ for these generated random weights.

   - Repeat the four previous sub-steps N_1 times and obtain N_1 sparse estimates $\hat{\boldsymbol{\beta}}^1, \cdots, \hat{\boldsymbol{\beta}}^{N_1}$.

   - For each x_i, if its frequency of appearing in $\hat{\boldsymbol{\beta}}^1, \cdots, \hat{\boldsymbol{\beta}}^{N_1}$ is less than the preset cutoff probability π_{thr}, set all the elements in the i^{th} row of all $\hat{\boldsymbol{\beta}}^1, \cdots, \hat{\boldsymbol{\beta}}^{N_1}$ to 0, say, $\hat{\boldsymbol{\beta}}^1_s, \cdots, \hat{\boldsymbol{\beta}}^{N_1}_s$. Then, the first k eigenvectors of $\sum_{i=1}^{N_1} \hat{\boldsymbol{\beta}}^i_s (\hat{\boldsymbol{\beta}}^i_s)'$ will be our estimate, say, $\hat{\boldsymbol{\beta}}^s$ for this subsample in the step.

5. Repeat Steps 1 to 4 N_2 times to obtain N_2 estimates $\hat{\boldsymbol{\beta}}^{s,1} \cdots, \hat{\boldsymbol{\beta}}^{s,N_2}$. For each x_i, if its frequency of appearing in $\hat{\boldsymbol{\beta}}^{s,1} \cdots, \hat{\boldsymbol{\beta}}^{s,N_2}$ is less than the preset cutoff probability π_{thr},

set all the elements in the i^{th} row of all $\hat{\boldsymbol{\beta}}^{s,1} \cdots , \hat{\boldsymbol{\beta}}^{s,N_2}$ to 0, say, $\hat{\boldsymbol{\beta}}_s^{s,1} \cdots , \hat{\boldsymbol{\beta}}_s^{s,N_2}$. Then the first k eigenvectors of $\sum_{i=1}^{N_2} \hat{\boldsymbol{\beta}}_s^{s,i}(\hat{\boldsymbol{\beta}}_s^{s,i})'$ will be the final estimate.

In the simulation studies, we use the BIC criterion of Li (2007) for selecting tuning parameters. We also set $N_1 = 500$, $N_2 = 500$, $u = 0.5$, $p_w = 0.8$, and $\pi_{thr}^{th} = 0.85$.

2.2.3 Ensemble approaches in dimension reduction

In this section, we introduce an ensemble approach to obtain a stable nonsparse estimate of the central subspace based on subsample with a threshold, and we also solve the well-known problem in the sliced inverse methods: the choice of number of slices, by using the ensemble idea.

2.2.3.1 Nonsparse estimation of the central subspace

Let $\mathbf{M}$ be a method-specific dimension reduction kernel matrix based on the standardized predictors $\mathbf{Z}$. We propose the following general algorithm to provide a nonsparse estimate of the CS with subsampling, assuming the structural dimension is known.

1. Draw a random subsample of size $\lfloor n/2 \rfloor$ without replacement from the original sample, and form a sample dimension reduction matrix $\mathbf{M}_{\frac{n}{2}}$ based on the standardized predictors $\mathbf{Z}$ of the subsample.

2. Test the structural dimension $\hat{k}$ of $\mathbf{M}_{\frac{n}{2}}$ based on the subsample by existing methods such as the stable procedure proposed in Section 2.2.1. If $\hat{k} = k$, proceed to the next step. Otherwise, go back to step 1 and draw another random subsample.

3. Obtain an estimated $p \times k$ basis matrix $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, ..., \hat{\boldsymbol{\beta}}_k)$ of the CS and covariance matrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{i},\ \mathbf{x}}$ for the accepted subsample.

4. Repeat Step 1 to Step 3 N_2 times to obtain N_2 estimated basis matrices $\hat{\boldsymbol{\beta}}_1, \cdots, \hat{\boldsymbol{\beta}}_{N_2}$ and covariance matrices $\hat{\boldsymbol{\Sigma}}_{1,x}, \cdots, \hat{\boldsymbol{\Sigma}}_{N_2,x}$. Let $\hat{\mathbf{M}} = \frac{1}{N_2}\sum_{i=1}^{N_2}(\hat{\boldsymbol{\Sigma}}_{i,x}^{-1/2}\hat{\boldsymbol{\beta}}_i)(\hat{\boldsymbol{\Sigma}}_{i,x}^{-1/2}\hat{\boldsymbol{\beta}}_i)'$. Conduct an eigenvalue decomposition of $\hat{\mathbf{M}}$. The final estimate of basis directions is $\hat{\boldsymbol{\beta}}^f = (\hat{\beta}_1^f, ..., \hat{\beta}_k^f)$ where $\hat{\beta}_j^f$'s, for $j = 1, \cdots, k$ are eigenvectors of $\hat{\mathbf{M}}$ corresponding to the k largest eigenvalues of $\hat{\mathbf{M}}$.

In this approach, we find a new use of subsampling in nosparse estimation. In addition, the second step is a threshold idea that filters out the estimates due to sampling bias. This adaptive step can significantly improve the estimation accuracy for the CS after combining results from low-noise level subsamples. The ensemble idea in Step 4 that aggregates subsampling estimates differs from that of Xia et al. (2002) where they aggregated local estimates to obtain a final solution, and that of Yin and Li (2011) where they assembled estimates with different methods from the entire data.

2.2.3.2 The number of slices in inverse dimension reduction methods

A well-known problem in the sliced inverse methods such as SIR (Li, 1991) and SAVE (Cook and Weisberg, 1991) is that the results are not stable if the number of slices changes. A referee asked a question that whether this is an issue in SED method and if so, whether we can stabilize it. This is also an issue in the SED method and so we adopt an ensemble idea to stabilize the solution and further improve it with the subsampling idea. Note that this ensemble idea has been used by Yin and Li (2011), in which intuitively the repeated use of the data will enhance the accuracy of the results.

We define the following ensemble dimension reduction matrix (using SIR as our example):

$$\mathbf{M} = \sum_{H \in \mathcal{H}} \mathbf{M}_H = \sum_{H \in \mathcal{H}} \left\{ \sum_{i=1}^{H} p_i E(\mathbf{Z}|Y) E(\mathbf{Z}|Y)' \right\} \tag{2.24}$$

where $\mathbf{M}_H$ is the dimension reduction kernel matrix using H slices and $\mathcal{H}$ is a range of reasonable choices for H. A reasonable choice for H is typically between 5 and 25, depending on the sample size. Since each $\mathbf{M}_H$ estimates the same CS, the aggregation in (2.24) is equivalent to calculating the expectation of $\mathbf{M}_H$ over $\mathcal{H}$. This approach is particularly useful in our proposed method because we use a penalized estimation procedure for nonzero eigenvalues. When adding the $\mathbf{M}_H$ together, the differences in the eigenvalues become relatively larger. In this way, the signals in the CS are amplified compared to random noise directions. Our stable procedure incorporating subsampling further improves this ensemble method, stabilizing the tuning parameters (see related simulation studies in Web Appendix H).

2.3 Simulations

In this section, we conduct numerical studies to evaluate the performance of our proposed stable methods. We first focus on the stable estimation of structural dimension of the CS. Then we demonstrate the stable procedure to estimate the CS in both the sparse case and the nonsparse case. Finally, we apply our proposed method to analyze a real data set.

2.3.1 Estimating structural dimension

We compare results between the original SED method and the proposed stable procedure for SIR, PHD and SAVE using the same models that were used by Zhu et al. (2010a). Consider the following six models which were used in Zhu et al. (2010a),

$$Y = 3X_1 + \sigma\varepsilon, \tag{2.25}$$

$$Y = (X_1 + X_2)^3 + 2\exp(X_3 + X_4) + \sigma\varepsilon, \tag{2.26}$$

$$Y = 2\cos(X_1) + \sigma\varepsilon, \tag{2.27}$$

$$Y = 2\cos(X_1) + \cos(X_2) + \sigma\varepsilon, \tag{2.28}$$

$$Y = 2X_1^2 + \sigma\varepsilon, \tag{2.29}$$

$$Y = 2X_1^2 + X_2^2 + \sigma\varepsilon. \tag{2.30}$$

The predictors $\mathbf{X} = (X_1, \cdots, X_p)$ are generated from a p-dimensional multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}_p$. In each case, the error ε is generated independently from the standard normal distribution. The noise level σ is set to be 0.25 according to Zhu et al. (2010a). We use $p = 10$ predictors with sample size $n = 400$ for all models. For example, in model (2.26), among all 10 predictors, only two linear combinations of all 10 predictors are associated with the response Y. Hence, in this case the structural dimension is 2. Models (2.25) and (2.26) are estimated by SIR; models (2.27) and (2.28) are estimated by PHD; and models (2.29) and (2.30) are estimated by SAVE. For SIR and SAVE, H = 10 slices are used.

Figure 2.1 is the plot of accuracy for identifying structural dimension over 100 data replicates versus the tuning parameter values for the six models respectively. We use the algorithm described in Section 2.2.1 with $N_{12} = 500 \times 500$. It shows that with a good choice of the penalization tuning parameter, the original SED procedure (dashed lines) can successfully estimate the structural dimension with high accuracy. For stable procedures (solid lines), high accuracy can be achieved by a much wider range of tuning parameters. Hence, the stable procedures significantly increase the chance of correctly estimating the structural dimension.

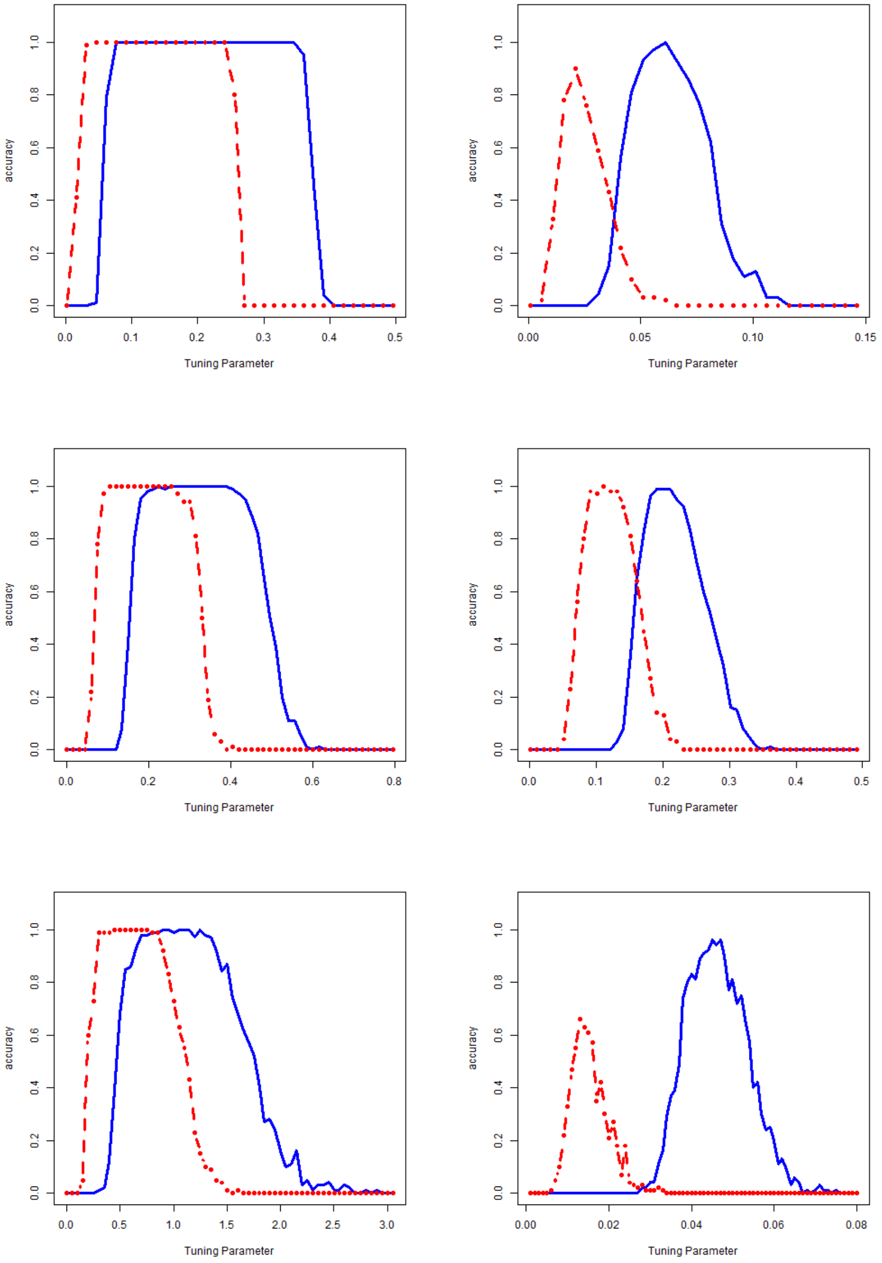

Figure 2.1: Accuracy of Stable SED for testing structural dimension

2.3.2 Estimating the central subspace

2.3.2.1 The sparse case

To illustrate the effectiveness of the stable procedure in obtaining a sparse estimate of the CS, we compare results from the original dimension reduction methods, GMSE, and stable GMSE (SGMSE) for the models used by Li (2007):

$$Y = \text{sign}(\boldsymbol{\beta}_1^T X)\log(|\boldsymbol{\beta}_2^T X + 5|) + 0.2\varepsilon, \tag{2.31}$$

$$Y = \cos(2\boldsymbol{\beta}_1^T X) - \cos(\boldsymbol{\beta}_2^T X) + 0.5\varepsilon. \tag{2.32}$$

Model (2.31) is simulated to compare results for SIR with $\boldsymbol{\beta}_1 = (1,1,1,1,0,...,0)^T$, $\boldsymbol{\beta}_2 = (0,...,0,1,1,1,1)^T$, and $p = 20$. Model (2.32) is simulated to compare results for PHD method, where $\boldsymbol{\beta}_1 = (1,0,...,0)^T$, $\boldsymbol{\beta}_2 = (0,1,0,...,0)^T$ and $p = 10$. The predictors X_i and the error term ε are generated independently from the standard normal distribution. For the SIR model, we use a sample size $n = 200$ with 10 slices. For the PHD model, the sample size is $n = 400$. In addition, we use SAVE with model (2.30) with the same settings in Section 2.3.1. The constant δ is set to be 0.01 in GMSE and SGMSE procedures.

Since the main goal of sparse solution is variable selection, we report the true positive rate (TPR): the ratio of the number of correctly identified active predictors to the number of truly active predictors, and the false positive rate (FPR): the ratio of the number of falsely identified active predictors to the number of true inactive predictors. Note that $0 \leq$ TPR, FPR ≤ 1, and a better estimate has bigger TPR and smaller FPR. Table 2.6 reports the results by using 100 data replicates with $N_{12} = 500 \times 500$ in the stable algorithm. Among the three methods, SGMSE is the best. Simulation results comparing GMSE to SSDR (Li, 2007) and their counterparts are reported in Web Appendix G.

Models	d	Methods	TPR	FPR
Model (2.31)	2	SIR	1.000	1.000
		GMSE SIR	1.000	0.023
		SGMSE SIR	1.000	0.000
Model (2.32)	2	PHD	1.000	1.000
		GMSE PHD	1.000	0.274
		SGMSE PHD	1.000	0.095
Model (2.30)	2	SAVE	1.000	1.000
		GMSE SAVE	1.000	0.263
		SGMSE SAVE	1.000	0.058

Table 2.1: Stable Grassmann manifold sparse estimation

2.3.2.2 The nonsparse case

For the same models considered in Section 2.3.1, assuming the structural dimension k is known, we apply corresponding methods to estimate the CS with the stable procedure described in Section 2.2.3.1. We compare the accuracy of our estimates based on the vector correlation coefficient that measures the distance between two subspaces $S(\mathbf{A})$ and $S(\mathbf{B})$, which are spanned by the columns of $p \times k$ matrices $\mathbf{A}$ and $\mathbf{B}$, respectively, with $\mathbf{A}^T\mathbf{A} = \mathbf{B}^T\mathbf{B} = \mathrm{I}_k$. The vector correlation coefficient q is defined by Hotelling (1936): $q = \sqrt{|\mathbf{B}^T\mathbf{A}\mathbf{A}^T\mathbf{B}|}$, where $|\cdot|$ is the determinant of a matrix. The range of q is between 0 and 1 and two subspaces with a bigger value of q are closer to each other. This measure has been used by Ye and Weiss (2003), and others. Similar distance measures between two subspaces may be used as well, including trace correlation (Hooper, 1959), and $a_\Delta(\mathbf{A},\mathbf{B}) = |\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T - \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T|$ (Li et al., 2005).

We generate 100 data replicates of size $n = 400$ for each model defined in Section 2.3.1. For each individual data replicate, $N_2 = 500$ subsamples with size of $\lfloor n/2 \rfloor$ are sampled, from all possible $\binom{400}{200}$ subsamples, to implement the stable procedure following the algorithm described in Section 2.2.3.1. We used 10 slices in SIR and SAVE and $p = 10$ for all models. The reported vector correlation coefficient in Table 2.2 is the average of 100 vector correlations between the estimated CS and the true CS for each model. For each model we compute the proportion of improvement by the stable procedure among 100 replicates. Except for model (2.27), the stable procedure can produce more accurate nonsparse estimates of the CS about 65% of the time. The average vector correlation coefficients are improved for almost all models expect for model (2.27) where the stable method provides exactly the same result as the original PHD method.

Models	(2.25)	(2.26)	(2.27)	(2.28)	(2.29)	(2.30)
Methods	SIR		PHD		SAVE	
Original Method	0.9964	0.9214	0.9910	0.9502	0.9866	0.9520
Stable Method	0.9968	0.9282	0.9910	0.9512	0.9868	0.9539
Percentage of Improvement	83%	73%	50%	67%	65%	75%

Table 2.2: Stable nonsparse estimates of the CS

2.3.3 Hitters' salary data

Zhu et al. (2010a) applied their methodology to the Hitters' salary data. The complete data set can be obtained at *http://www.psych.yorku.ca/lab/psy6140/bb/basedata.htm*. This data set contains the salary of 263 baseball hitters. There are 16 independent variables X_1 to X_{16}: times at bat, hits, home runs, runs, runs batted in and walks in 1986, years in major leagues, career times at bat, career hits, career home runs, career runs, career batted in, career walks, put outs, assistances and errors. The response variable Y, is the logarithm of annual salary in 1987.

We first use the stable SED with SIR to estimate the structural dimension for the Hitters' salary data. We obtained $\hat{k} = 2$, which agrees with what Zhu et al. (2010a) found. We then proceed to compare the estimates by SIR, GMSE-SIR and SGMSE-SIR, using $\hat{k} = 2$.

Table 2.3 summarizes the results. SIR produces non-sparse estimates and GMSE-SIR shrinks 2 components in the two directions to 0, while SGMSE-SIR shrinks 12 components in

	SIR		**GMSE SIR**		**SGMSE SIR**	
X	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
x_1	0.00098	0.14931	0.00646	0.03065	0.00670	0.15821
x_2	-0.01879	-0.50472	-0.00170	-0.26082	**0.00000**	$\boxed{-0.91943}$
x_3	-0.15003	-0.18736	0.05339	-0.83360	**0.00000**	**0.00000**
x_4	0.01264	0.03314	-0.00918	0.32519	**0.00000**	**0.00000**
x_5	0.05042	0.02288	-0.02336	0.27517	-0.03853	**0.00000**
x_6	-0.02434	-0.46535	0.00851	-0.15560	-0.00836	$\boxed{-0.31913}$
x_7	-0.97714	0.38114	0.99814	**0.00000**	$\boxed{0.99909}$	**0.00000**
x_8	0.00437	0.00207	-0.00471	-0.00458	-0.00533	-0.02455
x_9	-0.00522	0.00626	0.00645	0.04596	0.01260	0.12801
x_{10}	0.00429	0.03892	**0.00000**	0.11614	**0.00000**	**0.00000**
x_{11}	-0.00807	-0.09192	0.00689	-0.05288	-0.00199	-0.09054
x_{12}	-0.00182	-0.04479	. 0.00071	-0.04548	-0.00352	**0.00000**
x_{13}	0.00278	0.06029	-0.00150	0.02085	0.00227	0.05004
x_{14}	0.00002	-0.02131	0.00029	0.00045	0.00056	0.00062
x_{15}	0.00089	-0.03395	0.00214	-0.00559	0.00182	0.00933
x_{16}	-0.13747	0.55423	-0.00231	0.09754	**0.00000**	**0.00000**
	$v_\Delta = 0.2627$		$v_\Delta = 0.2919$		$v_\Delta = 0.1075$	

Table 2.3: Estimates for Hitters' data

the two directions to 0 from which variables x_3, x_4, x_{10} and x_{16} are completely noninformative.

To evaluate which estimate is better, we use a similar measure based on the idea of Ye and Weiss (2003) to measure the variation of each estimate. The variation (v_Δ) is computed as one minus the vector correlation between the two estimates based on the entire sample and a subsample ($v_\Delta = 1 - q$). The basic idea is that both estimates aim to estimate the true central subspace. A better estimate will have a bigger value of vector correlation, thus smaller v_Δ. We calculate the variation as the average of 100 variations (by using 100 subsamples). The variation for SIR is 0.2627, the variation for GMSE-SIR is 0.2919, while the variation for SGMSE-SIR is 0.1075. Thus the result of SGMSE-SIR is the best.

2.4 Discussion

In this paper, we proposed stable estimation procedures for four aspects of dimension reduction: estimating structural dimension, obtaining a sparse estimate of the CS, providing a nonsparse estimate of the CS and solving the number of slices issue in the sliced inverse methods. Our stable procedures in dimension reduction methods always gain accuracy as simulation studies indicated, compared with the usual approaches. In the subsampling step of the stable procedure, we used half $\lfloor n/2 \rfloor$ of the original sample as suggested by Meinshausen and Bühlmann (2010). However, different choices of subsample size may be used. On the other hand, the cutoff percentage π_{thr} of choosing important variables in the subsample is also arbitrary. A bigger value of π_{thr} gives more stable and more sparse selection results, but may miss some important variables. If π_{thr} is too small, the algorithm may not be effective in the variable selection sense. The proposed choices of the parameters by Meinshausen and Bühlmann (2010) worked well in our study, and the overall merits of subsampling and random weights used in penalized dimension reduction methods have been well illustrated by the simulation results in our investigation.

Our proposed stable methods work for the case where $n > p$ and the sample covariance matrix of predictors is invertible. An interesting question is how to establish stable dimension reduction methods for the case where $n < p$ or highly correlated predictors. Cook et al. (2007) and Li and Yin (2008) developed dimension reduction methods when the sample covariance matrix is not invertible, and stable procedures incorporating these methods will be a future research topic. However, when p becomes divergent, effective dimension reduction methods have yet to be developed before considering any stable estimation procedure. This can be an even more challenging problem.

Appendix A

The Sparse Riesz Condition controls the range of eigenvalues of covariance matrices of subsets of a fixed number of design vectors. We will show in Lemma 1 below that for design $\tilde{\mathbf{X}}^{\mathbf{w}}_A$ of rank m, in model (2.4), the Sparse Riesz Condition is satisfied, and in Lemma 2 that the Sparse Riesz Condition guarantees a finite bound.

Lemma 1 and Lemma 2 are needed to prove Theorem 1 and Theorem 3.

Lemma 1. For model (2.4), for any $\tilde{\mathbf{X}}^{\mathbf{w}}_A$ of rank $m \leq p$, where $A \subseteq \{1, 2, \cdots, p\}$ with $|A| = p$, the Sparse Riesz Condition (2.10) is satisfied.

Proof: In model (2.4), for design $\tilde{\mathbf{X}}^{\mathbf{w}}_A$ of rank m, we have

$$\frac{\parallel \tilde{\mathbf{X}}^{\mathbf{w}}_A \mathbf{v} \parallel^2}{p^2 \parallel \mathbf{v} \parallel^2} = \frac{(\tilde{\mathbf{X}}^{\mathbf{w}}_A \mathbf{v})' \tilde{\mathbf{X}}^{\mathbf{w}}_A \mathbf{v}}{p^2 \mathbf{v}'\mathbf{v}} = \frac{\mathbf{v}' \mathcal{D}_A \mathbf{v}}{p^2 \mathbf{v}'\mathbf{v}} = \frac{\sum\limits_{i=1}^{m} W_{Ai}^2 v_i^2}{p^2 \sum\limits_{i=1}^{m} v_i^2}.$$

The Sparse Riesz Condition is satisfied, because

$$\phi_{min}(m) = \min_{|A|=m, \mathbf{v} \in \mathbb{R}^m} \frac{\parallel \tilde{\mathbf{X}}^{\mathbf{w}}_A \mathbf{v} \parallel^2}{p^2 \parallel \mathbf{v} \parallel^2} = \min_{|A|=m, \mathbf{v} \in \mathbb{R}^m} \frac{\sum\limits_{i=1}^{m} W_{Ai}^2 v_i^2}{p^2 \sum\limits_{i=1}^{m} v_i^2} \geq \min_{|A|=m, \mathbf{v} \in \mathbb{R}^m} \frac{\sum\limits_{i=1}^{m} u^2 v_i^2}{p^2 \sum\limits_{i=1}^{m} v_i^2} = \frac{u^2}{p^2},$$

since W_{Ai} are generated from $[u, 1]$ for some $u > 0$. For the same reason,

$$\phi_{max}(m) = \max_{|A|=m, \mathbf{v} \in \mathbb{R}^m} \frac{\parallel \tilde{\mathbf{X}}^{\mathbf{w}}_A \mathbf{v} \parallel^2}{p^2 \parallel \mathbf{v} \parallel^2} = \max_{|A|=m, \mathbf{v} \in \mathbb{R}^m} \frac{\sum\limits_{i=1}^{m} W_{Ai}^2 v_i^2}{p^2 \sum\limits_{i=1}^{m} v_i^2} \leq \max_{|A|=m, \mathbf{v} \in \mathbb{R}^m} \frac{\sum\limits_{i=1}^{m} v_i^2}{p^2 \sum\limits_{i=1}^{m} v_i^2} = \frac{1}{p^2}.$$

Therefore, $0 < \frac{u^2}{p^2} \leq \phi_{min}(m) \leq \phi_{max}(m) \leq \frac{1}{p^2} < \infty$ for all $m \leq p$. □

Lemma 2. Let $\phi_{min}(m)$ and $\phi_{max}(m)$ be defined as in (2.9) and the Sparse Riesz Condition

(2.10) holds. Let $S_k \subset \{1, \cdots, p\}$, $\tilde{\mathbf{X}}^{\mathbf{w}}_k = (\tilde{\mathbf{x}}_j, j \in S_k)$ and $\boldsymbol{\Sigma}_{1k} = (\tilde{\mathbf{X}}^{\mathbf{w}}_1)'\tilde{\mathbf{X}}^{\mathbf{w}}_k/n$, Then

$$\frac{\| \mathbf{v} \|^2}{\phi^Z_{max}(|S_1|)} \leq \| \boldsymbol{\Sigma}_{11}^{-1/2}\mathbf{v} \|^2 \leq \frac{\| \mathbf{v} \|^2}{\phi^Z_{min}(|S_1|)}, \tag{2.33}$$

for all $\mathbf{v}$ of proper dimension, and that

$$\| \boldsymbol{\beta}_k \|_1^2 \leq \frac{\| \tilde{\mathbf{X}}^{\mathbf{w}}_k \boldsymbol{\beta}_k \|^2 \, |S_k|}{n\phi^Z_{min}(|S_k|)}. \tag{2.34}$$

Proof: Let $\mathbf{v}, \mathbf{h} \in \mathbb{R}^{|S_1|}$ and $\mathbf{v} = \boldsymbol{\Sigma}_{11}^{1/2}\mathbf{h}$. Hence, $\boldsymbol{\Sigma}_{11}^{-1/2}\mathbf{v} = \mathbf{h}$. By Lemma 1, since the Sparse Riesz Condition (2.10) holds, we have that $\phi_{min}(|S_1|) \leq \frac{\|\tilde{\mathbf{X}}^{\mathbf{w}}_1\mathbf{h}\|^2}{n\|\mathbf{h}\|^2} \leq \phi_{max}(|S_1|)$.

Note that $\| \tilde{\mathbf{X}}^{\mathbf{w}}_1 \mathbf{h} \|^2 = n \| \mathbf{v} \|^2$ and $\| \mathbf{h} \|^2 = \| \boldsymbol{\Sigma}_{11}^{-1/2}\mathbf{v} \|^2$.

So $\phi_{min}(|S_1|) \leq \frac{\|\mathbf{v}\|^2}{\|\boldsymbol{\Sigma}_{11}^{-1/2}\mathbf{v}\|^2} \leq \phi_{max}(|S_1|)$, and $\frac{1}{\phi_{max}(|S_1|)} \leq \frac{\|\boldsymbol{\Sigma}_{11}^{-1/2}\mathbf{v}\|^2}{\|\mathbf{v}\|^2} \leq \frac{1}{\phi_{min}(|S_1|)}$, which yields (2.33). The Cauchy-Schwarz inequality implies that $\| \boldsymbol{\beta}_k \|_1^2 \leq \| \boldsymbol{\beta}_k \|^2 \, |S_k|$. By the Sparse Riesz Condition, $\phi_{min}(|S_k|) \leq \frac{\|\tilde{\mathbf{X}}^{\mathbf{w}}_k\boldsymbol{\beta}_k\|^2}{n\|\boldsymbol{\beta}_k\|^2} \Longrightarrow \| \boldsymbol{\beta}_k \|^2 \leq \frac{\|\tilde{\mathbf{X}}^{\mathbf{w}}_k\boldsymbol{\beta}_k\|^2}{n\phi_{min}(|S_k|)}$, which yields (2.34). □

Appendix B

Proof of Theorem 1. The goal of Theorem 1 is to find upper bounds of (2.12), (2.13), and (2.14). Since $S = \{k : \beta_k \neq 0\}$ is the set of important predictors in the true model, we define the following sets in Table 2.4 to facilitate our proof.

	nonzero $\boldsymbol{\beta}_j : j \in S$	zero $\boldsymbol{\beta}_j : j \notin S$
S_1: selected j	S_3	S_4
$S_2 = S_1^c$	S_5	S_6

Table 2.4: Definitions of sets

In our case, $\hat{q} = q_1 = |S_1|$. Define $\mathbf{Q}_{kj}$ to be the selection of variables in S_k from S_j:

$$\mathbf{Q}_{kj}\boldsymbol{\beta}_j = \boldsymbol{\beta}_k, \quad \boldsymbol{\beta}_1' = \boldsymbol{\beta}_3'\mathbf{Q}_{31}, \quad \boldsymbol{\beta}_k = \{\beta_j, j \in S_k\}.$$

Let $\tilde{\mathbf{X}}_i^{\mathbf{w}} = (\tilde{\mathbf{x}}_j, j \in S_i)$, define

$$\boldsymbol{\Sigma}_{jk} = \frac{1}{n}(\tilde{\mathbf{X}}_j^{\mathbf{w}})'\tilde{\mathbf{X}}_k^{\mathbf{w}}, \qquad \mathbf{f}_j = (\tilde{\mathbf{X}}_{S_j}^{\mathbf{w}})'(Y - \tilde{\mathbf{X}}^{\mathbf{w}}\boldsymbol{\beta})/\lambda, \;\; j = 1, 3, 4. \tag{2.35}$$

By (2.5), we have

$$\boldsymbol{\Sigma}_{jk} = \begin{cases} \mathcal{D}_i/p^2 & j = k, \\ \mathbf{0} & j \neq k. \end{cases} \tag{2.36}$$

where $\mathcal{D}_i = \mathrm{diag}(W_{ij}^2)$, $j = 1, \cdots, q_i$. So

$$\boldsymbol{\Sigma}_{ii}^{-1} = p^2\mathcal{D}_i^{-1} = p^2\mathrm{diag}(W_{ij}^{-2}). \tag{2.37}$$

With $\mathbf{P}_1$ be the projection from $\mathbb{R}^n$ to the span of $\{\tilde{\mathbf{x}}_j : j \in S_1\}$, we define,

$$\mathbf{v}_{1j} = \frac{\lambda}{\sqrt{n}}\boldsymbol{\Sigma}_{11}^{-1/2}\mathbf{Q}_{j1}'\mathbf{f}_j, \qquad \mathbf{w}_k = (\mathbf{I} - \mathbf{P}_1)\tilde{\mathbf{X}}_k^{\mathbf{w}}\boldsymbol{\beta}_k. \tag{2.38}$$

Since $\tilde{\mathbf{X}}^{\mathbf{w}}\boldsymbol{\beta} = \tilde{\mathbf{X}}_1^{\mathbf{w}}\boldsymbol{\beta}_1 + \tilde{\mathbf{X}}_2^{\mathbf{w}}\boldsymbol{\beta}_2$ and $(\mathbf{I} - \mathbf{P}_1)\tilde{\mathbf{X}}_1^{\mathbf{w}}\boldsymbol{\beta}_1 = \mathbf{0}$, then by (2.36),

$$\| \mathbf{w}_2 \|^2 = \| (\mathbf{I} - \mathbf{P}_1)\tilde{\mathbf{X}}^{\mathbf{w}}\boldsymbol{\beta} \|^2 = \| (\mathbf{I} - \mathbf{P}_1)\tilde{\mathbf{X}}_2^{\mathbf{w}}\boldsymbol{\beta}_2 \|^2 = \| \tilde{\mathbf{X}}_2^{\mathbf{w}}\boldsymbol{\beta}_2 \|^2 = \| \mathcal{W}_2\boldsymbol{\beta}_2 \|^2,$$

where $\mathcal{W}_i = \mathcal{D}_i^{1/2} = \mathrm{diag}(W_{ij})$. The Karush-Kuhn-Tucker condition (KKT) states that a vector $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \cdots, \hat{\beta}_p)'$ is the solution of (2.6) if and only if

$$\begin{cases} \tilde{\mathbf{x}}_j'(Y - \tilde{\mathbf{X}}^{\mathbf{w}}\hat{\boldsymbol{\beta}}) = \mathrm{sgn}(\hat{\beta}_j)\lambda, & |\hat{\beta}_j| > 0; \\ |\tilde{\mathbf{x}}_j'(Y - \tilde{\mathbf{X}}^{\mathbf{w}}\hat{\boldsymbol{\beta}})| \leq \lambda, & \hat{\beta}_j = 0. \end{cases} \tag{2.39}$$

In our case, the Karush-Kuhn-Tucker condition reduces to:

$$\begin{cases} \tilde{\mathbf{x}}_j'(Y - \tilde{\mathbf{X}}^{\mathbf{w}}\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}) = \lambda, & \hat{\beta}_j > 0; \\ |\tilde{\mathbf{x}}_j'(Y - \tilde{\mathbf{X}}^{\mathbf{w}}\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}})| \leq \lambda, & \hat{\beta}_j = 0, \end{cases}$$

because $\hat{\beta}_j$'s are eigenvalues which are non-negative. Since $S_4 \in S_1$ contains variables of nonzero estimates, by the Karush-Kuhn-Tucker condition and (2.35), each component of $|\mathbf{f}_4|$ is 1. Hence, $\| \mathbf{f}_4 \|^2 = |S_4| = q_4$. Since $|S| = q$, $S_3 = S_1 \cap S$, we have $|S_3| \leq |S| = q$. So $q_1 = |S_1| = |S_3| + |S_4| \leq q + \| \mathbf{f}_4 \|^2 \Longrightarrow \| \mathbf{f}_4 \|^2 \geq q_1 - q$. Then by (2.38) and the property of $\mathbf{Q}_{41}$,

$$\begin{aligned} \| \mathbf{v}_{14} \|^2 &= \frac{\lambda^2}{p^2}\mathbf{f}_4'\mathbf{Q}_{41}(\boldsymbol{\Sigma}_{11}^{-1/2})'\boldsymbol{\Sigma}_{11}^{-1/2}\mathbf{Q}_{41}'\mathbf{f}_4 = \frac{\lambda^2}{p^2} \| \boldsymbol{\Sigma}_{11}^{-1/2}\mathbf{v} \|^2 \\ &\geq \frac{\lambda^2 \| \mathbf{v} \|^2}{n\phi_{max}(|S_1|)} = \frac{\lambda^2\mathbf{f}_4'\mathbf{Q}_{41}\mathbf{Q}_{41}'\mathbf{f}_4}{p^2\phi_{max}(|S_1|)} = \frac{\lambda^2 \| \mathbf{f}_4 \|^2}{p^2\phi_{max}(|S_1|)}, \end{aligned}$$

where the inequality follows (2.33) by setting $\mathbf{v} = \mathbf{Q}_{41}'\mathbf{f}_4$. Hence, we have

$$\| \mathbf{v}_{14} \|^2 \geq \frac{\lambda^2(q_1 - q)}{p^2\phi_{max}(|S_1|)}. \tag{2.40}$$

Next, we will establish the results in Theorem 1 in three steps.

Step 1: Establish an upper bound for $\| \mathbf{v}_{14} \|^2 + \| \mathbf{w}_2 \|^2$.

Note that $S_2 = \{j : \hat{\beta}_j = 0\}$. Hence, $\hat{\boldsymbol{\beta}}_2 = 0$ implies that $\tilde{\mathbf{X}}^{\mathbf{w}}\hat{\boldsymbol{\beta}} = \tilde{\mathbf{X}}_1^{\mathbf{w}}\hat{\boldsymbol{\beta}}_1 + \tilde{\mathbf{X}}_2^{\mathbf{w}}\hat{\boldsymbol{\beta}}_2 = \tilde{\mathbf{X}}_1^{\mathbf{w}}\hat{\boldsymbol{\beta}}_1$. From (2.35) we have, $\mathbf{f}_1\lambda = (\tilde{\mathbf{X}}_1^{\mathbf{w}})'(Y - \tilde{\mathbf{X}}^{\mathbf{w}}\hat{\boldsymbol{\beta}}) = (\tilde{\mathbf{X}}_1^{\mathbf{w}})'(Y - \tilde{\mathbf{X}}_1^{\mathbf{w}}\hat{\boldsymbol{\beta}}_1) \Longrightarrow (\tilde{\mathbf{X}}_1^{\mathbf{w}})'\tilde{\mathbf{X}}_1^{\mathbf{w}}\hat{\boldsymbol{\beta}}_1 =$

$(\tilde{\mathbf{X}}_1^{\mathbf{w}})'Y - \mathbf{f}_1\lambda$. Since $Y = \tilde{\mathbf{X}}^{\mathbf{w}}\boldsymbol{\beta} = \tilde{\mathbf{X}}_1^{\mathbf{w}}\boldsymbol{\beta}_1 + \tilde{\mathbf{X}}_2^{\mathbf{w}}\boldsymbol{\beta}_2$,

$$\begin{aligned}
(\tilde{\mathbf{X}}_1^{\mathbf{w}})'\tilde{\mathbf{X}}_1^{\mathbf{w}}\hat{\boldsymbol{\beta}}_1 &= (\tilde{\mathbf{X}}_1^{\mathbf{w}})'\tilde{\mathbf{X}}_1^{\mathbf{w}}\boldsymbol{\beta}_1 + (\tilde{\mathbf{X}}_1^{\mathbf{w}})'\tilde{\mathbf{X}}_2^{\mathbf{w}}\boldsymbol{\beta}_2 - \mathbf{f}_1\lambda, \\
\mathcal{D}_1\hat{\boldsymbol{\beta}}_1 &= \mathcal{D}_1\boldsymbol{\beta}_1 - \mathbf{f}_1\lambda, \\
\mathcal{D}_1^{-1}\mathbf{f}_1\lambda &= \boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1.
\end{aligned}$$

Now, $(\tilde{\mathbf{X}}_2^{\mathbf{w}})'(Y - \tilde{\mathbf{X}}^{\mathbf{w}}\hat{\boldsymbol{\beta}}) = (\tilde{\mathbf{X}}_2^{\mathbf{w}})'\mathbf{y} - (\tilde{\mathbf{X}}_2^{\mathbf{w}})'\tilde{\mathbf{X}}^{\mathbf{w}}\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}_2^{\mathbf{w}})'\tilde{\mathbf{X}}_1^{\mathbf{w}}\boldsymbol{\beta}_1 + (\tilde{\mathbf{X}}_2^{\mathbf{w}})'\tilde{\mathbf{X}}_2^{\mathbf{w}}\boldsymbol{\beta}_2 - (\tilde{\mathbf{X}}_2^{\mathbf{w}})'\tilde{\mathbf{X}}_1^{\mathbf{w}}\hat{\boldsymbol{\beta}}_1 = \mathcal{D}_2\boldsymbol{\beta}_2$. By the Karush-Kuhn-Tucker condition, $|(\tilde{\mathbf{X}}_2^{\mathbf{w}})'(\mathbf{y} - \tilde{\mathbf{X}}^{\mathbf{w}}\hat{\boldsymbol{\beta}})|$ is bounded above componentwise by λ, then $|\mathcal{D}_2\boldsymbol{\beta}_2|$ is bounded above componentwise by λ. Moreover, since $\mathcal{D}_2\boldsymbol{\beta}_2$ is positive in our case, $\mathcal{D}_2\boldsymbol{\beta}_2$ is bounded above componentwise by λ. In other words, $W_i^2\beta_{2i} \leq \lambda$ for $i = 1, \cdots, q_2$. Therefore,

$$\| \mathbf{w}_2 \|^2 = \| \mathcal{W}_2\boldsymbol{\beta}_2 \|^2 = \boldsymbol{\beta}_2'\mathcal{W}_2'\mathcal{W}_2\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2'\mathcal{D}_2\boldsymbol{\beta}_2 = \sum_{i=1}^{q_2} \beta_{2i}W_i^2\beta_{2i} \leq \sum_{i=1}^{q_2} \beta_{2i}\lambda = \| \boldsymbol{\beta}_2 \|_1 \lambda.$$

Next, we have,

$$\begin{aligned}
\mathbf{v}_{14}'(\mathbf{v}_{13} + \mathbf{v}_{14}) &= \frac{\lambda}{\sqrt{p^2}}\mathbf{f}_4'\mathbf{Q}_{41}(\boldsymbol{\Sigma}_{11}^{-1/2})' \left(\frac{\lambda}{\sqrt{p^2}}\boldsymbol{\Sigma}_{11}^{-1/2}\mathbf{Q}_{31}'\mathbf{f}_3 + \frac{\lambda}{\sqrt{p^2}}\boldsymbol{\Sigma}_{11}^{-1/2}\mathbf{Q}_{41}'\mathbf{f}_4 \right) \\
&= \frac{\lambda}{\sqrt{p^2}}\mathbf{f}_4'\mathbf{Q}_{41}(\boldsymbol{\Sigma}_{11}^{-1/2})' \left[\frac{\lambda}{\sqrt{p^2}}\boldsymbol{\Sigma}_{11}^{-1/2} (\mathbf{Q}_{31}'\mathbf{f}_3 + \mathbf{Q}_{41}'\mathbf{f}_4) \right] \\
&= \frac{\lambda}{\sqrt{p^2}}\mathbf{f}_4'\mathbf{Q}_{41}(\boldsymbol{\Sigma}_{11}^{-1/2})' \left(\frac{\lambda}{\sqrt{p^2}}\boldsymbol{\Sigma}_{11}^{-1/2}\mathbf{f}_1 \right) \\
&= \frac{\lambda^2}{p^2}\mathbf{f}_4'\mathbf{Q}_{41}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{f}_1 = \lambda^2\mathbf{f}_4'\mathbf{Q}_{41}\mathcal{D}_1^{-1}\mathbf{f}_1 \\
&= \lambda\mathbf{f}_4'\mathbf{Q}_{41}(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1) = \lambda\mathbf{f}_4'(\boldsymbol{\beta}_4 - \hat{\boldsymbol{\beta}}_4),
\end{aligned}$$

where the second to the last equality holds by $\mathcal{D}_1^{-1}\mathbf{f}_1\lambda = \boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1$. In our case, $\mathbf{f}_4$ is a vector of 1's and $\hat{\boldsymbol{\beta}}_4 \geq 0$ componentwise. So $\mathbf{f}_4'\hat{\boldsymbol{\beta}}_4 \geq 0$ implies that $\mathbf{v}_{14}'(\mathbf{v}_{13} + \mathbf{v}_{14}) \leq \lambda\mathbf{f}_4'\boldsymbol{\beta}_4$. Combining

$\mathbf{v}'_{14}(\mathbf{v}_{13} + \mathbf{v}_{14})$ and $\| \mathbf{w}_2 \|^2$, we have

$$\| \mathbf{v}_{14} \|^2 + \mathbf{v}'_{14}\mathbf{v}_{13} + \| \mathbf{w}_2 \|^2 \leq \| \boldsymbol{\beta}_2 \|_1 \lambda + \mathbf{f}'_4 \boldsymbol{\beta}_4 \lambda.$$

By the definitions of sets A_i in Table 2.4, we have

$$\| \boldsymbol{\beta}_2 \|_1 + \mathbf{f}'_4 \boldsymbol{\beta}_4 = \| \boldsymbol{\beta}_2 \|_1 + \| \boldsymbol{\beta}_4 \|_1 = \| \boldsymbol{\beta}_5 \|_1 + \| \boldsymbol{\beta}_0 \|_1 \leq \| \boldsymbol{\beta}_5 \|_1 .$$

Hence, $\| \mathbf{v}_{14} \|^2 + \| \mathbf{w}_2 \|^2 \leq \| \boldsymbol{\beta}_5 \|_1 \lambda + (-\mathbf{v}_{14})'\mathbf{v}_{13} \leq \| \boldsymbol{\beta}_5 \|_1 \lambda + \| \mathbf{v}_{14} \| \| \mathbf{v}_{13} \|$, where the last inequality is obtained by the Cauchy-Schwarz inequality.

Again by the Karush-Kuhn-Tucker condition, since $S_3 \in S_1$ contains variables of nonzero estimates, each component of $|\mathbf{f}_3|$ is 1. So $\| \mathbf{f}_3 \|^2 = |S_3| = q_3$. By the property of $\mathbf{Q}_{31}$, we have

$$\begin{aligned} \| \mathbf{v}_{13} \|^2 &= \frac{\lambda^2}{p^2} \mathbf{f}'_3 \mathbf{Q}_{31} \boldsymbol{\Sigma}_{11}^{-1} \mathbf{Q}'_{31} \mathbf{f}_3 = \frac{\lambda^2}{p^2} \| \boldsymbol{\Sigma}_{11}^{-1/2} \mathbf{v} \|^2 \\ &\leq \frac{\lambda^2 \| \mathbf{v} \|^2}{p^2 \phi_{min}(|S_1|)} = \frac{\lambda^2 \mathbf{f}'_3 \mathbf{Q}_{31} \mathbf{Q}'_{31} \mathbf{f}_3}{p^2 \phi_{min}(|S_1|)} = \frac{\lambda^2 \| \mathbf{f}_3 \|^2}{p^2 \phi_{min}(|S_1|)} \\ &= \frac{\lambda^2 |S_3|}{p^2 \phi_{min}(|S_1|)}, \end{aligned}$$

where the inequality follows (2.33) by setting $\mathbf{v} = \mathbf{Q}'_{31}\mathbf{s}_3$. Therefore, we have

$$\| \mathbf{v}_{14} \|^2 + \| \mathbf{w}_2 \|^2 \leq \| \boldsymbol{\beta}_5 \|_1 \lambda + \| \mathbf{v}_{14} \| \left(\frac{\lambda^2 |S_3|}{p^2 \phi_{min}(S_1)} \right)^{1/2} . \tag{2.41}$$

Define,

$$B_1 = \left(\frac{q\lambda^2}{p^2 \phi^*} \right)^{1/2}, \qquad B_2 = \left(\frac{q\lambda^2}{p^2 \phi_*} \right)^{1/2}, \qquad B_2^2 = CB_1^2,$$

where $\phi_* = \min_{m \leq p} \phi_{min}(m)$ and $\phi^* = \max_{m \leq p} \phi_{max}(m)$.

Step 2: Establish (2.12).

Assume S_1 contains all labels j for nonzero β_j:

$$S_1 = \{j : \hat{\beta}_j \neq 0 \text{ or } j \in S\}. \tag{2.42}$$

In this case, $S_5 = \varnothing$. So $\parallel \boldsymbol{\beta}_5 \parallel_1 = 0$, $S_3 = S$, and thus $|S_3| = q \leq q_1$. Because $\left(\frac{\lambda^2|S_3|}{p^2\phi_{min}(S_1)}\right)^{1/2} \leq B_2$ and $\parallel \mathbf{w}_2 \parallel^2 \geq 0$, together with (2.41), we have that $\parallel \mathbf{v}_{14} \parallel^2 + \parallel \mathbf{w}_2 \parallel^2 \leq \parallel \mathbf{v}_{14} \parallel B_2$, which implies $\parallel \mathbf{v}_{14} \parallel^2 \leq \parallel \mathbf{v}_{14} \parallel B_2$, and $\parallel \mathbf{v}_{14} \parallel \leq B_2$. Combining with (2.40), we have,

$$\begin{aligned}
\frac{(q_1 - q)\lambda^2}{p^2\phi^*} \leq \frac{\lambda^2(q_1 - q)}{p^2\phi_{max}(|S_1|)} &\leq \parallel \mathbf{v}_{14} \parallel^2 \leq B_2^2, \\
(q_1 - q) &\leq \frac{q\phi^*}{\phi_*}, \\
q_1 &\leq \frac{\phi^*}{\phi_*}q + q = Cq + q = r_1 q,
\end{aligned}$$

where r_1 is defined in (2.11). Under assumption (2.42), S_1 is taken as the largest possible set which contains $\tilde{q}$ elements. In general, S_1 doesn't necessarily select all the variables with nonzero coefficients in the true model. Hence,

$$\hat{q} = q_1 = |S_1| \leq \tilde{q} = \#\{j : \hat{\beta}_j \neq 0 \text{ or } j \in S\} \leq r_1 q,$$

which is (2.12).

Step 3: Establish (2.13) and (2.14).

By Lemma 2, we have that

$$\parallel \boldsymbol{\beta}_5 \parallel_1^2 \leq \frac{\parallel \tilde{\mathbf{X}}_5^{\mathbf{w}} \boldsymbol{\beta}_5 \parallel^2 |S_5|}{p^2\phi_{min}(|S_5|)} \leq \frac{\parallel \tilde{\mathbf{X}}_5^{\mathbf{w}} \boldsymbol{\beta}_5 \parallel^2 q}{p^2\phi_*},$$

because by Table 2.4 $|S_3|+|S_5|=|S|=q \Longrightarrow |S_5|\leq q$ and $|S_3|\leq q$. Note that $S_5\subseteq S_2$,

$$\parallel \tilde{\mathbf{X}}_5^{\mathbf{w}}\boldsymbol{\beta}_5 \parallel^2 \leq \parallel \tilde{\mathbf{X}}_2^{\mathbf{w}}\boldsymbol{\beta}_2 \parallel^2 = \parallel \mathbf{w}_2 \parallel^2 .$$

Combining the above two inequalities,

$$\parallel \boldsymbol{\beta}_5 \parallel_1 \lambda \leq \left(\frac{\parallel \tilde{\mathbf{X}}_5^{\mathbf{w}}\boldsymbol{\beta}_5 \parallel^2 q\lambda}{p^2\phi_*}\right)^{\frac{1}{2}} \leq \left(\frac{\parallel \mathbf{w}_2 \parallel^2 q\lambda}{p^2\phi_*}\right)^{\frac{1}{2}} \leq \parallel \mathbf{w}_2 \parallel B_2.$$

By the Cauchy-Schwarz inequality, $\parallel \mathbf{v}_{14} \parallel B_2 \leq \parallel \mathbf{v}_{14} \parallel^2 + \frac{B_2^2}{4}$. So based on (2.41) we have

$$\parallel \mathbf{v}_{14} \parallel^2 + \parallel \mathbf{w}_2 \parallel^2 \leq \parallel \mathbf{v}_{14} \parallel^2 + \frac{B_2^2}{4} + \parallel \mathbf{w}_2 \parallel B_2,$$
$$\parallel \mathbf{w}_2 \parallel^2 \leq \frac{B_2^2}{4} + \parallel \mathbf{w}_2 \parallel B_2.$$

One can easily show that $x^2 \leq c+2bx$ implies $x^2 \leq (b+\sqrt{b^2+c})^2 \leq 2c+4b^2$. Setting $x=\parallel \mathbf{w}_2 \parallel, c=\frac{B_2^2}{4}, 2b=B_2$, we obtain the result in (2.13),

$$\parallel \mathbf{w}_2 \parallel^2 \leq \frac{B_2^2}{2}+B_2^2 = \frac{3B_2^2}{2} = \frac{3C}{2}B_1^2 = r_2\left(\frac{q\lambda^2}{p^2\phi^*}\right), \tag{2.43}$$

where r_2 is defined in (2.11).

By the Sparse Riesz Condition, $\phi_{min}(|S_5|) \leq \frac{\|\tilde{\mathbf{X}}_5^{\mathbf{w}}\boldsymbol{\beta}_5\|^2}{p^2\|\boldsymbol{\beta}_5\|^2} \Longrightarrow \parallel \boldsymbol{\beta}_5 \parallel^2 \leq \frac{\|\tilde{\mathbf{X}}_5^{\mathbf{w}}\boldsymbol{\beta}_5\|^2}{p^2\phi_{min}(|S_5|)}$. Since $\parallel \tilde{\mathbf{X}}_5^{\mathbf{w}}\boldsymbol{\beta}_5 \parallel^2 \leq \parallel \mathbf{w}_2 \parallel^2$, we have $\parallel \boldsymbol{\beta}_5 \parallel^2 \leq \frac{\|\mathbf{w}_2\|^2}{p^2\phi_*}$, which directly gives the result in (2.14) after combining with (2.43): $\parallel \boldsymbol{\beta}_5 \parallel^2 \leq r_2\left(\frac{q\lambda^2}{p^4\phi^*\phi_*}\right)$. □

Appendix C

Proof of Theorem 2. Since we have shown that in model (2.4), the covariates $\mathbf{Z}$ satisfy the Sparse Riesz Condition, it follows directly from the result (2.14) of Theorem 1,

$$\sum_{j \in S} |\beta_j|^2 \mathbb{1}_{\{\hat{\beta}_j = 0\}} \leq r_2 \frac{q\lambda^2}{\phi^* \phi_* p^4} = 1.5 \frac{\phi^*}{\phi_*} \frac{q\lambda^2}{\phi^* \phi_* p^4} = \frac{1.5q\lambda^2}{\phi_*^2 p^4} \leq \frac{1.5q\lambda^2}{u^4},$$

because $0 < \frac{u^2}{p^2} \leq \phi_* \leq \phi^* \leq \frac{1}{p^2} < \infty$. Hence, $\forall j \in \hat{S}^\lambda$, $\beta_j > \sqrt{1.5q}\lambda/u^2$. By definition of S_{small}, we can conclude that $(S \setminus S_{small}) \subseteq \hat{S}^\lambda$, which verifies (2.16).

By Lemma 1 of Meinshausen and Bühlmann (2006), a variable $j \notin S$ is in the selected set $\hat{S}^\lambda$ only if

$$|\mathbf{z}_j'(Y - \tilde{\mathbf{X}}^{\mathbf{w}} \hat{\boldsymbol{\beta}}^{-j})| \geq \lambda, \tag{2.44}$$

where $\hat{\boldsymbol{\beta}}^{-j}$ is the solution to (2.6) under the constraint of $\hat{\beta}_j^{-j} = 0$. We can rewrite the left-hand side as

$$|\tilde{\mathbf{x}}_j' Y - \tilde{\mathbf{x}}_j' \tilde{\mathbf{X}}^{\mathbf{w}} \hat{\boldsymbol{\beta}}^{-j}| = |\tilde{\mathbf{x}}_j' \tilde{\mathbf{X}}^{\mathbf{w}} \boldsymbol{\beta} - \tilde{\mathbf{x}}_j' \tilde{\mathbf{X}}^{\mathbf{w}} \hat{\boldsymbol{\beta}}^{-j}| = |\tilde{\mathbf{x}}_j' \tilde{\mathbf{x}}_j \beta_j - \tilde{\mathbf{x}}_j' \tilde{\mathbf{x}}_j \hat{\beta}_j^{-j}| = 0.$$

The second equality is due to the orthogonal property of $\tilde{\mathbf{X}}^{\mathbf{w}}$ in (2.5). The last equality is because $\beta_j = 0$ and $\hat{\beta}_j^{-j} = 0$. Hence, the condition (2.44) will never be satisfied because $\lambda > 0$ which means that $\hat{S}^\lambda$ contains only variables in S. This completes the proof. □

Appendix D

In this section, we explain how a matrix in the Grassmann Manifold can be written as an eigen-decomposition solution, the link between the Grassmann Manifold and a dimension reduction matrix, and how to form an equivalent form for the purpose of obtaining a sparse estimate.

Grassmann Manifold

Suppose that a $p \times k$ matrix V is in the Grassmann Manifold with rank k.

First, we can extend it to a nonsingular $p \times p$ matrix $V^* = (V, V^{\perp})$.

Second, applying singular value decomposition on V^*, we have $V^* = L\Lambda R'$ where the columns of $p \times p$ matrix L and $p \times p$ matrix R are corresponding left and right eigenvectors of V^*, and $\Lambda = (\Lambda_k, \Lambda_{p-k})$ is a diagonal matrix with non-zero singular values of V^* being its diagonal elements. Let

$$G = L\Lambda^{-2}L', \tag{2.45}$$

we have $V^{*'}GV^* = R\Lambda' L' L\Lambda^{-2}L'L\Lambda R' = \mathbf{I}_p$.

Let $\mathbf{M} = GV^*DV^{*'}G$, where G is found by (2.45) and D can be any diagonal matrix with diagonal terms being $\rho_1 >, \cdots, > \rho_p > 0$, then V^*, G, and $\mathbf{M}$ will satisfy the basic eigenvalue decomposition as

$$\mathbf{M}V^* = GV^*D. \tag{2.46}$$

Hence, for any $p \times k$ matrix V in the Grassmann Manifold with rank k, the columns of V are the eigenvectors of a symmetric and positive definite matrix $\mathbf{M}$, whose corresponding eigenvalues are $\rho_1, \cdots, \rho_k$, as long as $\rho_1 > \cdots > \rho_k > 0$. Typically, we would choose D so that $\rho_1 > \cdots, > \rho_p > 0$, and all the eigenvalues are bounded below by 0 and above by ∞.

Link between the Grassmann Manifold and a dimension reduction matrix

We assume that the CS is sparse, which means that only some variables are related to the response. Note that sparsity is not generally transformed from one scale to another scale. That is, if a model is sparse in the $\mathbf{X}$-scale, it does not mean it is sparse in the $\mathbf{Z}$-scale.

Thus our discussion will focus on the original $\mathbf{X}$-scale. Suppose that the method specific kernel matrix $\mathbf{M}$ is obtained based on the original predictors $\mathbf{X}$. Basis directions are found by conducting the generalized eigenvalue problem of the form

$$\mathbf{M}V^* = GV^*D, \tag{2.47}$$

where columns of $V^* = (v_1, \cdots, v_p)$ are eigenvectors of $\mathbf{M}$ satisfying $V^{*'}GV^* = \mathbf{I}_p$ and D is a diagonal matrix with eigenvalues $\rho_1, \rho_2, \cdots, \rho_p$ of $\mathbf{M}$ in descending order. If the structural dimension of the CS is k, then the first k orthogonal eigenvectors, say, V, form an estimate of the central subspace.

Thus, in the Grassmann Manifold, we construct $\mathbf{M}$, D and G from V, while in dimension reduction, we have $\mathbf{M}$ and G to deduce V and D.

Our theoretical result in Section 2.2.2 in the paper requires that all eigenvalues are bounded below from 0 and above from ∞ to satisfy the Sparse Riesz Condition. This is not satisfied when we have a $p \times p$ dimension reduction matrix with k nonzero eigenvalues with $k < p$.

To fix this, from (2.47), we have $\mathbf{M} = \mathbf{M}V^*V^{*-1} = GV^*DV^{*'}G$. For some positive constant δ, let $\mathbf{M}_\delta = (\mathbf{M} + \delta G) = GV(D + \delta\mathbf{I}_p)V'G$. We will have similar eigenvalue decomposition on $\mathbf{M}_\delta$ as

$$\mathbf{M}_\delta V^* = GV^*D_\delta, \tag{2.48}$$

where D_δ is a diagonal matrix with eigenvalues $\rho_1+\delta, \rho_2+\delta, \cdots, \rho_p+\delta$. Since the eigenvectors of $\mathbf{M}_\delta$ are same as these of $\mathbf{M}$, we can work on $\mathbf{M}_\delta$ to estimate the basis directions of the central subspace. After some algebra, it requires the ratios of maximum and minimum eigenvalues for the matrix $\mathbf{M}_\delta$ to be bounded below from 0 and above from ∞. Let $m_1 \geq$

$\cdots \geq m_p$ be the eigenvalues of $\mathbf{M}_\delta$, we have

$$m_1 = \rho_1 + \delta, \qquad m_p = \rho_p + \delta.$$

Therefore, by choosing $\delta > 0$, the eigenvalues of $\mathbf{M}_\delta$ are bounded below from 0 and above from ∞ so that the Sparse Riesz Condition will be satisfied. Under this condition, Theorem 3 applies to the dimension reduction matrix $\mathbf{M}_\delta$.

Appendix E

Proof of Theorem 3. Since the optimization problem (2.20) was developed based on the generalized eigenvalue problem (2.48), we again have an "error-free" model in the population: $\tilde{Y} = \tilde{\mathbf{X}}\boldsymbol{\beta}$. If the Sparse Riesz Condition is satisfied, following (5.8) in Zhang and Huang (2008), we have

$$\| \mathbf{v}_{14} \|^2 + \| \mathbf{w}_2 \|^2 \leq \| \boldsymbol{\beta}_5 \|_1 \lambda + \| \mathbf{v}_{14} \| \left(\frac{\lambda^2 |S_3|}{p\phi^Z_{min}(S_1)} \right)^{1/2}. \tag{2.49}$$

where $\mathbf{v}_{14}$, $\mathbf{w}_2$, S_1, and S_2 are defined in the same ways as in the proof of Theorem 1. Following the same steps as in the proof of Theorem 1, we are able to obtain the following two upper bounds:

$$\hat{q}(\lambda) \leq \tilde{q} = \#\{j : \hat{\beta}^\lambda_j \neq 0 \text{ or } j \in S\} \leq r_1 q, \tag{2.50}$$

and

$$\sum_{j \in S} |\beta_j|^2 \mathbb{1}_{\{\hat{\beta}^\lambda_j = 0\}} \leq r_2 \frac{q\lambda^2}{\phi^* \phi_* p^2}, \tag{2.51}$$

where r_1, r_2, ϕ^*, and ϕ_* are defined in (2.11).

The rest of the proof follows mostly from the steps of Meinshausen and Bühlmann (2010) in the proof of their Theorem 2.

Lemma 3. Define C by $(1+C)q+1 = \bar{C}q^2$ and assume $q \geq 3$. Let weights W_k be generated randomly in $[u,1]$ as in (2.20), and let $\tilde{X}_k^{\mathbf{w}} = \tilde{X}^{\mathbf{w}} W_k$ for $k = 1, \cdots, p$ be the corresponding rescaled predictor variables. For $u^2 = \nu\phi_{min}(\bar{C}q^2)/\bar{C}q^2$ with $\nu > 0$, it holds under assumption (2.21) for all realizations W_k that

$$\frac{\phi^w_{max}(\bar{C}q^2)}{\phi^w_{min}(\bar{C}q^2)} \leq \frac{3C}{\kappa\sqrt{\nu}}. \tag{2.52}$$

Proof. We can follow exactly the same steps as in the proof of the Lemma 3 in Meinshausen and Bühlmann (2010). The only remark we need to make is that C in our notation is their $\bar{C}$ while our $\bar{C}$ is their C. Since the steps are similar, we omit the details.

Lemma 4. Let $\hat{S}^{\lambda,w}$ be the set $\{k : \hat{\beta}_k^{\lambda,w}\}$ of selected variables of the randomized lasso with $u \in (0,1]$ and randomly sampled weights $\mathbf{w}$. Suppose that $u^2 \geq (3/\kappa)^2\phi_{min}(\bar{C}q^2)/\bar{C}q^2$, we can show that

$$|\hat{S}^{\lambda,w} \cup S| \leq \bar{C}q^2 \text{ and } (S \setminus S_{small}) \subseteq \hat{S}^{\lambda,w}, \tag{2.53}$$

where $S_{small} = \{k : \beta_k \leq \sqrt{1.5}\bar{C}q^{3/2}\lambda\}$.

Proof. The proof of this lemma follows from Theorem 1 and Lemma 4 of Meinshausen and Bühlmann (2010). With Remark 2 in Zhang and Huang (2008), the equivalent condition of (2.21) requires the existence of some $C > 0$ such that

$$\frac{\phi_{max}((1+C)q+1)}{\phi_{min}((1+C)q+1)} < C,$$

where C is defined in (2.11). Hence, for all realizations W_i, if $u^2 \geq (3/\kappa)^2\phi_{min}(\bar{C}q^2)/\bar{C}q^2$, by Lemma 3, $\frac{\phi^w_{max}(\bar{C}q^2)}{\phi^w_{min}(\bar{C}q^2)}$ is bounded. Therefore, (2.50) and (2.51) hold which give us

$$|\hat{S}^{\lambda,w} \cup S| \leq (1+C)q \leq \bar{C}q^2,$$

and

$$\sum_{j \in S} |\beta_j|^2 \mathbb{1}_{\{\hat{\beta}^\lambda_j = 0\}} \leq (1.5C^2)q\lambda^2 \leq (\sqrt{1.5}\bar{C}q^{3/2}\lambda)^2, \tag{2.54}$$

where the first inequality uses the fact that $1/\phi^*\phi_* \leq C$ and the second inequality follows from $C \leq \bar{C}q$. Accordingly, (2.54) is equivalent to the second part of (2.53).

Lemma 5 Let p_w be the probability of choosing weight u for each variable and $1 - p_w$ the probability of choosing weight 1. Define $\tilde{p} = p_w(1-p_w)^{\bar{C}q^2}$ and let $\hat{\Pi}^\lambda_k$ be the probability of variable k being in the selected subset $\hat{S}^{\lambda,w}$ with respect to random sampling of the weights w. Under assumptions of Theorem 3, for any $\lambda \geq \inf\{\lambda : r_1 q + 1 \leq p\}$,

$$\max_{k \in N}(\hat{\Pi}^\lambda_k) < 1 - \tilde{p}, \tag{2.55}$$

$$\min_{k \in S \backslash S_{small}}(\hat{\Pi}^\lambda_k) \geq 1 - \tilde{p}, \tag{2.56}$$

where $S_{small} = \{k : \beta_k \leq \sqrt{1.5}\bar{C}q^{3/2}\lambda\}$.

Proof. Following Meinshausen and Bühlmann (2006), a variable $j \notin S$ is in the selected set $\hat{S}^{\lambda,w}$ only if

$$|\tilde{\mathbf{x}}'_j(Y - \tilde{\mathbf{X}}^w \hat{\boldsymbol{\beta}}^{-j})| \geq \lambda, \tag{2.57}$$

where $\hat{\boldsymbol{\beta}}^{-j}$ is the solution to (2.20) under the constraint of $\hat{\beta}^{-j}_j = 0$. Using Lemma 5 of Meinshausen and Bühlmann (2010) and Lemma 4 above, we can show that the left-hand side of (2.57) is bounded by $\| \, ((\tilde{\mathbf{X}}_B^{\mathbf{w}})'\tilde{\mathbf{X}}_B^{\mathbf{w}})^{-1}(\tilde{\mathbf{X}}_B^{\mathbf{w}})'\tilde{\mathbf{X}}_j^{\mathbf{w}} \, \|_1 \, \lambda \le 2^{-1/4}\lambda < \lambda$ with probability greater than or equal to $p_w(1-p_w)^{\bar{C}q^2}$, where set $B = \hat{S}^{\lambda,w} \cup S$ and the first inequality is based on Lemma 5 of Meinshausen and Bühlmann (2010). This leads to the result (2.55). The consequence of Lemma 4 directly yields (2.56). Since our Lemma 5 is equivalent to Theorem 3, the proof of Theorem 3 is complete. □

Appendix F

In Section 2.2.2 in the paper, we used $\delta > 0$ to have $\mathbf{M}_\delta = (\mathbf{M} + \delta G)$ for GMSE and SGMSE in order to satisfy the Sparse Riesz Condition. In this section, we investigate the choice of δ in GMSE and SGMSE. Our empirical evidences show that the choices of the positive constant δ have little effect on the final estimates. We ran simulations on the three models in Section 2.3.2.1 in the paper, with different choices of $\delta = 0.001, 0.01, 0.1, 0.5$. For the same model and method, varying δ does not greatly change the results. In addition, for all δ values, SGMSE is improved over GMSE. It seems that a smaller value of δ is preferable because it results in a lower false positive rate. Hence, a rule of thumb for appropriate δ to use in GMSE is between 0.001 to 0.01.

		$\delta = 0.001$		$\delta = 0.01$		$\delta = 0.1$		$\delta = 0.5$	
		TPR	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**
SIR	GMSE SIR	1.000	0.073	1.000	0.023	1.000	0.018	1.000	0.025
	SGMSE SIR	1.000	0.003	1.000	0.000	1.000	0.000	1.000	0.003
PHD	GMSE PHD	1.000	0.268	1.000	0.274	1.000	0.380	1.000	0.350
	SGMSE PHD	1.000	0.085	1.000	0.095	1.000	0.140	1.000	0.154
SAVE	GMSE SAVE	1.000	0.283	1.000	0.263	1.000	0.304	1.000	0.451
	SGMSE SAVE	1.000	0.063	1.000	0.058	1.000	0.069	1.000	0.151

Table 2.5: Additional simulation results for Section 2.3.2.1 I

Appendix G

In this appendix, we compare GMSE to the sparse sufficient dimension reduction (SSDR) method (Li, 2007), and SGMSE to stable SSDR (SSSDR).

Li's SSDR starts with an equivalent formulation of eigen-decomposition as

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \sum_{i=1}^{p} \parallel G^{-1}m_i - \boldsymbol{\alpha}\boldsymbol{\beta}^T m_i \parallel_G^2 + \lambda_2 tr(\boldsymbol{\beta}^T G \boldsymbol{\beta}) + \sum_{j=1}^{k} \lambda_{1j} \sum_{h=1}^{p} |\beta_{jh}|, \tag{2.58}$$

subject to $\boldsymbol{\alpha}^T G \boldsymbol{\alpha} = \mathrm{I}$, where the norm is the inner product with with respect to G. In (2.58), G takes the form of the covariance matrix Σ_x of $\mathbf{X}$, the values of m_i are columns of the square root of the method-specific dimension reduction matrix $\mathbf{M}$ and $\boldsymbol{\beta}$ is a $p \times k$ matrix of which the columns are the basis directions of the central subspace. The λ_2 and λ_{1j}'s are the tuning parameters corresponding to the L^1 and L^2 penalties. Then, Li (2007) showed that the optimization problem (2.58) can be solved in an alternative way by solving k independent LASSO problems for a given $\boldsymbol{\alpha}$ as:

$$\hat{\boldsymbol{\beta}}_j = \min_{\boldsymbol{\beta}_j} \left\{ \boldsymbol{\beta}_j^T (\mathbf{M} + \lambda_2 G) \boldsymbol{\beta}_j - 2\alpha_j^T M \boldsymbol{\beta}_j + \lambda_{1j} \sum_{h=1}^{p} |\beta_{jh}| \right\}, \tag{2.59}$$

subject to $\boldsymbol{\alpha}^T G \boldsymbol{\alpha} = \mathrm{I}$. For given $\boldsymbol{\beta}_j$'s, solving $\boldsymbol{\alpha}$ is just a usual OLS problem. Li (2007) also showed that (2.59) can be transformed into an equivalent problem as

$$\hat{\boldsymbol{\beta}}_j = \min_{\boldsymbol{\beta}_j} \left\{ \parallel u^* - m^* \boldsymbol{\beta}_j \parallel^2 + \lambda_{1j} \sum_{h=1}^{p} |\beta_{jh}| \right\}, \tag{2.60}$$

where,

$$m^* = \begin{pmatrix} \mathbf{M}^{1/2} \\ \sqrt{\lambda_2} G^{1/2} \end{pmatrix}, \qquad u^* = \begin{pmatrix} \mathbf{M}^{1/2}\alpha_j \\ 0 \end{pmatrix}.$$

In the above SSDR method, by introducing subsampling and random weight we also develop a Stable SSDR, which we call SSSDR.

However, the introduction of λ_2 in SSDR is only for the uniqueness of the eigenvectors. For this reason, Li's algorithm gives an invariant result for any $\lambda_2 > 0$. However, under this formulation, the Sparse Riesz Condition is not always satisfied, even though λ_2 is a nuisance parameter. Thus, we are not able to prove the theoretical result, even if we believe the result holds. In addition, with an information criterion to select λ_2, it slows down the computing speed. Nevertheless, the table below shows that in our simulations the two approaches have very comparable results.

	Original		SSDR		GMSE		SSSDR		SGMSE	
	TPR	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**
SIR	1.000	1.000	1.000	0.012	1.000	0.022	1.000	0.001	1.000	0.002
PHD	1.000	1.000	1.000	0.171	1.000	0.249	1.000	0.044	1.000	0.058
SAVE	1.000	1.000	1.000	0.264	1.000	0.191	1.000	0.115	1.000	0.005

Table 2.6: Additional simulation results for Section 2.3.2.1 II

Appendix H

In this section, we include two simulation studies: one simulation shows the sensitivity of sliced inverse methods to the choices of H; another simulation illustrates the stability of the results using the ensemble method proposed in Section 2.2.3.2 in the paper. We used model (2.25) for SIR and model (2.29) for SAVE as in Section 2.3 in the paper.

Figure 2.2 shows that for each fixed $H = 5, 10, 15, 20$, the results for SIR do vary but not as much as these of SAVE (left column); stable procedures show significant improvement. While results for SIR do vary, but results vary more for SAVE (right column). We now use these four different numbers of slices to develop one aggregated dimension reduction matrix as proposed in Section 2.2.3.2. Figure 2.3 shows that the ensemble method gives better and more stable results (left column), which are further improved by our newly developed stable procedure (right column).

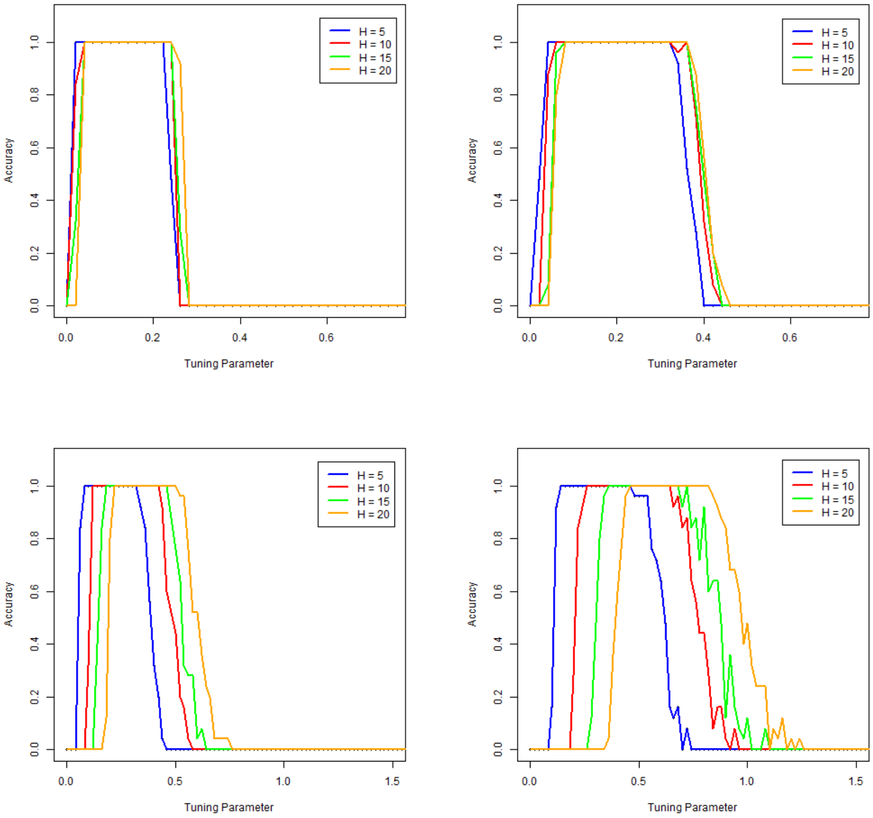

Figure 2.2: Sensitivity of sliced inverse methods to the choices of H

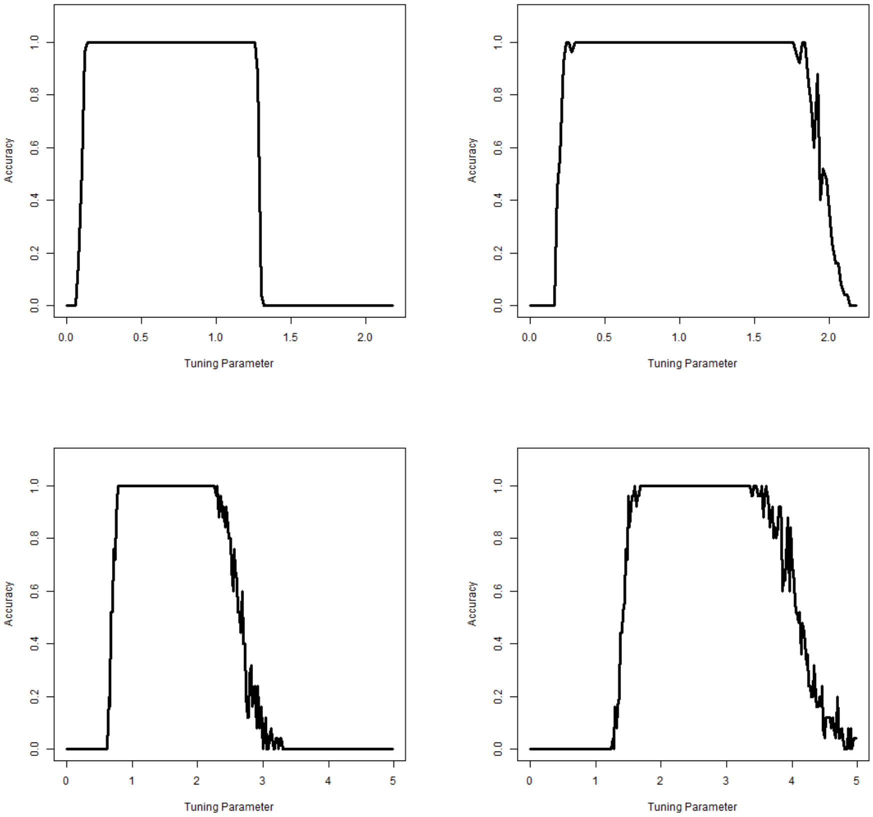

Figure 2.3: Stabilizing the choice of H for inverse methods

Chapter 3

Pseudo sufficient dimension reduction and sufficient variable selection

3.1 Introduction

3.1.1 Sufficient dimension reduction and sufficient variable selection

Sufficient dimension reduction (SDR) has been one of the popular topics in statistics over the past two decades or so. Suppose that Y is a response variable and $\mathbf{X}$ is a $p \times 1$ predictor vector. The basic idea of SDR is based on the SDR subspace. Let $\mathbf{B}$ be a $p \times d$ matrix, a dimension reduction subspace is a space spanned by the columns of $\mathbf{B}$ such that $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{B}^T\mathbf{X}$, which means that for a given value of $\mathbf{B}^T\mathbf{X}$, the distribution of Y is independent of $\mathbf{X}$. Hence, $\mathbf{B}^T\mathbf{X}$ contains all the regression information of $Y|\mathbf{X}$. When the intersection of all dimension-reduction subspaces itself is a dimension-reduction subspace, it is called the *central subspace* (CS) and denoted by $S_{Y|\mathbf{X}}$ (Li, 1991; Cook, 1994, 1996). The number of columns for $\mathbf{B}$ is called the structural dimension or dimensionality of the CS. The existence of the CS can be guaranteed under mild conditions (Cook, 1998; Yin et al., 2008). The CS is designed to give a complete picture of the dependence of Y on $\mathbf{X}$. Subspaces with specific interests are also introduced by Cook and Li (2002) for the mean function, Yin and Cook (2002) for the kth moment function, Zhu and Zhu (2009) for the variance function, and by Luo et al. (2014) with a general function. Rich literature is available on the inferences of the CS (Li, 1991, 1992; Cook and Weisberg, 1991; Cook, 1998).

Variable selection is another important problem in the literature of SDR. Yin and Hilafu (2015) discussed the differences between the model selection and the variable selection. They also formally defined the central variable selection space, $S^V_{Y|\mathbf{X}}$, to be the column space of $\boldsymbol{\alpha}$ where the columns of $\boldsymbol{\alpha}$ consist of unit vectors $\mathbf{e}_j$ with j^{th} element 1 and 0 otherwise, such that $Y \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\alpha}^T\mathbf{X}$. Penalized approach is a popular way to achieve sufficient variable selection. Many existing methods (Ni et al., 2005; Li, 2007; Zhou and He, 2008; Wang and Yin, 2008; Wu and Yin, 2015b; Chen et al., 2010) impose a penalization to the estimated $S_{Y|\mathbf{X}}$ in order to obtain $S^V_{Y|\mathbf{X}}$.

Most of the SDR methods require the inverse of the sample covariance matrix $\hat{\Sigma}_{\mathbf{x}}$ of the predictors. However, in many applications, $\hat{\Sigma}_x$ may be singular, hence, not invertible, due to high correlations such as in Chemometrics, or with a large p small n data such as in microarray studies. To handle the singularity of the covariance matrix, Chiaromonte and Martinelli (2002) and Li and Li (2004) used principal component analysis (PCA) on $\hat{\Sigma}_{\mathbf{x}}$ prior to applying SDR methods. As an ad hoc approach, it worked well, but it lacks theoretical support to show why the estimated directions are in the CS. Li et al. (2007) and Cook et al. (2007) combined a partial least squares approach with SDR, and they showed that such estimated directions are in the CS. Zhong et al. (2005) used a ridge type sample covariance estimate to avoid inverting the problematic sample covariance matrix. Li and Yin (2008) transformed the dimension reduction problem into a least-squares formulation and then applied a penalized approach to handle the singularity of the sample covariance matrix.

Indeed, many SDR methods used a tiny tuning parameter as a ridge type estimator to deal with possible singularity of the sample covariance matrix, such as those methods developed by Xia et al. (2002), Wang and Xia (2008), Yin and Li (2011), Li et al. (2012, 2014); Cook and Zhang (2014), Fukumizu et al. (2009) and Fukumizu and Leng (2014). However, whether a ridge type estimator is in the correct CS $S_{Y|\mathbf{X}}$ remains to be a question. The answer we propose to this question is that one cannot guarantee a ridge type estimator theoretically to be in the right CS $S_{Y|\mathbf{X}}$, but its sample estimate can be better than the usual estimate.

3.1.2 Review of ridge regression and measurement error

Ridge regression (Hoerl and Kennard, 1970) is one of the most well-known technique to handle multi-collinearity. Consider a linear model,

$$\mathbf{Y} = \mathcal{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.1}$$

where $\mathbf{Y}$ is an $n \times 1$ response vector, $\mathcal{X}$ is an $n \times p$ design matrix, $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector, and $\boldsymbol{\varepsilon}^T = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$ are i.i.d. random noise. With a tuning parameter $\lambda > 0$, the ridge estimator using the L^2 penalty can be found explicitly as

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = (\mathcal{X}'\mathcal{X} + n\lambda\mathbf{I})^{-1}\mathcal{X}'\mathbf{Y}, \tag{3.2}$$

where $\mathbf{I}$ is the identity matrix. Its population version is $\boldsymbol{\beta}^{\text{ridge}} = (E[\mathbf{X}\mathbf{X}'] + \lambda\mathbf{I})^{-1}E[\mathbf{X}Y]$. Hoerl and Kennard (1970) showed that there exists a λ such that this estimator is better than the OLS estimator in terms of mean squared errors (MSE). However, the choice of λ is subjective and difficult (Gibbons, 1981; Kibria, 2003). More importantly, the existing selection methods rely on a linear model assumption.

In our study, we will show that the ridge regression is closely linked to SDR for measurement error regression. In measurement error regression (Fuller, 1987), a surrogate variable $\mathbf{W}$ that contains measurement error is obtained, whose relation to $\mathbf{X}$ is assumed to be the following:

$$\mathbf{W} = \gamma + \Gamma\mathbf{X} + \boldsymbol{\delta}, \tag{3.3}$$

where γ is a q-dimensional vector and Γ is a $q \times p$ matrix, which can be either known or unknown in different cases. And $\boldsymbol{\delta}$ is a q dimensional error vector which is independent of $\mathbf{X}$ and Y. To further simplify our discussion, we focus on an important case: $p = q$, $\gamma = 0$ and $\Gamma = \mathbf{I}$. Carroll and Li (1992) studied SDR in measurement regression for sliced inverse regression (SIR; Li 1991) and Lue (2004) investigated the case for principal hessian directions (PHD; Li 1992). Let $\mathbf{U} = \Sigma_{\mathbf{xw}}\Sigma_{\mathbf{w}}^{-1}\mathbf{W}$, where $\Sigma_{\mathbf{xw}}$ is the covariance matrix between $\mathbf{X}$ and $\mathbf{W}$, under a general condition, Li and Yin (2007) established the following result: $S_{Y|\mathbf{U}} = S_{Y|\mathbf{X}}$, which can be equivalently written as

$$\Sigma_{\mathbf{w}} S_{Y|\mathbf{W}} = \Sigma_{\mathbf{x}} S_{Y|\mathbf{X}}. \tag{3.4}$$

Li and Yin (2007) assumed that an auxiliary sample which provides information about the relationship between the true predictor $\mathbf{X}$ and the surrogate predictor $\mathbf{W}$ is usually available to estimate $\Sigma_{\mathbf{xw}}$. Hence, one can recover $S_{Y|\mathbf{X}}$ by working on $\mathbf{U}$. If $\mathbf{W}$ and $\mathbf{X}$ are related as in (3.3) with $\gamma = 0$, immediately one can establish the following result.

Proposition 1 (Ridge equivalence) *Assume that* $\mathrm{E}[\boldsymbol{X}] = 0$, *and the added error* $\boldsymbol{\delta}$ *follows a multivariate normal distribution* $N_p(\boldsymbol{0}, \lambda\boldsymbol{I})$ *with* $\lambda > 0$. *Let* $\boldsymbol{\beta}_{\boldsymbol{w}}^{OLS}$ *denote the OLS estimate based on the error-prone predictor* $\boldsymbol{W}$ *and let* $\boldsymbol{\beta}_{\boldsymbol{x}}^{ridge}$ *denote the ridge estimate based on the* *true predictor* $\boldsymbol{X}$*, then*

$$\boldsymbol{\beta}_{\boldsymbol{w}}^{OLS} = \boldsymbol{\beta}_{\boldsymbol{x}}^{ridge}. \tag{3.5}$$

The equivalence reveals the true nature of ridge regression. In the population sense, the ridge estimate based on the observed $\mathbf{X}$ with tuning parameter λ is same as the OLS estimate based on $\mathbf{W}$ under (3.3) with the specific error structure from $N_p(\mathbf{0}, \lambda\mathbf{I})$. Hence, under certain conditions, the estimated direction $\boldsymbol{\beta}_{\mathbf{w}}^{\text{OLS}}$ is in the CS, $S_{Y|\mathbf{W}}$, rather than in $S_{Y|\mathbf{X}}$ as we wish.

3.1.3 Remarks

In ridge regression, since we have a model, the accuracy can be measured by the mean square error (MSE), where MSE $= \mathrm{E}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$, in which $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ are the respective population parameter and its estimate. However, in dimension reduction, only the CS is unique while the basis matrices may be different. Thus we use the criterion of Li et al. (2005)

$$\Delta(\hat{\mathbf{B}}, \mathbf{B}) = \parallel \hat{\mathbf{B}}(\hat{\mathbf{B}}^T\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^T - \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T \parallel \tag{3.6}$$

to evaluate the accuracy of the estimated basis matrix $\hat{\mathbf{B}}$ if the true basis $\mathbf{B}$ is known. The norm $\parallel \cdot \parallel$ is taken as the Frobenius norm such that $\parallel \mathbf{D} \parallel = \sqrt{\text{trace}(\mathbf{D}\mathbf{D}^T)}$. A subtle point is that if the direction is the most accurate, then we consider its estimation method to be the best. This is because once the direction is estimated, many different approaches can be used to build a model. For instance, James and Stein (1961) showed that a shrinkage estimator is better than the OLS estimator in terms of MSE, though both directions are the same.

When a sample $\mathcal{X}$ of size n is observed, the sample covariance matrix of $\mathcal{X}$ is $\mathcal{X}'\mathcal{X}/n$. It is common to scale the data as $\mathcal{X}_{scale} = \mathcal{X}/\sqrt{n}$ so that the diagonal elements of $\mathcal{X}'_{scale}\mathcal{X}_{scale}$ are the sample variances of $\mathcal{X}$. Under model (3.1), if we also scale the response as $\mathbf{Y}_{scale} = \mathbf{Y}/\sqrt{n}$, it is invariant to the use of ($\mathcal{X}_{scale}$, $\mathbf{Y}_{scale}$) or ($\mathcal{X}$, $\mathbf{Y}$) in the estimation. However, the ridge parameter in (3.2) will be different as $\lambda_{scale} = \lambda/n$, where λ is the ridge parameter using $\mathcal{X}$ and λ_{scale} is the ridge parameter using $\mathcal{X}_{scale}$. For our discussion later, the data are scaled.

3.1.4 Contributions and outline

In this paper, we explore the ridge regression by linking it to the measurement error regression and sufficient dimension reduction (SDR). Such a connection not only helps us to discuss the essence of ridge estimate from measurement error point of view which provides a deeper understanding for an important, classical and well-known method, but also inspires us to propose a new concept of pseudo SDR and develop a new procedure which is aimed at a different parameter space than the parameter space of interest but produces better estimate in practice. Our method doesn't require a parametric assumption on the underlying model between the predictors and response variable. Empirical evidence shows that when a preselected ridge tuning parameter is available, our method can always improve it. If there is no prior knowledge on the ridge tuning parameter, we can still improve the original estimate. Hence, we bypass the difficulties in choosing a possible optimal ridge tuning parameter using ensemble. Our approach is model-free and method-free. Thus it can be uniquely applied to any SDR methods. Moreover, our method not only provides better estimate for highly correlated predictors, but also works for the case when correlations between predictors are mild or small. In addition, we answer an interesting question in SDR: are ridge type estimators for SDR methods are in the correct subspace? For the prostate cancer data, our proposed method using a combination of SDR and ridge regression not only improves the accuracy over existing methods but also sufficiently justifies the assumption of using a linear model which was assumed by other studies. In Section 3.2, we introduce a new concept of pseudo SDR and pseudo SVS based on the measurement error idea. We discuss a plan to select the tuning parameter and propose a stable approach using an ensemble of multiple pseudo estimates. We will also discuss how to obtain a pseudo estimate for the structure dimension in SDR. In Section 3.3, we provide algorithms to obtain pseudo estimates based on OLS and SIR, as well as an algorithm for pseudo SVS. In Section 3.4, we show some numerical results using simulations, analyze the prostate cancer data with highly correlated predictors, and study the managerial role performance data (Warren et al., 1974) when measurement error are contained in the predictors. Further comments are arranged in Section 3.5.

3.2 Pseudo SDR and Pseudo SVS

3.2.1 Pseudo SDR

When $\hat{\Sigma}_{\mathbf{x}}$ is not invertible, our approach is to add an error term to the predictor vector so that the problem becomes to a measurement error regression as previously discussed in Section 3.1.2. Unfortunately, because $\hat{\Sigma}_{\mathbf{x}}$ is not invertible, using the transformed $\mathbf{U}$-scale predictors as in Li and Yin (2007) will still fail to recover $S_{Y|\mathbf{X}}$. Therefore, we propose, as defined below, to use the $\mathbf{W}$ scale for estimating $S_{Y|\mathbf{W}}$ as an approximation to the estimator of $S_{Y|\mathbf{X}}$, based on (3.4). In general, we know that $S_{Y|\mathbf{W}}$ and $S_{Y|\mathbf{X}}$ are two different subspaces. They are coincident, if $\Sigma_{\mathbf{w}} = constant \times \Sigma_{\mathbf{x}}$. In the next section, we will discuss how to obtain a good estimate for $S_{Y|\mathbf{W}}$ such that it can be a good estimate for $S_{Y|\mathbf{X}}$ as well.

Definition 1 (Pseudo SDR) *Suppose that $\boldsymbol{W} = \boldsymbol{X} + \boldsymbol{\delta}$, where $\boldsymbol{\delta} \sim N(\boldsymbol{0}, \Sigma_\delta)$, then the estimated $S_{Y|\boldsymbol{W}}$ is called the pseudo SDR estimator of $S_{Y|\boldsymbol{X}}$, if it is treated as the estimated $S_{Y|\boldsymbol{X}}$.*

As indicated by Proposition 1, the ridge estimator is a pseudo estimator for the OLS estimator. However, we would like to point out that James-Stein's estimator is not a pseudo SDR estimator but a shrinkage estimator, as its direction is correct in $S_{Y|\mathbf{X}}$. Estimators using diagonal sample covariance matrices are not pseudo SDR estimators either, because the relationship in (3.4) does not hold for them in general.

When a sample $\mathcal{X}$ is observed, we can always create a surrogate sample $\mathcal{W}$ as described in Definition 1 and obtain an estimate based on $\mathcal{W}$ as a pseudo estimate. Proposition 1 shows that a pseudo estimate for OLS is of the form

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{\text{pseudo}} = (\mathcal{X}'\mathcal{X} + n\Sigma_\delta)^{-1}\mathcal{X}'Y. \tag{3.7}$$

We can see that ridge estimator is a special case of (3.7) with $\Sigma_\delta = \lambda\mathbf{I}$. In this paper, in order to ease the discussion, we particularly choose $\Sigma_\delta = \lambda\mathbf{I}$. If $\lambda = 0$, then $S_{Y|\mathbf{W}} = S_{Y|\mathbf{X}}$. In Section 3.2.4.1, we will discuss on how to choose a proper λ for a given sample with fixed p.

3.2.2 Pseudo SVS

Similar to the pseudo SDR, variable selection can be done using $\mathbf{W}$ instead. Therefore, a pseudo central variable selection space, $S^V_{Y|\mathbf{W}}$, is a sufficient variable selection space based on $\mathbf{W}$, where $\mathbf{W} = \mathbf{X} + \boldsymbol{\delta}$ with $\boldsymbol{\delta} \sim N(\mathbf{0}, \Sigma_\delta)$. Once $\mathbf{W}$ is created, there are many ways to implement variable selection. For example, one can directly use a penalized approach on $\mathbf{W}$ to obtain a sparse estimate. Note that although many penalized methods for the linear model do not require inverting the sample covariance matrix of $\mathbf{X}$, most of penalized dimension reduction methods which are based on the least squares forms still need an inverse of the sample covariance matrix. In such a case, pseudo variable selection becomes useful when having difficulty inverting the sample covariance matrix. However, using a single $\mathbf{W}$ to obtain one sparse estimate may not be accurate. We will hence propose the the following non-penalized approach by creating many $\mathbf{W}$s.

Bootstrap samples are useful to estimate the standard error of parameters and make related inferences. In the regression, bootstrap typically can be done in two ways: (A) treating the regressors as random and selecting bootstrap samples directly from the observations $\mathbf{z}_i' = [Y_i, \mathbf{X}_{i1}, \cdots, \mathbf{X}_{ip}]$ or (B) treating the regressors as fixed and resampling from the residuals e_i of the fitted regression model. In (B), bootstrap observations are constructed as $Y_{bi}^* = \hat{Y}_i + e_{bi}^*$, where the $\hat{Y}_i$ are the fitted values from the original regression and the e_{bi}^* are the resampled residuals for the b^{th} bootstrap sample. However, there are situations when bootstrap becomes inadequate in making inferences in regression analysis. For example, when the ratio between the number of predictors p and the number of observations n become closer and closer, i.e. $n/p \to 1$, the performance of bootstrap estimates start to break down very soon. Also when high correlations among predictors exist, bootstrap estimates are also questionable. Therefore, obtaining multiple pseudo estimates is a suitable alternative to bootstrap samples in these situations. To obtain a sparse estimate of $\boldsymbol{\beta}$, we will discuss how an empirical confidence interval based on multiple pseudo estimates can be used to determine which element of $\boldsymbol{\beta}$ is zero so that sparse estimation can be achieved without using a penalized procedure.

In the linear model, it is known that the distribution of the regression coefficient β_k follows that

$$\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}_{\hat{\beta}_k}} \sim \mathbf{t}_{n-p}.$$

Therefore, if we can estimate the standard error of $\hat{\beta}_k$ by surrogate samples, we can construct empirical confidence intervals using extrinsic samples.

Definition 2 (**Pseudo confidence interval**)

If $\boldsymbol{z}_i = \{Y_i, \boldsymbol{X}_i\}$ for $i = 1, \cdots, n$ is the original observed sample where $\boldsymbol{X}_i = (X_1, \cdots, X_p)$, let $\boldsymbol{z}_{bi}^* = \{Y_i, \boldsymbol{W}_i\}$ with $\boldsymbol{W}_i = \boldsymbol{X} + \boldsymbol{\delta}_i$ where $\boldsymbol{\delta}_i \sim N(\boldsymbol{0}, \lambda \boldsymbol{I})$ for $i = 1, \cdots, n$ and $b = 1, \cdots, N$

be N surrogate samples of size n. Suppose that $\hat{\beta}_k$ is the estimated regression coefficient for X_k based on original observed sample and $\hat{\beta}^*_{bk}$ is the estimated regression coefficient for X_k base on each surrogate sample, the standard error of $\hat{\beta}_k$ can be estimated by $S.E.^*(\hat{\beta}^*_k) = \sum_{b=1}^{N} \frac{(\hat{\beta}^*_k - \bar{\beta}^*)^2}{(N-1)}$, where $\bar{\beta}^*$ is the average of $\hat{\beta}^*_{bk}$'s. If N is sufficiently large, a $100 \times (1-\alpha)\%$ Pseudo Confidence Interval (PCI) can be obtained by $\hat{\beta}_k \pm z_{\alpha/2} S.E.^*(\hat{\beta}^*_k)$.

Definition 2 is very similar to the definition of bootstrap confidence interval except for each $\mathbf{W}$, the effective sample size is much greater than the bootstrap samples in calculating $\hat{\beta}^*_{bk}$. The enlarged sample size significantly reduces the bias and variation caused by the bootstrap samples especially for the case when the number of predictors p is close to the sample size n or the correlations among predictors are high.

In many situations, the coefficient vector $\boldsymbol{\beta}$ in (3.1) is sparse meaning that some of its components are zeros. Let $S = \{i : \beta_i \neq 0\}$ be the set of indices of nonzero components of $\boldsymbol{\beta}$. One goal in statistics is to infer S through a variable selection procedure (Tibshirani, 1996; Meinshausen and Bühlmann, 2010; Shao and Deng, 2012). Using PCI as defined above can also provide a sparse estimate for variable selection purpose. To be more specific, let $\hat{\boldsymbol{\beta}}$ be the pseudo estimator as defined in (3.7), set $\tilde{\beta}_i = 0$ if the $100 \times (1-\alpha)\%$ PCI of $\hat{\beta}_i$ contains 0, otherwise set $\tilde{\beta}_i = \hat{\beta}_i$. Let $\hat{S} = \{i : \tilde{\beta}_i \neq 0\}$ be the set of indices of components selected using PCI. We call this variable selection procedure based on PCI the pseudo SVS.

3.2.3 Estimating dimensionality

Sometimes it is reasonable to assume that the dimension d is known. For example in ridge regression, $d = 1$. However, in general, d has to be estimated. Equation (3.4) indicates that the dimension of $S_{Y|\mathbf{W}}$ is same as that of $S_{Y|\mathbf{X}}$. Thus we propose the following pseudo SDR approach to estimate d. We first randomly generated a $\tilde{\lambda}_k$ from $U(l, u)$, where l and u are small positive constants. We then use an available method, for example the sequential chi-square test (Li, 1991), to estimate the dimensionality of $S_{Y|\mathbf{W}}$, say d_{λ_k}. We repeat this step m times and we estimate d by the mode of these d_{λ_k}'s for $k = 1, ..., m$. Note that since the dimension of $S_{Y|\mathbf{W}}$ is same as that of $S_{Y|\mathbf{X}}$, theoretically the estimation of dimensionality based on $\mathbf{W}$ is independent of the choice of λ_k. By repeating m times, our estimate is stabilized as being seen in (Wu and Yin, 2015b). Moreover, when the sample $\hat{\Sigma}_{\mathbf{x}}$ is non-invertible, working on $\mathbf{W}$ is a feasible alternative.

3.2.4 Selection of tuning parameter

The most important thing in pseudo SDR is the selection of λ. A wise choice of λ ensures the pseudo estimator is indeed a good estimator. Recall that even for the ridge estimator, the most well-known and well-studied method, selection of tuning parameter is very difficult. For instance, in the last several decades, many criteria are proposed by Hoerl et al. (1975); Goldstein and Smith (1974); Obenchain (1975); Lindley and Smith (1972); Lawless and Wang (1976); Dempster et al. (1977). Comprehensive comparisons by Gibbons (1981) and Kibria (2003) showed that none of the above methods consistently outperforms the others. In such a simplest model-based case, the selection of the tuning parameter is quite unsatisfactory to a certain degree. Hence, it is difficult to develop an explicit formula to obtain the optimal λ for a pseudo SDR estimator.

3.2.4.1 Selection procedure

Our goal is to have a rule of thumb to select the tuning parameter λ so that it works in general for any dimension reduction method, when the sample covariance matrix is difficult to be inverted. As previously pointed out, even in the ridge regression case, where we have a particular model, selecting an optimal λ is a difficult task. As Gibbons (1981) and Kibria (2003) studied, there is no universal criterion for choosing the best λ in this simplest case. We expect more difficulties to define a universal criterion for selecting the best λ for a pseudo

SDR estimator. Since $S_{Y|\mathbf{W}} = S_{Y|\mathbf{X}+\delta}$, when λ becomes too large, $S_{Y|\mathbf{W}}$ departs from $S_{Y|\mathbf{X}}$ and becomes closer to $S_{Y|\delta}$ which is the null space as $Y \perp\!\!\!\perp \delta$. Therefore, we want to keep λ to be small enough so that $S_{Y|\mathbf{W}}$ can still be a good approximation to $S_{Y|\mathbf{X}}$. Empirical evidence showed that an effective range for λ is between 0.0001 and 0.01.

We use the following example to illustrate our conjecture using a small quantity λ to obtain the pseudo estimate. Consider a linear model:

$$Y = 3X_1 + 5X_2 + c\varepsilon.$$

The predictors $\mathbf{X} = (X_1, \cdots, X_p)$, $p = 10$, are generated from a p-dimensional multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma_x}$ with diagonal elements $\sigma_{ii} = 1$ and off-diagonal elements $\sigma_{ij} = 0$ except for $\sigma_{13} = \sigma_{31} = \rho$, where $\rho \in (0, 1)$ is the correlation between X_1 and X_3. We use two setups: (A), without measurement error and (B), with measurement error by setting $\tilde{\mathbf{X}} = \mathbf{X} + \delta$, where $\delta \sim N(\mathbf{0}, 0.25\mathbf{I})$. Figure 3.1 shows the box plots for the two setups: the left panel is for (A) and the right panel is for (B). For (A), we

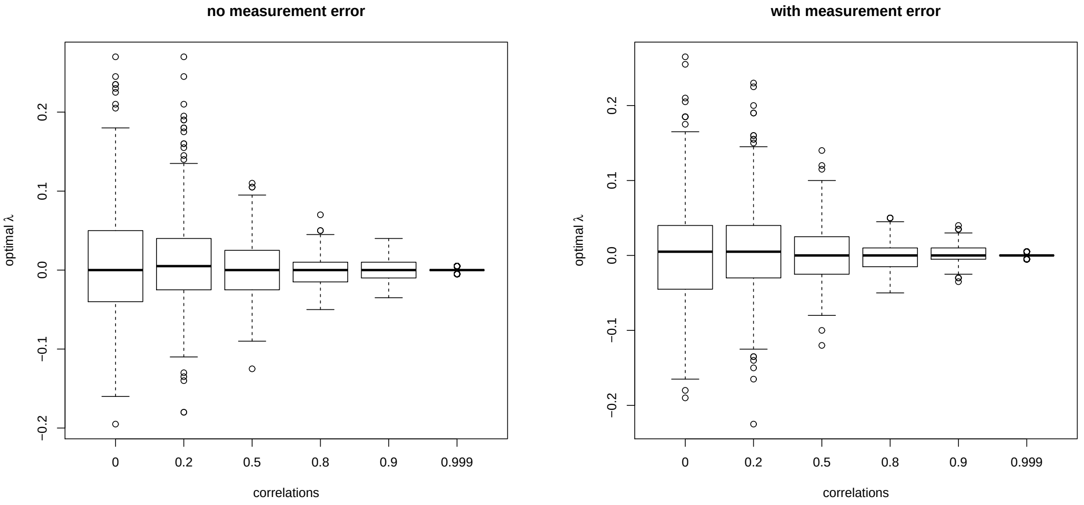

Figure 3.1: Box plots of optimal ridge tuning parameter using $\mathbf{X}$

use the ridge formula (3.2) on $\mathbf{X}$ and run λ over the range [-1,1]. From there we choose λ to be optimal if it has the smallest measure of accuracy. For (B), we first transform $\tilde{\mathbf{X}}$ into $\mathbf{U}$ using method proposed by Li and Yin (2007) and then repeat the same procedures as for (A), except using $\mathbf{U}$ instead of $\mathbf{X}$. We run 500 simulations for each setup. We can see that when there is no measurement error, the optimal λ is around zero, and the range of optimal λ becomes smaller when correlation is getting higher; when measurement errors are already contained in the observed sample, the optimal tuning parameter is still near 0, similar to that of the no measurement error case.

3.2.4.2 Ensemble approach

Section 3.2.4.1 provides an ad hoc method for choosing a tuning parameter. We realize that a particular choice of small λ may not be the optimum. To partly circumvent the possibility of choosing a bad tuning parameter, we propose an ensemble idea. Note that for any choice of λ, equation (3.4) guarantees that the estimator using such a λ is a pseudo SDR estimator. If a possible choice of λ can lead to a reasonable pseudo estimator, varying such λ will lead to many reasonable pseudo estimators. We then can aggregate all these possible estimators together to obtain a more stable estimate.

This idea is similar to that of Wu and Yin (2015b), where subsamples were used to obtain a stable and better estimate. However, their approach cannot deal with collinear data due to repeatedly using part of the same data. We call a method an intrinsic method if it repeatedly uses part of or the entire data multiple times. The class of intrinsic methods includes the work by Meinshausen and Bühlmann (2010); Zhu et al. (2010b); Yin and Li (2011), and more recently by Cook and Zhang (2014). These intrinsic methods use part of or the entire data multiple times to enhance the accuracy of the estimates by estimating the same population parameter space, while our proposed method is to use a small perturbation of the data to enhance the accuracy by estimating a variation of the targeted parameter space. We call the latter an extrinsic method. Note that our extrinsic method is specifically designed to deal with ill conditioned covariance matrix. Therefore, this idea leads us to the new procedures described in Section 3.3.

3.2.4.3 Improve ridge estimate with a given parameter

As summarized by Gibbons (1981) and Kibria (2003), many procedures have been proposed to select a working ridge tuning parameter from different point of views. All these methods assume a linear model and none of them is consistently better than others in different situations. We propose to improve ridge estimate based on a preselected tuning parameter λ_0 by aggregating different pseudo estimates based on different tuning parameters λ_i which are randomly generated from the neighborhood of λ_0, say $(\lambda_0 - 0.0005, \lambda_0 + 0.0005)$. Our empirical results in Section 3.4.2 show that this ensemble approach consistently improve upon the ridge estimate using λ_0. Compared with the procedure described in the last two sections, when λ_0 is preselected by some methods, we have prior knowledge for the range of the optimal tuning parameter. When we don't have a preselected tuning parameter, we simply use a small quantity to start with, as discussed in Section 3.2.4.1. This idea is analogous to selecting the prior distribution in a Bayesian context. A noninformative prior is used unless there is a specific distribution preferred to be the prior.

3.3 Illustrative algorithms

Suppose that the dimensionality d is known, our procedure is as follows: we generate a random sample $\tilde{\lambda}_k$, $k = 1, ..., m$ from (l, u) uniformly, where l and u are small positive constants and set $\lambda_k = \tilde{\lambda}_k^2$. Let $\mathbf{B}_{\lambda_k}$ be the pseudo SDR estimator. Then the d eigenvectors of $\sum_{k=1}^m \mathbf{B}_{\lambda_k}\mathbf{B}_{\lambda_k}^T$ corresponding to the d largest eigenvalues will be our final estimate.

To be more specific, we illustrate the pseudo SDR idea by using SIR. Assuming the dimensionality d is known, a pseudo SIR estimator can be obtained using the following procedure:

Algorithm 1 (**Pseudo SDR**)

1. *Let l and u be small positive constants (say $l = 0.001$ and $u = 0.005$ for example). Obtain m random numbers $\tilde{\lambda}_k$ from $U(l, u)$.*

2. *For each $\tilde{\lambda}_k$, let $\mathcal{W} = \mathcal{X} + \tilde{\lambda}_k \tilde{\Delta}$ where $\tilde{\Delta}$ is an $n \times p$ matrix with rows being generated from $\delta \sim N(\boldsymbol{0}, \boldsymbol{I})$, and obtain the SIR matrix $\boldsymbol{M}$ (Li, 1991) based on $\mathcal{W}$.*

3. *Solve the eigenvalue problem of the form $\boldsymbol{M}\nu_i = \phi_i \hat{\Sigma}_w \nu_i$ for $i = 1, \cdots, d$ where ϕ_i and ν_i are eigenvalues and eigenvectors of $\boldsymbol{M}$ with $\phi_1 > \cdots > \phi_p$ and $\hat{\Sigma}_w$ is the sample covariance of $\mathcal{W}$. One pseudo SIR estimate can be obtained as $\boldsymbol{B}_{\lambda_k} = (\nu_1, \cdots, \nu_d)$.*

4. *The final pseudo SIR estimate is obtained as the d eigenvectors of $\sum_{k=1}^{m} \boldsymbol{B}_{\lambda_k} \boldsymbol{B}_{\lambda_k}^T$ corresponding to its d largest eigenvalue.*

Similarly, pseudo OLS estimate can be obtained using above steps by setting $d = 1$ and $\mathbf{B}_{\lambda_k} = (\mathcal{W}'\mathcal{W})^{-1}\mathcal{W}'\mathbf{Y}$. To implement pseudo SVS, we use the following steps:

Algorithm 2 (**Pseudo SVS**)

1. *For the observed sample $\boldsymbol{z} = \{\boldsymbol{Y}, \mathcal{X}\}$, create m extrinsic samples of size n as $\boldsymbol{z}^* = \{\boldsymbol{Y}, \mathcal{X}\}$ with $\mathcal{W} = \mathcal{X} + \tilde{\lambda}_k \tilde{\Delta}$ where $\tilde{\Delta}$ is an $n \times p$ matrix with rows being generated from $\delta \sim N(\boldsymbol{0}, \boldsymbol{I})$. Then obtain an ordinary estimate $\hat{\boldsymbol{\beta}}_j^*$ based on the augmented sample $\boldsymbol{z}^*$.*

2. *Repeat above steps N times to obtain N estimates $\hat{\boldsymbol{\beta}}_j^*$ for $j = 1, \cdots, N$ based on N enlarged samples.*

3. *To obtain sparse estimate, for any predictor k, we can obtain its empirical quantiles based on the N estimates. If $\hat{\boldsymbol{\beta}}^*_{\alpha/2} \geq 0 \geq \hat{\boldsymbol{\beta}}^*_{1-\alpha/2}$, where $\hat{\boldsymbol{\beta}}^*_{\alpha/2}$ is the $(100 \times \alpha/2)^{th}$ percentile of the N estimates $\hat{\boldsymbol{\beta}}^*_j$, we set all the k^{th} element of $\hat{\boldsymbol{\beta}}^*_j$ to be 0. Then the eigenvectors of $\boldsymbol{M} = \sum_{i=1}^N \hat{\boldsymbol{\beta}}^*_j(\hat{\boldsymbol{\beta}}^*_j)'$ will be used as our final estimate.*

For linear models, OLS can be used and the first eigenvector $\hat{\boldsymbol{\beta}}^*$ of $\mathbf{M}$ can be used as the scaled final estimate. For nonlinear models, traditional SDR methods can be used and the first d eigenvector $\hat{\boldsymbol{\beta}}^*$ of $\mathbf{M}$ can be used as the basis for CS, where d is the structural dimension of the CS that is known or can be estimated. In Step 4, we can also use a hard threshold that is similar to Shao and Deng (2012) or replace $\hat{\boldsymbol{\beta}}^*_{\alpha/2}$ and $\hat{\boldsymbol{\beta}}^*_{1-\alpha/2}$ by $\bar{\boldsymbol{\beta}}^* \pm \Phi^{-1}(1-\alpha/2)\text{S.E.}^*(\boldsymbol{\beta}^*_j)$ where $\text{S.E.}^*(\hat{\boldsymbol{\beta}}^*_j) = \sum_{b=1}^N \frac{(\hat{\boldsymbol{\beta}}^*_j - \bar{\boldsymbol{\beta}}^*)^2}{(N-1)}$, $\bar{\boldsymbol{\beta}}^*$ is the average of $\hat{\boldsymbol{\beta}}^*_j$, and $\Phi(\cdot)$ is the c.d.f. of standard normal distribution.

Ensemble approach can help us obtain a more stable variable selection result. In specific, one can simply repeat the steps in Algorithm 2 T times to obtain T sparse estimate. For any predictor k, if the proportion of its occurrence among the T estimates is smaller than a preselected cutoff point π, we set all the k^{th} element of $\hat{\boldsymbol{\beta}}^*_j$ to be 0. Then the eigenvectors of $\mathbf{M} = \sum_{i=1}^T \hat{\boldsymbol{\beta}}^*_j(\hat{\boldsymbol{\beta}}^*_j)'$ will be used as our final stable estimate. In our simulations, we set $\pi = 0.85$.

3.4 Numerical study

In this section, we use simulations and two real data analyses to illustrate the idea of pseudo estimates and advantages of pseudo methods including pseudo non-sparse estimation and pseudo variable selection. In addition, we show the accuracy of estimating dimensionality by using pseudo SDR approach introduced in Section 3.2.3.

3.4.1 Estimation accuracy by Pseudo OLS and Pseudo SIR

To make comparisons between OLS and its corresponding pseudo method, and SIR and its corresponding pseudo method, we consider three linear models for OLS:

$$\text{Linear Model 1: } Y = 3X_1 + 5X_2 + c\varepsilon, \tag{3.8}$$

$$\text{Linear Model 2: } Y = 3X_1 + 1.5X_2 + 2X_5 + c\varepsilon, \tag{3.9}$$

$$\text{Linear Model 3: } Y = X_2 + 2(X_3 + X_4) + c\varepsilon, \tag{3.10}$$

and three nonlinear models for SIR:

$$\text{SIR Model 1: } Y = (X_1 + X_2)^3 + e^{(X_3+X_4)} + c\varepsilon, \tag{3.11}$$

$$\text{SIR Model 2: } Y = X_1/(0.5 + (5X_2 + 1.5)^2) + c\varepsilon, \tag{3.12}$$

$$\text{SIR Model 3: } Y = X_1(X_1 + X_2 + 1) + c\varepsilon. \tag{3.13}$$

The predictors $\mathbf{X} = (X_1, \cdots, X_p)$ are generated from a p-dimensional multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma_x}$ of which diagonal elements $\sigma_{ii} = 1$ and off-diagonal elements $\sigma_{ij} = 0$ except for $\sigma_{12} = \sigma_{21} = \rho$, where $\rho \in (0, 1)$ is the correlation between X_1 and X_2. The errors ε are generated independently from the standard normal distribution. The noise level c is set to be 0.25 and we use $p = 10$ predictors with sample size $n = 400$ for both models. We simulate two scenarios for each model to illustrate the effectiveness of pseudo estimates. The first scenario is when the predictor data $\mathcal{X}$ contain no measurement errors. In such a case, we directly obtain pseudo estimates based on $\mathcal{X}$ and compare our results to the respective OLS and SIR estimates. The second scenario is when the predictor data $\tilde{\mathcal{X}}$ contain measurement errors. To simulate this situation, we artificially created measurement error predictors $\tilde{\mathcal{X}}$ by adding small errors to the original $\mathcal{X}$. When measurement errors are present, we will use a result of Li and Yin (2007)'s **U** approach to obtain pseudo estimates based on $\tilde{\mathcal{X}}$. In our simulations, we set $l = 0.0001$ and $u = 0.005$, then aggregate over 500 estimates. Note that when there is no measurement error, $\mathcal{U}$ is identical to $\mathcal{X}$.

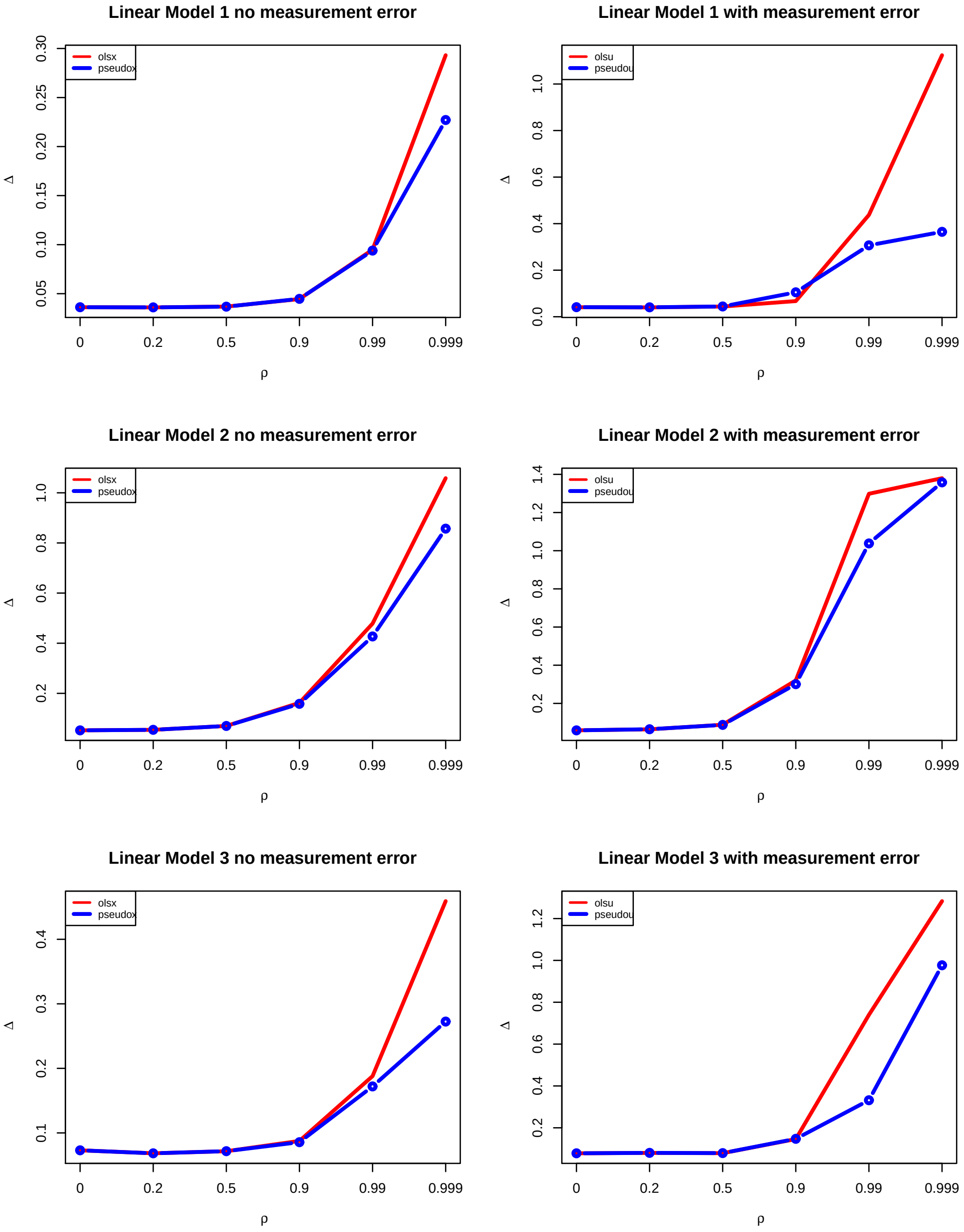

Figure 3.2: Estimation accuracy of linear models

Figure 3.2 illustrates the estimation accuracy of linear models (3.8) - (3.10) with and without measurement errors using (3.6). For all three models, the left column shows that when there are no measurement errors in the predictors, the pseudo estimates and OLS estimates are similar when correlations between predictors are low. However, the pseudo estimates outperform the OLS estimates when correlations between predictors become large. When measurement errors are contained in the predictors, the right column shows that the proposed pseudo estimates based on $\mathcal{U}$ are better than the Li and Yin (2007)'s estimates, especially when correlations between predictors are high.

In our simulation studies, we generate $\tilde{\lambda}_k$ from a very small range $(0.0001, 0.005)$ following the reason discussed in Section 3.2.4.1. In order to validate our conclusion, we apply several existing methods to choose the optimal tuning parameters: tuning parameters chosen by Hoerl et al. (1975) which is denoted by (HKB), by Lawless and Wang (1976) which is denoted by (LW), by Hocking et al. (1976) which is denoted by (HSL), and by Kibria (2003) which is denoted by (K), respectively. These methods represent different points of views. For example, HKB is a modified version of the original method proposed by Hoerl and Kennard (1970) based on the structure of MSE, and LW is proposed from the Bayesian point of view. As indicated by Table 3.1, the optimal tuning parameters for each of the linear models (3.8)-(3.10) should indeed be small. We can also see from Table 3.1 that each method suggests different tuning parameters. Therefore, by our ensemble step in the algorithm, we reduce the risk of using a bad tuning parameter value.

Figure 3.3 illustrates similar results for nonlinear models (3.11) - (3.13) using pseudo SIR estimates. For models without measurement error, the left column shows that, like linear models, when correlations are low between predictors, the SIR estimates and the pseudo SIR estimates are similar. However a significant improvement using the pseudo SIR estimates can be found when the correlations between variables start to increase. When predictors have measurement errors, the right column shows that the pseudo SIR estimates based on $\mathbf{U}$ are

					r		
		0	**0.2**	**0.5**	**0.9**	**0.99**	**0.999**
Model 1 no error	**HKB**	0.00072	0.00072	0.00072	0.00072	0.00071	0.00066
	LW	0.00060	0.00058	0.00055	0.00051	0.00049	0.00045
	HSL	0.00007	0.00007	0.00008	0.00008	0.00008	0.00008
	K	0.00277	0.00417	0.00961	0.01902	0.02388	0.02906
Model 1 with error	**HKB**	0.00071	0.00072	0.00072	0.00071	0.00072	0.00069
	LW	0.00059	0.00058	0.00055	0.00050	0.00049	0.00046
	HSL	0.00007	0.00007	0.00008	0.00008	0.00008	0.00008
	K	0.00250	0.00481	0.01802	0.02080	0.02296	0.02580
Model 2 no error	**HKB**	0.00160	0.00159	0.00158	0.00158	0.00141	0.00080
	LW	0.00134	0.00113	0.00075	0.00035	0.00029	0.00074
	HSL	0.00016	0.00017	0.00022	0.00042	0.00056	0.00058
	K	0.00632	0.00653	0.00408	0.00386	0.00358	0.00231
Model 2 with error	**HKB**	0.00158	0.00160	0.00160	0.00159	0.00140	0.00077
	LW	0.00133	0.00115	0.00077	0.00036	0.00029	0.00150
	HSL	0.00016	0.00017	0.00022	0.00042	0.00055	0.00058
	K	0.00598	0.00540	0.00447	0.00384	0.00362	0.00215
Model 3 no error	**HKB**	0.00276	0.00272	0.00269	0.00268	0.00273	0.00226
	LW	0.00250	0.00242	0.00240	0.00238	0.00242	0.00188
	HSL	0.00028	0.00028	0.00028	0.00030	0.00031	0.00031
	K	0.01005	0.01010	0.00827	0.00780	0.00959	0.00812
Model 3 with error	**HKB**	0.00276	0.00268	0.00269	0.00272	0.00268	0.00227
	LW	0.00249	0.00239	0.00239	0.00240	0.00236	0.00189
	HSL	0.00028	0.00027	0.00028	0.00030	0.00031	0.00031
	K	0.01091	0.00909	0.00837	0.00757	0.00900	0.00962

Table 3.1: Optimal ridge tuning parameters using existing methods

consistently better than the original SIR estimates based on $\mathbf{U}$, especially when correlations between predictors are high.

3.4.2 With a given tuning parameter

Corresponding to our discussion in Section 3.2.4.3, in this simulation, we illustrate how pseudo estimates can be used to improve the existing ridge estimates with a tuning parameter obtained by some criteria. We compare four different ridge estimates for OLS models using tuning parameters chosen by the methods that used in Section 3.4.1.

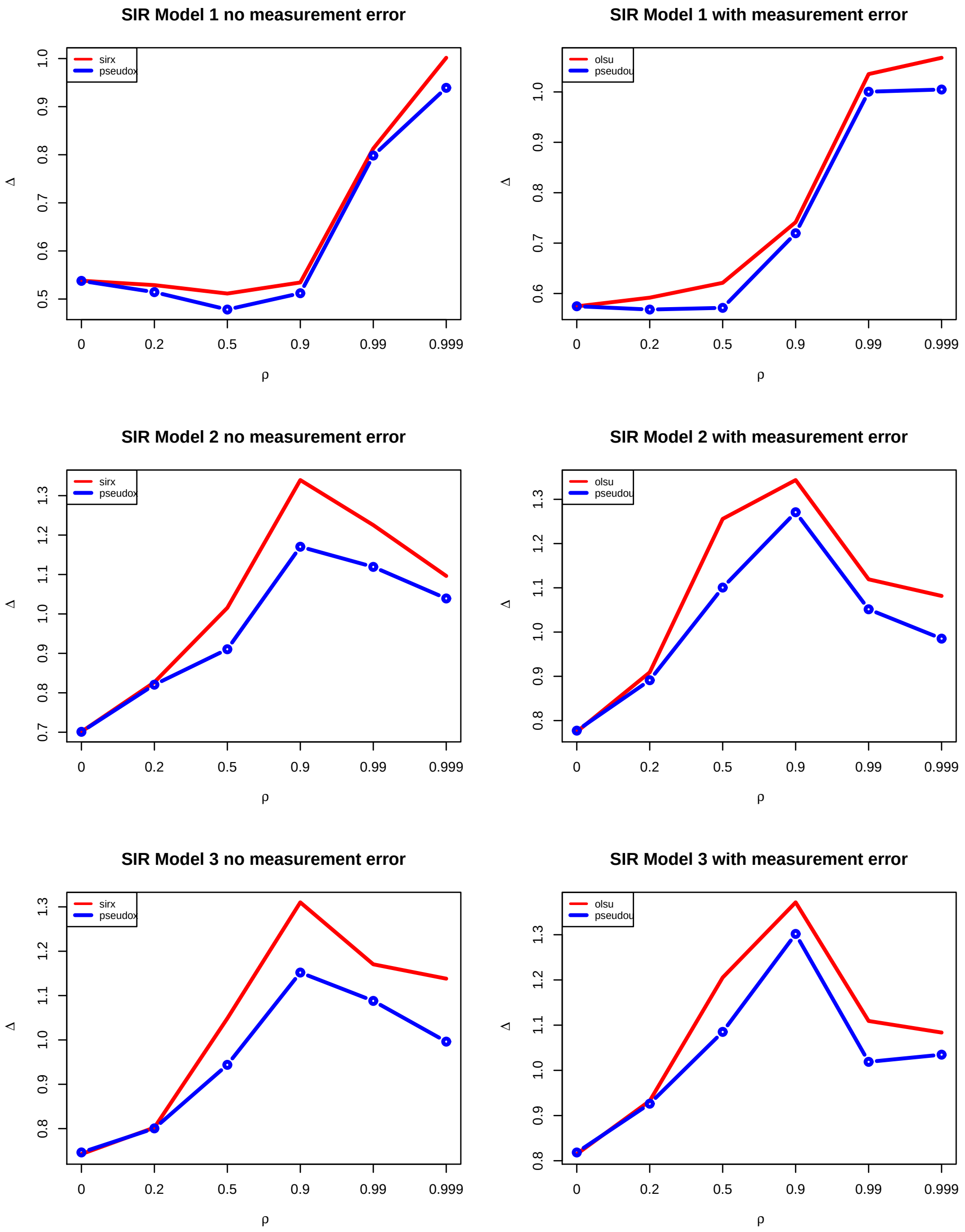

Figure 3.3: Estimation accuracy of SIR models

To make comparisons, we report ridge estimates using different tuning parameters chosen by these methods, and pseudo ridge estimates assemble multiple estimates based on tuning parameters in the neighborhood, e.g. ±0.005, of the chosen tuning parameter. We also report the OLS estimate and its corresponding pseudo estimate as discussed in Section 3.3. Using Model (3.8), Table 3.2 shows that all four ridge estimates improve the original OLS estimates. In addition, all pseudo estimates improve the respective original estimates, except one case for Kibria (2003)'s criteria where data are not correlated at all. Therefore, in general, if there is no prior choice for the ridge tuning parameter, we can use the pseudo estimate to improve the original OLS estimate. If there is a choice for ridge tuning parameter, we can still obtain a better pseudo ridge estimate using the chosen tuning parameter.

	r = 0.0		**r = 0.5**		**r = 0.9**	
	Original	**Pseudo**	**Original**	**Pseudo**	**Original**	**Pseudo**
HKB	0.1164	0.1094	0.1461	0.1454	0.3752	0.3540
LW	0.1181	0.1175	0.1462	0.1456	0.3878	0.3813
HSL	0.1193	0.1183	0.1463	0.1462	0.3860	0.3791
K	0.1102	0.1172	0.1460	0.1440	0.3698	0.3531
OLS	0.1198	0.1164	0.1463	0.1456	0.3912	0.3825

Table 3.2: Improvement over existing ridge estimates

3.4.3 Variable selection by pseudo SVS

In order to demonstrate pseudo variable selection based on multiple **W**s for linear model, we will use the following model from Zou (2006):

$$Y = 3X_1 + 1.5X_2 + 2X_5 + c\varepsilon. \tag{3.14}$$

And for nonlinear model using SIR, we use the the following model from Li (1991):

$$Y = X_1(X_2 + 1.5) + c\varepsilon. \tag{3.15}$$

The predictors $\mathbf{X}_i$ (i = 1,..., n) were i.i.d. standard normal vectors as well as the noise vector $\boldsymbol{\varepsilon}$. We set the pairwise correlation between $\mathbf{X}_j$ and $\mathbf{X}_k$ to be $cor(\mathbf{X}_j, \mathbf{X}_k) = \rho^{|j-k|}$, in which we vary the correlation coefficient between 0 and 0.999. The noise level c is set to be 0.1. Since pseudo variable selection is useful for the case when correlations between variables are high and the sample size is relatively small comparing to the number of predictors. We simulate different situations of sample size and number of predictors (n, p) for (25, 10), (50, 20), (100, 40) and (200, 80) so that ratio between sample size and number of predictors are 2.5 for all cases. As mentioned in Section 3.2.2, pseudo variable selection is achieved based on pseudo confidence interval which is very similar to bootstrap confidence interval. For comparison purpose, we report the true positive rates (TPR) and false positive rates (FPR) for variable selections based on pseudo confidence interval, bootstrap confidence interval, and the empirical confidence interval using subsamples.

Table 3.3 and Table 3.4 illustate that when the correlations between predictors cause no problem for inverting the sample covariance matrix. even for high correlation r = 0.9, variable selections based on all empirical intervals in general perform very well for all (n, p) combinations, but the FPRs for pseudo variable selections are lower than FPRs based on bootstrap intervals and subsample intervals. However, when the correlation is too high, r = 0.999, the sample covariance matrix is close to be singular, in such a case, the TPRs decrease for all (n, p) combinations. For this situation, pseudo variable selections perform much better than variable selections based on bootstrap and subsample intervals. This advantage becomes significant when the sample size is sufficiently large.

3.4.4 Estimating dimensionality by pseudo SIR

We now demonstrate the estimation of dimensionality for a nonlinear model (3.12) using the pseudo SIR method. As mentioned in Section 3.2.3, the pseudo estimate approach in estimating dimensionality is theoretically stable with different added errors, and is extremely

Table 3.3: TPRs and FPRs of pseudo SVS for linear model

		PSVS		**Bootstrap**		**Subsample**	
	ρ	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**
n = 25 p = 10	**0**	1.000	0.000	1.000	0.046	1.000	0.007
	0.2	1.000	0.000	1.000	0.053	1.000	0.016
	0.5	1.000	0.001	1.000	0.064	1.000	0.010
	0.8	1.000	0.004	1.000	0.043	1.000	0.006
	0.9	1.000	0.000	1.000	0.047	1.000	0.009
	0.999	0.463	0.021	0.330	0.047	0.180	0.013
n = 50 p = 20	**0**	1.000	0.000	1.000	0.028	1.000	0.003
	0.2	1.000	0.000	1.000	0.021	1.000	0.001
	0.5	1.000	0.001	1.000	0.024	1.000	0.001
	0.8	1.000	0.002	1.000	0.025	1.000	0.002
	0.9	1.000	0.004	1.000	0.017	1.000	0.002
	0.999	0.637	0.021	0.377	0.016	0.110	0.001
n = 100 p = 40	**0**	1.000	0.000	1.000	0.018	1.000	0.001
	0.2	1.000	0.000	1.000	0.016	1.000	0.001
	0.5	1.000	0.000	1.000	0.020	1.000	0.001
	0.8	1.000	0.001	1.000	0.018	1.000	0.001
	0.9	1.000	0.002	1.000	0.020	1.000	0.001
	0.999	0.837	0.040	0.570	0.017	0.300	0.001
n = 200 p = 80	**0**	1.000	0.000	1.000	0.012	1.000	0.001
	0.2	1.000	0.000	1.000	0.012	1.000	0.001
	0.5	1.000	0.000	1.000	0.012	1.000	0.001
	0.8	1.000	0.001	1.000	0.011	1.000	0.001
	0.9	1.000	0.002	1.000	0.013	1.000	0.001
	0.999	0.973	0.029	0.797	0.012	0.453	0.001

useful when the original methods fail. Again we choose $l = 0.001, u = 0.005$. The true model (3.12) clearly has two dimensions. To make the comparison, we set correlations between two of the predictors in each situation to be (0, 0.2, 0.5, 0.9, 0.999, 1). When the correlation is equal to 1, it means that the two variables are the same and using predictor data $\mathcal{X}$ to test the dimensionality will fail. For this part of our simulation, we use three different extreme cases: $X_1 = X_2$, $X_1 = X_3$, and $X_3 = X_4$ representing situations where (1) important variables are correlated with each other, (2) one important variable is correlated with unimportant variables, and (3) unimportant variables are correlated with each other.

Table 3.4: TPRs and FPRs of pseudo SVS for SIR model

		PSVS		**Subsamples**		**Bootstrap**	
	ρ	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**
n = 25 p = 10	**0**	0.850	0.125	0.625	0.000	0.400	0.000
	0.2	0.875	0.131	0.575	0.000	0.300	0.000
	0.5	0.825	0.038	0.500	0.000	0.125	0.000
	0.8	0.850	0.000	0.475	0.000	0.075	0.000
	0.9	0.850	0.000	0.375	0.000	0.000	0.000
	0.999	1.000	0.000	0.000	0.000	0.000	0.000
n = 50 p = 20	**0**	0.960	0.083	0.620	0.000	0.215	0.000
	0.2	0.900	0.043	0.585	0.000	0.195	0.000
	0.5	0.890	0.068	0.545	0.000	0.115	0.000
	0.8	0.940	0.009	0.500	0.000	0.015	0.000
	0.9	0.990	0.011	0.490	0.000	0.000	0.000
	0.999	1.000	0.014	0.005	0.000	0.000	0.000
n = 100 p = 40	**0**	0.995	0.044	0.935	0.000	0.540	0.000
	0.2	1.000	0.038	0.905	0.000	0.510	0.000
	0.5	1.000	0.039	0.845	0.000	0.500	0.000
	0.8	1.000	0.011	0.570	0.000	0.500	0.000
	0.9	1.000	0.011	0.515	0.000	0.435	0.000
	0.999	1.000	0.009	0.035	0.000	0.000	0.000
n = 200 p = 80	**0**	1.000	0.024	1.000	0.000	0.885	0.000
	0.2	1.000	0.019	1.000	0.001	0.860	0.000
	0.5	1.000	0.020	0.985	0.001	0.655	0.000
	0.8	1.000	0.007	0.795	0.001	0.510	0.000
	0.9	1.000	0.010	0.565	0.001	0.500	0.000
	0.999	1.000	0.008	0.125	0.000	0.005	0.000

Table 3.5 illustrates the testing accuracy of dimensionality by applying the sequential chi-square test (Li, 1991) on $\mathcal{X}$ and using pseudo SIR with the sequential chi-square test on $\mathcal{W}$ as described in Section 3.2.3. For comparison purpose, we also include the sparse eigenvalue decomposition (SED, Zhu et al. 2010a) approach using pseudo SIR. Table 3.5 indicates that both chi-square tests and SED based on the SIR matrix of $\mathcal{X}$ are very consistent for cases when one important variable is correlated with unimportant variables and unimportant variables are correlated with each other even for very high correlations. However, when $\mathcal{X}'\mathcal{X}$ becomes non-invertible, testing based on $\mathcal{X}$ fails but the testing based on $\mathcal{W}$ is still very consistent. When there are some high correlations between important variables, chi-square

Table 3.5: Estimating structural dimension by pseudo SIR

						ρ			
			0	**0.5**	**0.9**	**0.99**	**0.999**	**0.9999**	**1**
X1 and X2	**Chi Square**	**SIR**	0.930	0.985	0.280	0.045	0.040	0.060	0.000
		pseudo SIR	0.930	0.985	0.260	0.035	0.030	0.050	0.030
	SED	**SIR**	0.995	0.995	0.530	0.595	0.685	0.675	0.000
		pseudo SIR	0.995	0.995	0.820	0.820	0.870	0.850	0.860
X1 and X3	**Chi Square**	**SIR**	0.955	0.965	0.935	0.965	0.945	0.945	0.000
		pseudo SIR	0.955	0.965	0.935	0.975	0.945	0.965	0.970
	SED	**SIR**	1.000	1.000	1.000	1.000	0.995	1.000	0.000
		pseudo SIR	1.000	1.000	1.000	1.000	1.000	1.000	1.000
X2 and X3	**Chi Square**	**SIR**	0.960	0.955	0.955	0.955	0.940	0.950	0.000
		pseudo SIR	0.965	0.955	0.960	0.970	0.940	0.945	0.950
	SED	**SIR**	1.000	1.000	0.990	1.000	1.000	0.995	0.000
		pseudo SIR	1.000	1.000	1.000	1.000	1.000	1.000	1.000

tests of both the usual and pseudo approaches underestimate the dimension because two directions converge to one direction as the correlation between the two important variables increase. In such a case, it is expected that two important variables are indistinguishable.

3.4.5 Real Data

3.4.5.1 Prostate cancer data

The prostate cancer data have been studied many times in the literature (Stamey et al., 1989; Tibshirani, 1996; Zou and Hastie, 2005; Fu, 1998). A total of 97 male patients aged from 41 to 79 have been examined to study the association between the level of prostate specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy. There are eight predictors including log cancer volume (lcavol), log prostate weight (lweight), age of patient, log of benign prostatic hyperplasia amount (lbph), presence or absence of seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason grade (gleason), and percent Gleason grade 4 or 5 (pgg45). The response variable is the log prostate specific antigen level (lpsa). These data have been studied by Fu (1998) and Tutz and Binder (2007) for ridge regression, and have also been analyzed by many others for variable selection purposes (Tibshirani, 1996; Zou and Hastie, 2005). Before fitting an OLS or ridge regression, we should always validate the assumption of fitting a linear model. Using the method described in Section 3.2.3 to test the dimensionality of the underlying model, we have $\hat{d} = 1$ which validates the single index assumption. A plot of the response versus the first direction of SIR in Figure 3.4 shows a clear linear trend. The plot shows that a curve using a spline with 5 degrees of freedom, a curve using a LOWESS smoothing and the OLS line agree with each other well. Furthermore, we fitted a 3rd degree polynomial regression. It turned out that only the linear term is significant. Therefore, this evidence supports that fitting a linear model is sufficient. The advantage of using a pseudo SDR method is that it doesn't assume a parametric model. For variable selection purpose, we compare our pseudo SVS result to LASSO. We compute the prediction error based on leave-

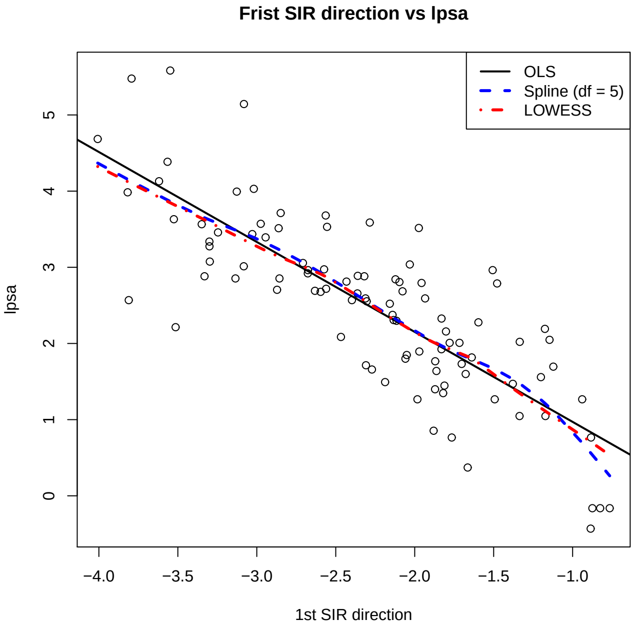

Figure 3.4: Scatter plot of lpsa vs the 1^{st} SIR direction

one-out cross validation for each method, as well we the mean squared error based on using all 97 observations.

The third column in Table 3.6 shows that the pseudo estimate consistently outperforms all other methods. As a result of variable selection, LASSO excludes lcp and gleason from the model while pseudo SVS excludes lcp, gleason and pgg45 from the model. However, the estimated coefficient for pgg45 is very small in LASSO estimate ($\hat{\beta}_{pgg45} = 0.0024$). Therefore, these two variable selection methods mostly agree with each other.

3.4.5.2 Managerial role performance data

We use managerial role performance data to illustrate how pseudo estimates work when the predictors already contain measurement error. The dataset was studied by Warren et al. (1974), Fuller (1987) and Li and Yin (2007). The data contain a random sample of 98 Iowa farmers whose role performance as managers are measured as the response variable (Y). The four predictors are: knowledge of the economic phases of management directed toward profit-making (X_1), tendency to rationally evaluate means to an economic end (X_2), gratification obtained (X_3), and amount of formal education (X_4). The first three predictors and the response variable are measured with questionnaires filled out by the managers and contain measurement errors. The amount of formal education is measured without error. Since the predictors in this data contain measurement errors, we will compare results using pseudo methods to the results using the **U** approach proposed by Li and Yin (2007). Based on Li and Yin (2007), we can obtain both $\Sigma_{\mathbf{w}}$ and Σ_δ. Therefore, we can directly obtain **U** scale predictors as $\mathbf{U} = (I - \hat{\Sigma}_\delta \hat{\Sigma}_{\mathbf{w}}^{-1})\mathbf{W}$. Our pseudo SIR method agrees with Li and Yin (2007) that the estimated structural dimension is 1. To obtain pseudo estimate, we set $l = 0.0005$ and $u = 0.005$ and then ensemble over 500 estimates. A plot of the response versus the respective estimated predictor variable (not reported here) shows significant linear relationship. By fitting a linear model, the last column in Table 3.6 reports the prediction error using leave-one-out cross-validation. Again, it indicates that the two pseudo estimates are very consistent and have better accuracy than the corresponding **U** approach of Li and Yin (2007). For this data, we also use LASSO and pseudo SVS for variable selection. But neither approach excludes any variable from the model, indicating that all the predictors are related to the response variable.

Table 3.6: PE for prostate cancer data and managerial role performance data

	Prostate cancer		**Managerial role performance**	
	MSE	**PE**		**PE**
Pseudo	0.444	0.532	**U-SIR**	0.0155
OLS	0.444	0.541	**U-OLS**	0.0161
Lasso	0.452	0.549	**pseudo SIR**	0.0152
Ridge	0.445	0.537	**pseudo SIR**	0.0152

3.5 Discussion

The pseudo SDR methods introduced in this article can be applied to a broad range of applications in which the predictors are correlated. Although the pseudo SDR directions are laying in a subspace that is different from the true parameter subspace in the population, it can lead to a better estimate based on an observed sample. This relaxation on the parameter space opens the door to search for a better estimate in a much larger space. Our study showed that regardless of whether the observed predictors contain measurement errors or not, the pseudo estimates can improve the classic estimates. In this paper, we only focus on collinearity. However, we can extend the pseudo SDR to the "large p, small n" problem, as its sample covariance matrix is not invertible either. Such a consideration is under current investigation.

It is perhaps interesting and intuitively correct if we relax the positiveness of λ, especially in terms of measurement error data. In highly correlated data, Hoerl and Kennard (1970) showed that there exists a $\lambda > 0$ such that the ridge estimator is better than the OLS estimator. When we directly adopt the result of Li and Yin (2007), it is obvious that $\lambda > 0$, as the covariance matrix of $\boldsymbol{\delta}$ is positive definite. However, suppose that the true variables in $\mathbf{X}$ are highly correlated, but $\mathbf{X}$ already has measurement error in the observations. Then the covariance matrix of $\mathbf{X}$ is as if highly correlated with an error term. In such a case, if this error term is too big, then based on the argument of Hoerl and Kennard (1970), instead of adding a positive λ, we may need to add a negative λ so that the MSE will achieve its minimum. The estimator should not be shrunk but expended. To illustrate this idea, we use the same linear model and same setup on $\mathbf{X}$ in Section 3.2.4.1, instead of working on $\mathbf{U}$, we directly work on measurement error predictor $\tilde{\mathbf{X}} = \mathbf{X} + \boldsymbol{\delta}$, where $\boldsymbol{\delta} \sim N(\mathbf{0}, 0.25\mathbf{I})$. Figure 3.5 shows the box plots for this setup. We use the ridge formula (3.2) on $\tilde{\mathbf{X}}$ and run λ over the range [-1,1]. From there we choose the λ to be the optimal if it has the smallest measure of accuracy for the ridge estimate. We run 500 simulations for each setup. We can see that when measurement errors are already contained in the observed sample, the optimal tuning parameter is around a negative number. Its range is becoming smaller as the correlation among predictors increases. Hence, in order to find a better pseudo estimate, we

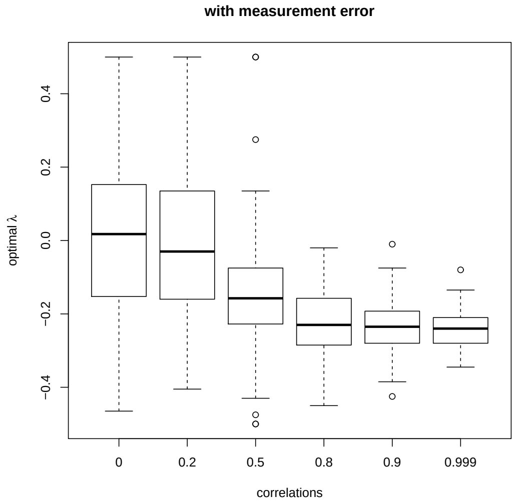

Figure 3.5: Box plot of optimal ridge tuning parameter using $\tilde{\mathbf{X}}$.

may relax the positiveness of the tuning parameter so that λ can be negative. Unfortunately, this approach fails in practice because the method is too sensitive with respect to choosing the “correct” value of the measurement error. In other words, the results are not stable in the sense that they cannot consistently improve the classical estimates.

Appendix

Proof of Proposition 1. If $\gamma = 0$ and $\Gamma = \mathbf{I}$, the relationship in (3.3) reduces to $\mathbf{W} = \mathbf{X}+\boldsymbol{\delta}$. In such a case, $E[\mathbf{W}] = E[\mathbf{X}] + E[\boldsymbol{\delta}] = 0$ and

$$\begin{aligned}\Sigma_{\mathbf{w}} &= E[\mathbf{W}\mathbf{W}'] = E[(\mathbf{X}+\delta)(\mathbf{X}+\delta)'] = E[\mathbf{X}\mathbf{X}' + \delta\mathbf{X}' + \mathbf{X}\delta' + \delta\delta'] \\ &= E[\mathbf{X}\mathbf{X}'] + E[\delta\mathbf{X}'] + E[\mathbf{X}\delta'] + E[\delta\delta'] = E[\mathbf{X}\mathbf{X}'] + E[\delta\delta'] \\ &= \Sigma_{\mathbf{x}} + \Sigma_{\delta}, \qquad (3.16)\end{aligned}$$

because $\mathbf{X}$ and δ are independent so that $E[\delta\mathbf{X}'] = E[\mathbf{X}\delta'] = 0$. Moreover, since δ is also independent of Y,

$$E[\mathbf{W}Y] = E[(\mathbf{X}+\delta)Y] = E[\mathbf{X}Y] + E[\delta Y] = E[\mathbf{X}Y]. \qquad (3.17)$$

For the surrogate predictors $\mathbf{W}$, using (3.16) and (3.17), we can obtain the OLS estimate based on $\mathbf{W}$ as

$$\begin{aligned}\boldsymbol{\beta}^o_{\mathbf{w}} &= (E[\mathbf{W}\mathbf{W}'])^{-1}E[\mathbf{W}Y] = (\Sigma_{\mathbf{x}} + \Sigma_{\delta})^{-1}E[\mathbf{W}Y] \\ &= (\Sigma_{\mathbf{x}} + \Sigma_{\delta})^{-1}E[\mathbf{X}Y]\end{aligned}$$

If $\Sigma_{\delta} = \lambda\mathbf{I}$, $\boldsymbol{\beta}^o_{\mathbf{w}} = \boldsymbol{\beta}^r_{\mathbf{x}}$. □

Chapter 4

Extrinsic sufficient dimension reduction

4.1 Introduction

Among many active research areas in statistics, the “small n, large p” problem attracts more and more attention with emergence of neuroimage data and micro-array data. Nowadays it is common to have a dataset with only a few observations but hundreds of predictors or covariates. If a underlying linear structure is assumed, the model is usually of the form:

$$Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{4.1}$$

where Y is an $n \times 1$ response vector, $\mathbf{X}$ is an $n \times p$ design matrix, $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector, and $\boldsymbol{\varepsilon}^T = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$ are the i.i.d. random noises with $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$. When a dataset is observed with many predictors, $\boldsymbol{\beta}$ is often sparse in a sense that $s < p$ components of it are non-zero. To estimate the structure, one goal is to infer the set of non-zero components of $\boldsymbol{\beta}$ which we denote by $S = \{i : \beta_i \neq 0\}$.

Penalized approach has become a main stream of research to infer the set S from the data (Tibshirani, 1996; Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006). With a proper amount of regularization, L_1 penalty has shown its usefulness in many applications. However, how to select the proper amount of regularization remains a controversial problem. To conquer the challenge, Meinshausen and Bühlmann (2010) proposed a stability selection by using a subsampling approach. Stability selection conservatively controls a certain familywise type I error rate in multiple testing for finite sample size. In conjunction with penalized estimation, stability selection has established a new usage of resampling in studying high dimensional data beyond its traditional role as a tool for asymptotic statistical inference in terms of standard errors, confidence intervals and statistical testing. An extension of stability selection to sufficient dimension reduction (SDR) methods was studied by Wu and Yin (2015b). Let $\mathbf{B}$ be a $p \times d$ matrix such that

$$Y \perp\!\!\!\perp \mathbf{X} | \mathbf{B}^T \mathbf{X} \tag{4.2}$$

where $\perp\!\!\!\perp$ stands for independence. The space spanned by the columns of $\mathbf{B}$ is called a SDR subspace. If the intersection of all the SDR subspaces itself is a SDR subspace, it is called the central subspace (CS, Cook 1994, 1996). Inference in SDR are focused on estimating the basis matrix $\mathbf{B}$ of the CS. Typical estimators are usually obtained by using dimension reduction matrices. These methods include well-known sliced inverse regression (SIR; Li 1991), principal Hessian directions (PHD; Li 1992) and sliced inverse variance estimate (SAVE; Cook and Weisberg 1991). Rich literatures on SDR methods are available (Cook, 1998; Li, 1991, 1992; Cook and Weisberg, 1991; Yin, 2010).

In this paper, we take a completely different approach to effectively use observed sample to facilitate estimating the underlying structure for high dimensional data. Instead of bootstrapping or subsampling, we create surrogate samples by adding small amount of random noises to the original sample so that we are not limited to the original observed data. This idea is related to measurement error regression (Fuller, 1987) and surrogate samples (Li and

Yin, 2007). Since we create surrogate samples of data points which differ from the original observed data, we call it an extrinsic sampling method, comparing with the traditional subsampling or bootstrap which we call an intrinsic sampling method. In Section 4.2, we define the intrinsic and extrinsic sampling schemes and discuss their differences and connections. We also propose a way of obtaining extrinsic samples by adding noises with known structure to the original sample. In Section 4.3, we discuss how to use extrinsic sampling to obtain better penalized estimates for a $n < p$ problem. In Section 4.4, we propose an innovative approach to expand the original sample by using extrinsic sampling to increase the effective sample size such that we can make a $n < p$ problem become a $n > p$ problem. Numerical studies will be included in Section 4.5. We will conclude with a discussion in Section 4.6.

4.2 Extrinsic and intrinsic sampling

Re-sampling is a useful method to estimate known population parameter by reusing available data. Jackknife was proposed by Tukey (1958) based on the idea of Quenouille (1949, 1956) to estimate the bias and standard error (variance) of a statistic. Efron (1979) proposed the bootstrap method as a generalization of jackknife. Bootstrap can be used to estimate the sampling distribution of an estimator by resampling with replacement. Both parametric and nonparametric bootstrap exist under different assumptions. When the parametric inference is complicated or impossible for the calculation of standard errors, bootstrap is particularly useful as a robust alternative to estimate the standard errors. Both bootstrapping and jackknife can be used to study the asymptotic behavior of regression coefficients by estimating standard errors, building confidence intervals, and conducting statistical testing. In-depth study of bootstrap estimate of regression coefficients can be found in Bickel and Freedman (1981) and Freedman (1981).

One limitation of resampling with or without replacement is that the resulting sample has a smaller effective sample size than the original sample. Since every unit in a sample of size n has probability $1-(1-1/n)^n$ to appear in a sufficient bootstrap resample. So, the expected length of a sufficient bootstrap resample is $n^* = [1-(1-1/n)^n] \times n$. For example, if $n = 400$, then $n^* = 253$. This drawback of traditional resampling amplifies as the original sample size decreases and the correlation between predictors increases. Hence, we propose an alternative extrinsic resampling approach.

4.2.1 Differences between extrinsic and intrinsic resampling

If $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n\}$ is the original sample with n observations, aforementioned resampling methods draw a sample $\mathcal{X}^* = \{\mathbf{X}_1^*, \mathbf{X}_2^*, \cdots, \mathbf{X}_n^*\}$ from $\mathcal{X}$ such that $\mathbf{X}_i^* \in \mathcal{X}$. We call any resampling method $\boldsymbol{\mathcal{S}}^I$ that satisfy above property an Intrinsic Resampling (IR) method. Bootstrap is an IR scheme with the functionality of $\boldsymbol{\mathcal{S}}^I$ to be sample with replacement and subsampling is also an IR method with the functionality of $\boldsymbol{\mathcal{S}}^I$ to be sample without replacement.

We now propose a different resampling method as follows. Let $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n\}$ be the original sample, an Extrinsic Resampling (ER) method $\boldsymbol{\mathcal{S}}^E$ generates a sample $\mathcal{X}^* = \{\mathbf{X}_1^*, \mathbf{X}_2^*, \cdots, \mathbf{X}_n^*\}$ that is related to the original sample $\mathcal{X}$, say, $\mathbf{X}_i^* = \mathbf{X}_i + \boldsymbol{\delta}_i$, where $\boldsymbol{\delta}_i$ follows a distribution of F_δ, and $\mathbf{X}_i^*$ may or may not be in the original sample $\mathcal{X}$. Although different methods can be used to obtain extrinsic samples $\mathcal{X}^*$, its relationship with the original sample $\mathcal{X}$ is very important because the principle of obtaining extrinsic samples $\mathcal{X}^*$ is similar to obtaining the intrinsic samples, that is to ensure the behavior of estimates based on the resamples are similar to the estimates that are based on the original sample. However, since the observations in the extrinsic samples are different from the original sample, the estimates based on extrinsic samples may contain some bias.

4.2.2 Obtaining extrinsic samples

Our method to achieve such ER is to create surrogate samples $\mathbf{W}$ by enforcing known noises to the original sample, which creates samples containing "measurement error" $\boldsymbol{\delta}$ with known distribution. There are many references and techniques available in the literature (Fuller, 1987) to study linear models for measurement error predictors. Usually, $\mathbf{W}$ and $\mathbf{X}$ are assumed to be related through a linear model:

$$\mathbf{W} = \gamma + \Gamma\mathbf{X} + \boldsymbol{\delta}, \tag{4.3}$$

where γ is a q-dimensional vector and Γ is a $q \times p$ matrix, which can be either known or unknown in different cases. And $\boldsymbol{\delta}$ is a q dimensional error vector which is independent of $\mathbf{X}$ and Y. The relationship between the response Y and the true predictors $\mathbf{X}$ can be found through the measurement error predictors $\mathbf{W}$. A typical approach is transforming $\mathbf{W}$ to $\mathbf{U}$ as $\mathbf{U} = \Sigma_{\mathbf{xw}}\Sigma_{\mathbf{w}}^{-1}\mathbf{W}$, where $\Sigma_{\mathbf{xw}}$ is the covariance matrix between $\mathbf{X}$ and $\mathbf{W}$. This is essentially a linear regression of $\mathbf{X}$ on $\mathbf{W}$. An auxiliary sample which provides information about the relation between the original predictor $\mathbf{X}$ and the surrogate predictor $\mathbf{W}$ is usually available to estimate $\Sigma_{\mathbf{xw}}$. After this linear transformation, one may proceed with the analysis as if the errors are free in $\mathbf{U}$. Extensive studies of estimation based on surrogate samples related to SDR can be found in Carroll and Li (1992); Lue (2004); Li and Yin (2007) and Wu and Yin (2015a).

An important case of (4.3) is when $p = q$ and $\Gamma = \mathbf{I}$. In such a case, if we generate $\boldsymbol{\delta}$ from a known distribution (e.g. N(0, $\sigma^2\mathbf{I}$)), a new sample can be created by ER scheme as $\mathbf{X}^* = \mathbf{X} + \boldsymbol{\delta}$. Let $\mathbf{z}_i = \{Y_i, \mathbf{X}_i\}$ for $i = 1, \cdots, n$ be the originally observed sample of size n, an extrinsic sample of size n can be drawn as $\mathbf{z}_i^* = \{Y_i, \mathbf{X}_i^*\}$ with $\mathbf{X}_i^* = \mathbf{X}_i + \boldsymbol{\delta}_i$ where $\boldsymbol{\delta}_i \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ and σ^2 is a small constant for $i = 1, \cdots, n$.

4.3 Penalized extrinsic sampling approaches

In this section, we will introduce how extrinsic samples can be used to obtain a stable sparse estimation when the original sample size is small and correlations between predictors are high. One benefit of obtaining extrinsic samples following the way described in Section 4.2.2 is to reduce the estimation bias for highly correlated predictors. Wu and Yin (2015a) studied the underlying connection between measurement error regression and ridge regression. They showed that estimation based on the measurement error predictors as pseudo estimates are better than using original predictors when predictors are highly correlated. Meanwhile, ER approach can also be useful in penalized methods. We follow exactly the same idea of pseudo estimates which uses selected variables of $\mathbf{X}^*$ as the selected variables of $\mathbf{X}$.

Meinshausen and Bühlmann (2010) proposed stability selection by using subsamples of $\lfloor n/2 \rfloor$ to achieve stable variable selection. The idea of stability selection is to select variables based on their empirical selection probability which is the frequency of each variable that been selected using penalized procedure for each subsample. The selection probability for a set $K \subseteq \{1, ..., p\}$ and selected tuning parameter λ is defined to be

$$\hat{\Pi}_K^\lambda = P^*\{K \subseteq \hat{S}^\lambda(\mathcal{I})\}. \tag{4.4}$$

For a given cutoff probability π_{thr} and a tuning parameter range Λ_0, the set of stable predictors is

$$\hat{S}^{stable} = \{k : \max_{\lambda \in \Lambda_0}(\hat{\Pi}_k^\lambda) \geq \pi_{thr}\}. \tag{4.5}$$

Meinshausen and Bühlmann (2010) suggested the subsample size to be $\lfloor n/2 \rfloor$ and a reasonable range for π_{thr} to be $[1/2,\ 1)$. They showed that the stability selection results are little sensitive to the choices of π_{thr}. To select variables for each subsample, Meinshausen and Meinshausen and Bühlmann (2010) proposed to use randomized LASSO be solving the penalized optimization problem,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\| Y - \mathbf{X}\boldsymbol{\beta} \|^2 + \lambda \sum_{i=1}^{p} \frac{|\beta_i|}{w_i} \right), \tag{4.6}$$

where w_i are generated from Uniform$[u, 1]$ for some $u \in (0, 1)$. Wu and Yin (2015b) extend stability selection idea to a general nonlinear model setting of SDR.

However, an obvious limitation of using $\lfloor n/2 \rfloor$ as the sample size for the subsamples is that if the original sample size n is very small, the variance of estimates based on each subsample will be very large so that the final variable selection results may not be accurate. Moreover, if the correlations between predictors are high, original stability selection method may fail to select the correct sets of variable as a known pitfall of LASSO. Therefore, using ER in stability selection brings advantages in two situations when the original method fails. One is that the original sample size is small, and the other is that the correlations among predictors are relatively high. To implement ER in the stability selection, following algorithm can be used:

Algorithm 3 (**Extrinsic stability selection**)

1. *For the observed sample $\boldsymbol{z}_i = \{Y_i, \boldsymbol{X}_i\}$ for $i = 1, \cdots, n$, create N extrinsic samples of size n as $\boldsymbol{z}^*_{bi} = \{Y_i, \boldsymbol{X}^*_i\}$ with $\boldsymbol{X}^*_i = \boldsymbol{X} + \boldsymbol{\delta}_i$ where $\boldsymbol{\delta}_i \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ for $i = 1, \cdots, n$ and $b = 1, \cdots, N$.*

2. *For each extrinsic sample $\boldsymbol{z}^*_{bi}$, obtain the sparse estimated regression coefficient $\hat{\boldsymbol{\beta}}^*_b = (\hat{\beta}^*_{b1}, \cdots, \hat{\beta}^*_{bp})$ by solving (4.6).*

3. *For any predictor k, if its frequency of appearing in $\hat{\boldsymbol{\beta}}^*_b$'s is higher than the preset cutoff probability π_{thr}, then $\boldsymbol{X}^*_b$ is kept in the variable selection procedure, otherwise excluded.*

4. *We select $\boldsymbol{X}_b$ as an active variable if $\boldsymbol{X}^*_b$ is selected.*

4.4 Data augmentation by ER

We should note that for linear models, when $n < p$, traditional penalized method, such as LASSO (Tibshirani, 1996), still work for original sample $\mathbf{X}$, but using extrinsic samples can improve the results when predictors are highly correlated. For dimension reduction, the stable method proposed by Wu and Yin (2015a) does not work on $\mathbf{X}$ for $n < p$ case as their methods require an inverse of the covariance matrix of $\mathbf{X}$. In this section, we will discuss how to increase the effective sample size for estimation by ER approach.

Following the procedure described in Section 4.2.2, we can repeatedly obtain extrinsic samples $\tilde{\mathcal{X}}$ so that when we pool them together, we can obtain an extrinsic sample with a much larger sample size than the original $\mathcal{X}$. Therefore, for a $n < p$ problem, we can always obtain an extrinsic sample $\mathcal{X}^*$ with sample size $n^* > p$. Then we can apply traditional method to obtain an estimate. This approach transforms an unsolvable problem to a solvable problem. Moreover, since multiple extrinsic samples of size n^* can be generated, an ensemble approach can be used to improve the accuracy of estimation.

Algorithm 4 (**Estimation based on data augmentation**)

1. *For the observed sample $\boldsymbol{z}_i = \{Y_i, \boldsymbol{X}_i\}$ for $i = 1, \cdots, n$, create m extrinsic samples of size n as $\boldsymbol{z}^*_{bi} = \{Y_i, \boldsymbol{X}^*_i\}$ with $\boldsymbol{X}^*_i = \boldsymbol{X} + \boldsymbol{\delta}_i$ where $\boldsymbol{\delta}_i \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ for $i = 1, \cdots, n$ and $b = 1, \cdots, m$.*

2. *Combine m extrinsic samples together to obtain an augmented extrinsic sample $\tilde{\boldsymbol{z}}^* = \{\boldsymbol{z}^*_{1i}, \boldsymbol{z}^*_{2i}, \cdots, \boldsymbol{z}^*_{mi}\}$. Then obtain an ordinary estimate $\hat{\boldsymbol{\beta}}^*_j$ based on the augmented sample $\tilde{\boldsymbol{z}}^*$.*

3. *Repeat above steps N times to obtain N estimates $\hat{\boldsymbol{\beta}}^*_j$ for $j = 1, \cdots, N$ based on N enlarged extrinsic samples.*

4.

- *A: To obtain nonsparse estimate, ensemble method can be used to obtain the eigenvectors of* $\boldsymbol{M} = \sum_{i=1}^{N} \hat{\boldsymbol{\beta}}_j^*(\hat{\boldsymbol{\beta}}_j^*)'$.
- *B: To obtain sparse estimate, for any predictor k, we can obtain its empirical quantiles based on the N estimates. If* $\hat{\boldsymbol{\beta}}_{\alpha/2}^* < 0 < \hat{\boldsymbol{\beta}}_{1-\alpha/2}^*$, *where* $\hat{\boldsymbol{\beta}}_{\alpha/2}^*$ *is the* $(100 \times \alpha/2)^{th}$ *percentile of the N estimates* $\hat{\boldsymbol{\beta}}_j^*$, *we set all the* k^{th} *element of* $\hat{\boldsymbol{\beta}}_j^*$ *to be 0. Then the eigenvectors of* $\boldsymbol{M} = \sum_{i=1}^{N} \hat{\boldsymbol{\beta}}_j^*(\hat{\boldsymbol{\beta}}_j^*)'$ *will be used as our final estimate.*

In Step 4, our final estimates as based on $\mathbf{X}^*$ for both sparse and nonsparse approaches. This idea is closely related to the pseudo estimate that introduced by Wu and Yin (2015a). Although the estimate based on each extrinsic sample may contain a small bias, we treat them as the estimates for the correct parameter space. For linear models, OLS can be used and the first eigenvector $\hat{\boldsymbol{\beta}}^*$ of $\mathbf{M}$ can be used as the scaled final estimate. For nonlinear models, traditional SDR methods can be used and the first d eigenvector $\hat{\boldsymbol{\beta}}^*$ of $\mathbf{M}$ can be used as the basis for CS, where d is the structural dimension of the CS that is known or can be estimated. In Step 4B, we can also use a hard threshold that is similar to Shao and Deng (2012) or replace $\hat{\boldsymbol{\beta}}_{\alpha/2}^*$ and $\hat{\boldsymbol{\beta}}_{1-\alpha/2}^*$ by $\bar{\boldsymbol{\beta}}^* \pm \Phi^{-1}(1-\alpha/2)\text{S.E.}^*(\boldsymbol{\beta}_j^*)$ where $\text{S.E.}^*(\hat{\boldsymbol{\beta}}_j^*) = \sum_{b=1}^{N} \frac{(\hat{\boldsymbol{\beta}}_j^* - \bar{\boldsymbol{\beta}}^*)^2}{(N-1)}$, $\bar{\boldsymbol{\beta}}^*$ is the average of $\hat{\boldsymbol{\beta}}_j^*$, and $\Phi(\cdot)$ is the c.d.f. of standard normal distribution. The way of obtaining sparse estimate in Step 4B is very different from the way in Step 3 of Algorithm 3. As mentioned earlier, Algorithm 3 can only be used for OLS when $n < p$, but Algorithm 4 can be used to obtain sparse dimension reduction estimates when $n < p$ because the sample covariance matrix of the predictors in the augmented sample becomes invertible. Since the enlarged sample makes the traditional estimation methods become available, one can also adopt stable approach (Wu and Yin, 2015b) for each augmented sample in Step 4B to obtain sparse estimate. Similarly, we can repeat the entire Algorithm 4 multiple times to obtain a more stable sparse estimation result.

4.5 Numerical study

In this section, we will use simulations to illustrate the effectiveness of our proposed methods. We will use the following model from Zou (2006):

$$\text{Model 1:} \qquad Y = 3X_1 + 1.5X_2 + 2X_5 + c\varepsilon, \tag{4.7}$$

to compare the results of stability selection using ER, subsampling as originally proposed by Meinshausen and Bühlmann (2010), and bootstrap sampling p to illustrate the method proposed in Section 4.3.

To demonstrate the method in Section 4.4 using SIR, we consider the the following nonlinear model from Li (1991):

$$\text{Model 2:} \qquad Y = X_1(X_2 + 1.5) + c\varepsilon. \tag{4.8}$$

For both models, the predictors $\mathbf{X}_i = (X_{i1}, X_{i2}, \cdots, X_{ip})'$ for $i = 1, \cdots, n$ were i.i.d. standard multivariate normal vectors, as well as the noise vector $\boldsymbol{\varepsilon}$. We set the pairwise correlation between X_k and X_j to be $\text{cor}(X_k, X_j) = \rho^{|k-j|}$, in which we vary the correlation coefficient between 0 and 0.9. The noise level c is set to be 0.1. Since both methods introduced in Section 4.3 and Section 4.4 provide sparse estimates for variable selection purposes, we report the true positive rate (TPR): the ratio of the number of correctly identified active predictors to the number of truly active predictors, and the false positive rate (FPR): the ratio of the number of falsely identified active predictors to the number of true inactive predictors. A better estimate should have bigger TPR and smaller FPR. Following the algorithms in Section 4.3 and Section 4.4, when create the extrinsic samples, we set the variance of the noise to be 0.005.

Table 4.1: Stability selection with extrinsic sampling, subsampling, and bootstrap

		Extrinsic Sampling		**Sub-sampling**		**Bootstrap**	
p	ρ	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**
10	**0**	1.0000	0.0143	1.0000	0.0014	0.9900	0.0000
	0.2	1.0000	0.0114	1.0000	0.0000	0.9967	0.0000
	0.5	1.0000	0.0057	1.0000	0.0000	0.9933	0.0000
	0.8	0.9967	0.0300	0.9400	0.0029	0.8533	0.0000
	0.9	0.9467	0.0486	0.8100	0.0071	0.6600	0.0014
50	**0**	1.0000	0.0036	0.9867	0.0000	0.7067	0.0000
	0.2	1.0000	0.0017	0.9933	0.0000	0.7900	0.0000
	0.5	1.0000	0.0023	0.9967	0.0000	0.9333	0.0000
	0.8	1.0000	0.0034	0.9400	0.0002	0.8567	0.0000
	0.9	0.9667	0.0064	0.6933	0.0004	0.5767	0.0000
100	**0**	1.0000	0.0016	0.9300	0.0000	0.4167	0.0000
	0.2	0.9967	0.0012	0.9733	0.0000	0.5933	0.0000
	0.5	1.0000	0.0008	0.9967	0.0000	0.8433	0.0000
	0.8	0.9833	0.0014	0.8967	0.0001	0.7800	0.0000
	0.9	0.9333	0.0029	0.7133	0.0000	0.5900	0.0000
300	**0**	0.9900	0.0006	0.7367	0.0000	0.2067	0.0000
	0.2	1.0000	0.0004	0.8067	0.0000	0.2333	0.0000
	0.5	0.9933	0.0003	0.9400	0.0000	0.5767	0.0000
	0.8	0.9967	0.0005	0.8967	0.0001	0.7533	0.0000
	0.9	0.9067	0.0006	0.7233	0.0003	0.5767	0.0000

Table 4.1 provides TPRs and FPRs of using stability selection for model (4.7). Since the effectiveness of using extrinsic samples in stability becomes significant when the original sample size is relatively small and the correlations between predictors are relatively high, we fix the samples size of each data that we generated to be $n = 40$ and set correlations ρ between variables to be $(0, 0.2, 0.5, 0.8, 0.9)$ respectively. To simulate $n < p$ scenario, we set the number of predictors to be $(10, 50, 100, 300)$ respectively with $p = 10$ for comparison purpose. The TPRs and FPRs are averaged over 500 simulated data for each situation.

From the output, we can see that when the correlation between predictors increase, using ER in stability selection maintain a good performance while using subsampling and bootstrap become insufficient especially as the number of predictors also increase. In such a case, TPRs of subsampling and bootstrap sampling significantly decrease. Stability selection using ER outperform the other two sampling methods in almost all (n, p) combinations. We should also note that, the reason for subsampling and bootstrap having small FPRs is because when p gets large, both methods fail to select any variable. This is why the TPR and FPR decrease the same time.

The data augmentation approach can be used for obtaining both sparse OLS and SIR estimates when $n < p$. For linear model (4.7) using OLS, since LASSO can also solve the $n < p$ problem, we compare the TPRs and FPRs for LASSO with data augmentation approach as described in Section 4.4. We also include the TPRs and FPRs by using stable data augmentation approach. To obtain a stable estimate based on data augmentation, one can simply repeat the steps in Algorithm 4 T times to obtain T sparse estimate. To obtain a stable estimate, for any predictor k, if the proportion of k^{th} elements among the T estimates is smaller than a preselected cutoff point π, we set all the k^{th} element of $\hat{\boldsymbol{\beta}}_j^*$ to be 0. Then the eigenvectors of $\mathbf{M} = \sum_{i=1}^{T} \hat{\boldsymbol{\beta}}_j^* (\hat{\boldsymbol{\beta}}_j^*)'$ will be used as our final stable estimate. In our simulations, we set $\pi = 0.85$.

Table 4.2 includes the TPRs and FPRs of LASSO estimate and sparse estimate based on data augmentation using extrinsic samples and stable estimate as described above. We fix the samples size of each data we generated to be $n = 80$ and set correlations between variables ρ to be $(0, 0.2, 0.5, 0.8, 0.9)$ respectively. To simulate $n < p$ scenario, we set the number of predictors to be $(10, 40, 80, 120, 160)$ respectively with $p = 10$ and $p = 40$ for comparison purpose. The TPRs and FPRs are averaged over 500 simulated data for each situation. According to the results, we can see that when the number of predictors is relatively small, the data augmentation approaches are significantly better than LASSO in terms of FPRs. As the number of predictors increase, data augmentation approaches consistently remain good performance and are as good as LASSO. We should keep in mind that since the data augmentation approach (without the stabilizing step) is computationally much simpler than LASSO, it is very effective to provide variable selection for $n < p$ problem. We should

Table 4.2: TPR and FPR of variable selection for linear model

		LASSO		**Extrinsic**		**Stable Extrinsic**	
p	**r**	**TPR**	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**
10	**0**	1.0000	0.3875	1.0000	0.0113	1.0000	0.0088
	0.2	1.0000	0.3463	1.0000	0.0138	1.0000	0.0063
	0.5	1.0000	0.2963	1.0000	0.0350	1.0000	0.0200
	0.8	1.0000	0.2113	1.0000	0.1250	1.0000	0.1150
	0.9	1.0000	0.1338	1.0000	0.1688	1.0000	0.1438
40	**0**	1.0000	0.1400	1.0000	0.0234	1.0000	0.0121
	0.2	1.0000	0.1416	1.0000	0.0137	1.0000	0.0055
	0.5	1.0000	0.1213	1.0000	0.0113	1.0000	0.0068
	0.8	1.0000	0.1134	1.0000	0.0239	1.0000	0.0224
	0.9	1.0000	0.0763	1.0000	0.0308	1.0000	0.0279
80	**0**	1.0000	0.1001	1.0000	0.0327	1.0000	0.0174
	0.2	1.0000	0.0801	1.0000	0.0231	1.0000	0.0103
	0.5	1.0000	0.0774	1.0000	0.0085	1.0000	0.0024
	0.8	1.0000	0.0931	1.0000	0.0122	1.0000	0.0112
	0.9	1.0000	0.0865	1.0000	0.0151	1.0000	0.0136
120	**0**	1.0000	0.0564	1.0000	0.1159	1.0000	0.0849
	0.2	1.0000	0.0453	1.0000	0.0972	1.0000	0.0729
	0.5	1.0000	0.0381	1.0000	0.0618	1.0000	0.0426
	0.8	1.0000	0.0297	1.0000	0.0215	1.0000	0.0153
	0.9	1.0000	0.0285	1.0000	0.0217	1.0000	0.0179
160	**0**	1.0000	0.0336	1.0000	0.0587	1.0000	0.0411
	0.2	1.0000	0.0316	1.0000	0.0529	1.0000	0.0362
	0.5	1.0000	0.0265	1.0000	0.0351	1.0000	0.0241
	0.8	1.0000	0.0184	1.0000	0.0159	1.0000	0.0109
	0.9	1.0000	0.0189	1.0000	0.0158	1.0000	0.0141

also notice that although the stability step requires additional computations, it result much smaller FPRs in variable selection.

Table 4.3 provides the TPRs and FPRs for SIR model (4.8). Similar to Table 4.2, the data augmentation approach selects variables with bigger TPRs and smaller FPRs.

Table 4.3: TPR and FPR of variable selection for SIR model

		Extrinsic		**Stable Extrinsic**	
p	ρ	**TPR**	**FPR**	**TPR**	**FPR**
10	**0**	1.000	0.056	1.000	0.031
	0.2	1.000	0.081	1.000	0.044
	0.5	1.000	0.106	1.000	0.056
	0.8	1.000	0.081	1.000	0.069
	0.9	1.000	0.156	1.000	0.119
40	**0**	1.000	0.042	1.000	0.033
	0.2	1.000	0.037	1.000	0.029
	0.5	1.000	0.033	1.000	0.028
	0.8	1.000	0.026	1.000	0.017
	0.9	1.000	0.033	1.000	0.026
80	**0**	1.000	0.017	1.000	0.013
	0.2	1.000	0.018	1.000	0.013
	0.5	1.000	0.016	1.000	0.009
	0.8	1.000	0.011	1.000	0.005
	0.9	1.000	0.014	1.000	0.012
120	**0**	0.975	0.014	0.975	0.012
	0.2	1.000	0.018	1.000	0.015
	0.5	1.000	0.020	1.000	0.017
	0.8	1.000	0.021	1.000	0.020
	0.9	1.000	0.024	1.000	0.020
160	**0**	0.850	0.009	0.900	0.007
	0.2	0.925	0.010	0.900	0.009
	0.5	1.000	0.012	1.000	0.012
	0.8	1.000	0.017	1.000	0.015
	0.9	1.000	0.018	1.000	0.015

4.6 Discussion and future work

In this paper, we propose extrinsic sampling in stability selection for linear models in order to improve the variable selection accuracy for $n < p$ case. Our approach is especially useful for the case when the size of observed sample is limited and predictors are correlated with each other. Our empirical study shows that stability selection using ER outperforms using subsampling as originally proposed by Meinshausen and Meinshausen and Bühlmann (2010). We also propose to pool multiple extrinsic samples together to increase the effective sample size in order to make traditional SDR methods work for $n < p$ problems. With multiple enlarged samples generated, sparse estimates can be obtained based on empirical distribution of estimates from each enlarged sample. Meanwhile, with estimates obtained from multiple extrinsic samples, an ensemble step can improve accuracies of both sparse and nonsparse estimates further.

In our simulations, we pool multiple extrinsic samples together to make sure that the final enlarged sample size is $m = 3$ times larger than the number of predictors p. However, choosing an optimal m could be one of our future investigations. Intuitively, the choice of m should rely on the ratio between n and p. To create each extrinsic sample, we uses a small quantity $\boldsymbol{\delta} = 0.005$ as the noise variance, which follows the suggestion by Wu and Yin (2015a). However, we could also generate $\boldsymbol{\delta}$ from an interval and use ensemble approach in order to avoid using a possible bad $\boldsymbol{\delta}$.

The data augmentation approach creates an opportunity for statisticians to apply traditional methods for a $n < p$ problem. While extra bias are introduced when extrinsic samples are created, one can always refine their estimation methods to seek for a balance between the bias and estimation accuracy because this bias is known and can be controlled by practitioners. We treat the final estimate as a pseudo estimate based on $\mathbf{X}^*$, the theory of linking this pseudo estimate to the estimate based on $\mathbf{X}$ is another future research topic.

Bibliography

Bellman, R. (1961). Adaptive control processes. *Princeton University Press*.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for bootstrap. *The Annals of Statistics*, 9(6):1196–1217.

Carroll, R. J. and Li, K.-C. (1992). Measurement error regression with unknown link: Dimension reduction and data visualization. *Journal of the American Statistical Association*, 87:1040–1050.

Chen, X., Zou, C., and Cook, R. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics*, 38:3696–3723.

Chiaromonte, F. and Martinelli, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. *Mathematical Biosciences*, 176:123144.

Cook, R. (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. *Proceedings of the Section on Physical and Engineering Sciences*, pages 18–25.

Cook, R. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, 91(435):983–992.

Cook, R. (1998). *Regression Graphics: Ideas for studying regressions through graphics*. Wiley: New York.

Cook, R. and Li, B. (2002). Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30(2):455–474.

Cook, R., Li, B., and Chiaromonte, F. (2007). Dimension reduction in regression without matrix inversion. *Biometrika*, 94(3):569–584.

Cook, R. and Weisberg, S. (1991). Discussion of ”sliced inverse regression for dimension reduction”. *Journal of the American Statistical Association*, 86:328–332.

Cook, R. and Zhang, X. (2014). Fused estimators of the central subspace in sufficient dimension reduction. *Journal of the American Statistical Association*, 109:815–827.

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *The Annals of Statistics*, 32(3):1061–1092.

Dempster, A. P., Schatzoff, M., and Wermuth, N. (1977). A simulation study of alternatives to ordinary least squares. *Journal of the American Statistical Association*, 72:77–91.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Freedman, D. A. (1981). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 9(6):1218–1228.

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7:397–416.

Fukumizu, K., Bach, F. R., and Jordan, M. I. (2009). Kernel dimension reduction in regression. *The Annals of Statistics*, 37(4):1871–1905.

Fukumizu, K. and Leng, C. (2014). Kernel dimension reduction in regression. *Journal of the American Statistical Association*, 109(500):359–370.

Fuller, W. A. (1987). *Measurement Error Models.* Wiley: New York.

Gibbons, D. G. (1981). A simulation study of some ridge estimators. *Journal of the American Statistical Association*, 76(373):131–139.

Goldstein, M. and Smith, A. F. M. (1974). Ridge-type estimators for regression analysis. *Journal of the Royal Statistical Society. Series B*, 36(2):284–291.

Hocking, R. R., Speed, F. M., and Lynn, M. J. (1976). A class of biased estimators in linear regression. *Technometrics*, 18(4):425–437.

Hoerl, A. E., Kannard, R. W., and Baldwin, K. F. (1975). Ridge regression: some simulation. *Communications in Statistics*, 4(2):105–123.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Hooper, J. W. (1959). Simulation equations and canonical correlation theory. *Econometrica*, 27(2):245–256.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. on Math. Statist. and Prob.*, pages 361–379.

Kibria, B. M. G. (2003). Performance of some new ridge regression estimators. *Communications in Statistics - Simulation and Computation*, 32(2):419–435.

Lawless, J. F. and Wang, P. (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics - Theory and Methods*, 5(4):307–323.

Li, B., Chun, H., and Zhao, H. (2012). Sparse estimation of conditional graphical models with application to gene networks. *Journal of the American Statistical Association*, 107(497):152–167.

Li, B., Chun, H., and Zhao, H. (2014). On an additive semi-graphoid model for statistical networks with application to pathway analysis. *Journal of the American Statistical Association*, 109(507):1188–1204.

Li, B. and Yin, X. (2007). On surrogate dimension reduction or measurement error regression: an invariance law. *The Annals of Statistics*, 35(5):2143–2172.

Li, B., Zha, H., and Chiaromonte, F. (2005). Contour regression: A general approach to dimension reduction. *The Annals of Statistics*, 33(4):1580–1616.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86:316–342.

Li, K.-C. (1992). On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association*, 87:1025–1039.

Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika*, 94:603–613.

Li, L., Cook, R. D., and Tsai, C.-L. (2007). Partial inverse regression. *Biometrika*, 94(3):615–625.

Li, L. and Li, H. (2004). Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics*, 20:3406–3412.

Li, L. and Yin, X. (2008). Sliced inverse regression with regulations. *Biometrics*, 64:124–131.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B*, 34:1–41.

Lue, H.-H. (2004). Principal hessian directions for regression with measurement error. *Biometrika*, 91(2):409–423.

Luo, W., Li, B., and Yin, X. (2014). On efficient dimension reduction with respect to a statistical functional of interest. *The Annals of Statistics*, 42(1):382–412.

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34:1436–1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society. Series B*, 72:417–473.

Ni, L., Cook, R. D., and Tsai, C.-L. (2005). A note on shrinkage sliced inverse regression. *Biometrika*, 92(1):242–247.

Obenchain, R. L. (1975). Ridge analysis following a preliminary test of the shrunken hypothesis. *Technometrics*, 17(4):431–441.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society. Series B*, 11:68–84.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, 43:353–360.

Shao, J. and Deng, X. (2012). Estimation in high-dimensional linear models with deterministic design matrices. *The Annals of Statistics*, 40:812–831.

Stamey, T., Kabalin, J., McNeal, J., Johnston, I., Freiha, F., Redwine, E., and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate ii. radical prostatectomy treated patients. *Journal of Urologys*, 16:1076–1083.

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

Tukey, J. W. (1958). Notes on bias in estimation. *The Annals of Mathematical Statistics*, 29:614.

Tutz, G. and Binder, H. (2007). Boosting ridge regression. *Computational Statistics & Data Analysis*, 51(12):6044 – 6059.

Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association*, 103(482):811–821.

Wang, Q. and Yin, X. (2008). A nonlinear multi-dimensional variable selection method for high dimensional data: sparse mave. *Computational Statistics & Data Analysis*, 52(9):4512–4520.

Warren, R. D., White, J. K., and Fuller, W. A. (1974). An errors-in-variables analysis of managerial role performance. *Journal of the American Statistical Association*, 69(348):886–893.

Wu, W. and Yin, X. (2015a). Pseudo sufficient dimension reduction and sufficient variable selection. *Manuscript*.

Wu, W. and Yin, X. (2015b). Stable estimation in dimension reduction. *Journal of Computational and Graphical Statistics*, 24(1):104–120.

Xia, Y., Tong, H., Li, W. K., and Zhu, L. X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society. Series B*, 64(3):363–410.

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, 98:968–979.

Yin, X. (2010). Sufficient dimension reduction in regression. In Cai, T. and Shen, X., editors, *High-Dimensional Data Analysis*. World Scientific, New Jersey.

Yin, X. and Cook, R. D. (2002). Dimension reduction for the conditional kth moment in regression. *Journal of the Royal Statistical Society. Series B*, 64:159–175.

Yin, X. and Hilafu, H. (2015). Sequential sufficient dimension reduction for large p, small n problems. *Journal of the Royal Statistical Society. Series B*, In Press.

Yin, X. and Li, B. (2011). Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *The Annals of Statistics*, 39(6):3392–3416.

Yin, X., Li, B., and Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*, 99(8):1733 – 1757.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594.

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

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

Zhou, J. and He, X. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. *The Annals of Statistics*, 36(4):1649–1668.

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

Zhu, L.-P., Yu, Z., and Zhu, L.-X. (2010a). A sparse eigen-decomposition estimation in semiparametric regression. *Computational Statistics & Data Analysis*, 54(476):976–986.

Zhu, L.-P. and Zhu, L.-X. (2009). Dimension reduction for conditional variance in regression. *Statistica Sinica*, 19(4):869–883.

Zhu, L.-P., Zhu, L.-X., and Feng, Z.-H. (2010b). Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association*, 105(492):1455–1466.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B*, 67:310–320.