Sufficient Dimension Reduction and Variable Selection

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Xin Chen

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

R. Dennis Cook, Adviser

November, 2010

Acknowledgements

I would like to thank my adviser, Professor R. Dennis Cook who taught me how to do research in Statistics and guided me throughout the whole process. I was inspired a lot by his great scientific insights and amused by his humors during the period of the supervision under him. I am also thankful for the many years of financial support I was granted in the School of Statistics.

Dedication

To my parents: Yishan Chen and Yuying Kuan

Abstract

Sufficient dimension reduction (SDR) in regression was first introduced by Cook (2004). It reduces the dimension of the predictor space without loss of information and it is very helpful when the number of predictors is large. It alleviates the "curse of dimensionality" for many statistical methods. In this thesis, we study the properties of a dimension reduction method named "continuum regression"; we propose a unified method – coordinate-independent sparse estimation (CISE) – that can simultaneously achieve sparse sufficient dimension reduction and screen out irrelevant and redundant variables efficiently; we also introduce a new dimension reduction method called "principal envelope models".

Contents

Acknowledgements i

Dedication ii

Abstract iii

List of Tables vi

List of Figures vii

1 Introduction 1

2 Some insights into continuum regression 4

2.1 Asymptotic properties of continuum regression 5

2.2 Discussion 8

2.3 Appendix 9

3 Variable selection in sufficient dimension reduction 14

3.1 Theory and Methodology 16

3.1.1 Motivation: generalized eigenvalue problems revisited 16

3.1.2 A coordinate-independent penalty function 18

3.1.3 Coordinate-independent sparse estimation 19

3.1.4 Algorithm 20

3.1.5 Oracle property 21

3.1.6 Choice of tuning parameters 24

3.2 Simulation studies 25

3.3 Boston housing data . . . 28
3.3.1 Variable screening . . . 28
3.3.2 Bootstrap study . . . 30
3.4 A Matlab package . . . 30
3.5 Discussion . . . 30
3.6 Appendix . . . 31

4 Principal envelope model **43**
4.1 Principal envelope model . . . 44
4.1.1 Probabilistic principal component analysis revisited . . . 44
4.1.2 Motivation: general error structure . . . 46
4.1.3 Specific principal envelope models . . . 48
4.1.4 Selection of the dimension u . . . 51
4.1.5 Simulation studies . . . 51
4.2 Feature selection . . . 53
4.2.1 A simple coordinate-independent penalty function . . . 54
4.2.2 Methodology . . . 54
4.2.3 Algorithm . . . 55
4.2.4 Simulation studies . . . 56
4.3 Data analysis . . . 58
4.4 Discussion . . . 59
4.5 Appendix . . . 59

References **65**

List of Tables

3.1 The generalized eigenvalue formulations 16
3.2 Variable selection summary of Study 1 26
3.3 Variable selection summary of Study 2 26
3.4 Variable selection summary of Study 3 27
3.5 Variable selection summary of Study 4 27
3.6 Estimated bases of the central subspace in Boston housing data 29
3.7 Variable selection in bootstrapping Boston housing data 30
4.1 Rates of feature selection from a PPCA model 57
4.2 Rates of feature selection from a principal envelope model 57

List of Figures

2.1 Boxplots of angles when $\boldsymbol{\nu}(y) = y$ 7
2.2 Average angles versus n when $\nu(y) = y$ and $\nu(y) = y + 1.5y^2$ 8
4.1 Average angles versus n when $p = 20$ 52
4.2 Recover the square using PEM and PCA methods 53

Chapter 1

Introduction

Consider the regression of a univariate response y on p random predictors $\mathbf{x} = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$, with the general goal of inferring about the conditional distribution of $y|\mathbf{x}$. When p is large, most statistical methods face the "curse of dimensionality", and thus dimension reduction is desirable. There has been a long history of dimension reduction in the statistical literature. For example, principal components is very popular in multivariate analysis, especially in applied science; partial least squares (Wold, 1975; Helland, 1990) and continuum regression (Stone and Brooks, 1990) are used frequently in Chemometrics.

In this thesis, we work under the framework of the sufficient dimension reduction paradigm. Sufficient dimension reduction (SDR) introduced by Cook (1994; 1998a) is important in both theory and practice. It strives to reduce the dimension of $\mathbf{x}$ by replacing it with a minimal set of linear combinations of $\mathbf{x}$, without loss of information on the conditional distribution of $y|\mathbf{x}$. If a predictor subspace $\mathcal{S} \subseteq \mathbb{R}^p$ satisfies

$$y \perp\!\!\!\perp \mathbf{x}|P_{\mathcal{S}}\mathbf{x},$$

where $\perp\!\!\!\perp$ stands for independence and $P_{(\cdot)}$ represents the projection matrix with respect to the standard inner product, then $\mathcal{S}$ is called a dimension reduction space. The central subspace $\mathcal{S}_{y|\mathbf{x}}$, which is the intersection of all dimension reduction spaces, is an essential concept of SDR. Under mild conditions, it can be shown that $\mathcal{S}_{y|\mathbf{x}}$ is itself a dimension reduction subspace (Cook 1994; 1998a), which we assume throughout this thesis, and then it is taken as the parameter of interest. The dimension d of $\mathcal{S}_{y|\mathbf{x}}$, usually far less

than p, has to be estimated in application. Through this thesis, we assume that the sample size n is larger than p.

There has been considerable interest in dimension reduction methods since the introduction of sliced inverse regression (SIR; Li 1991) and sliced average variance estimation (SAVE; Cook and Weisberg 1991). Li (1992) and Cook (1998b) proposed and studied the method of principal Hessian directions (PHD), and the related method of iterative Hessian transformations was proposed by Cook and Li (2002). Chiaromonte et al. (2002) proposed partial sliced inverse regression for estimating a partial central subspace. Yin and Cook (2002) introduced a covariance method for estimating the central kth moment subspace. Most of these and many other dimension reduction methods are based on the first two conditional moments and as a class are called F2M methods (Cook and Forzani 2009). They provide exhaustive estimation of $\mathcal{S}_{y|\mathbf{x}}$ under mild conditions. B. Li and Wang (2007) proposed another F2M method called directional regression (DR). They argued that DR is more accurate than or competitive with all of the previous F2M dimension reduction proposals. In contrast to these and other moment-based SDR approaches, Cook (2007) introduced a likelihood-based paradigm for SDR that requires a model for the inverse regression of $\mathbf{x}$ on y. This paradigm, which is broadly referred to as principal fitted components (PFC), was developed further by Cook and Forzani (2009). Likelihood-based SDR inherits properties and methods from general likelihood theory and can be very efficient in estimating the central subspace.

In Chapter 2, we first study an ad-hoc dimension reduction method called "continuum regression" in order to demonstrate the importance of SDR. Continuum regression (CR) encompasses ordinary least squares regression, partial least squares regression and principal component regression under the same umbrella using a non-negative parameter γ. However there seems to be no literature discussing the asymptotic properties for arbitrary continuum regression parameter γ. We establish a relation between continuum regression and sufficient dimension reduction and study the asymptotic properties of continuum regression for arbitrary γ under inverse regression models. Theoretical and simulation results show that the continuum seems unnecessary when the conditional distribution of the predictors given the response follows the multivariate normal distribution.

In Chapter 3, we propose a unified method – coordinate-independent sparse estimation (CISE) – that can simultaneously achieve sparse sufficient dimension reduction and screen out irrelevant and redundant variables efficiently. CISE is subspace oriented in the sense that it incorporates a coordinate-independent penalty term with a broad series of model-based and model-free SDR approaches. This results in a Grassmann manifold optimization problem and a fast algorithm is suggested. Under mild conditions, based on manifold theories and techniques, it can be shown that CISE would perform asymptotically as well as if the true irrelevant predictors were known, which is referred to as the oracle property. Simulation studies and a real-data example demonstrate the effectiveness and efficiency of the proposed approach.

Motivated by a general error structure in probabilistic principal component analysis (Tipping and Bishop, 1999) and incorporating the novel idea of an "envelope" proposed by Cook et. al (2010), we construct principal envelope models in Chapter 4 and demonstrate the possibility that any subset of principal components could retain most of the sample's information. The useful principal components can be found through maximum likelihood approaches. We also introduced a new method to select features based on the likelihood function.

Chapter 2

Some insights into continuum regression

Stone and Brooks (1990) introduced the method of continuum regression in which they suggest a spectrum of possible regressors controlled by a non-negative parameter γ that includes ordinary least squares, partial least squares (Helland, 1988, 1990) and principal component regression (Jolliffe, 2002). They also extended continuum regression for multiple predictands (Brooks and Stone, 1994). Continuum regression has been used in chemometrics, especially when the number of predictors is large and dimension reduction is needed.

Suppose we are studying the regression of y on $\mathbf{x}$ with n samples, where $y \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^p$ and $(\mathbf{x}, y)$ has a joint density. The construction rule for the regressors in continuum regression can be expressed as follows. Continuum regression first defines an objective function of $\mathbf{c} \in \mathbb{R}^p$,

$$T(\mathbf{c}) = (\mathbf{c}^T \hat{\mathbf{s}})^2 (\mathbf{c}^T \hat{\mathbf{\Sigma}} \mathbf{c})^{\gamma - 1}, \tag{2.1}$$

where the tuning parameter $\gamma \geq 0$, $\hat{\mathbf{\Sigma}}$ denotes the sample covariance matrix of $\mathbf{x}$ and $\hat{\mathbf{s}}$ denotes the sample covariance vector of $\mathbf{x}$ and y. The definition of $T(\mathbf{c})$ requires that the rank of $\hat{\mathbf{\Sigma}}$ equals p when $\gamma < 1$. A sequence of $\hat{\mathbf{c}}_i$ is obtained by maximizing $T(\mathbf{c})$ under the constraints that $\|\hat{\mathbf{c}}_i\| = 1$ and $\hat{\mathbf{c}}_j^T \hat{\mathbf{\Sigma}} \hat{\mathbf{c}}_i = 0$ for $j = 1, \ldots, i-1$. The first $\hat{\mathbf{c}}_1$ is obtained only under the constraint that $\|\hat{\mathbf{c}}_1\| = 1$. Stone and Brooks showed that $\gamma = 0$ corresponds to ordinary least squares, $\gamma = 1$ to partial least squares and $\gamma \to \infty$

to principal component regression.

Assume that $\omega < p$ regressors are chosen. Then new regressor(s) $\hat{\mathbf{c}}_1^T\mathbf{x}, \ldots, \hat{\mathbf{c}}_\omega^T\mathbf{x}$ are used for the study of the regression. The linear transformation $R(\mathbf{x}) = (\hat{\mathbf{c}}_1 \ldots, \hat{\mathbf{c}}_\omega)^T\mathbf{x}$ serves as a dimension reduction and $\text{span}(\hat{\mathbf{c}}_1 \ldots, \hat{\mathbf{c}}_\omega)$ serves as a dimension reduction space. Stone and Brooks suggested cross validation to select the number of regressors ω. A reduction $R(\mathbf{x})$ is called sufficient if it satisfies $y \perp\!\!\!\perp \mathbf{x}|R(\mathbf{x})$ (Cook, 1998a). In this way, we establish a relationship between continuum regression and sufficient dimension reduction. A very important concept of sufficient dimension reduction is the central subspace $\mathcal{S}_{y|\mathbf{x}}$ which is defined as the intersection of all subspaces $\mathcal{S} \in \mathbb{R}^p$ with the property that y is conditionally independent of $\mathbf{x}$ given the projection of $\mathbf{x}$ onto $\mathcal{S}$.

The Li-Duan Proposition shows that the subspace spanned by the ordinary least squares solution is a consistent estimate of the central subspace when the dimension of $\mathcal{S}_{y|\mathbf{x}}$ equals one and a linearity condition holds (Cook, 1998a). Naik and Tsai (2000) showed that the partial least squares estimator has the same property in single-index models. As far as we know, there is no literature discussing the asymptotic properties for an arbitrary continuum regression parameter γ. In Section 2.1, we study continuum regression with arbitrary γ under the inverse regression models developed by Cook (2007) in which the central subspace is well defined. Our studies suggest that the "continuum" is unnecessary under certain conditions. Concluding remarks can be found in Section 2.2. The technical details are given in the Appendix.

2.1 Asymptotic properties of continuum regression

Cook (2007) suggested inverse regression models to achieve sufficient dimension reduction. Suppose that the conditional distribution of $\mathbf{x}$ given y can be modeled as follows:

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\nu}(y) + \boldsymbol{\Delta}^{1/2}\epsilon, \tag{2.2}$$

where $\boldsymbol{\mu} = E(\mathbf{x})$, $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d}$, $d < p$, $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \mathbf{I}_d$ and d has to be estimated in application but is assumed to be known here. The error vector $\epsilon \in \mathbb{R}^p$ is assumed to be normally distributed with mean 0 and identity covariance matrix and to be independent of y. We use $\boldsymbol{\Delta}$ to denote the conditional covariance matrix $\text{var}(\mathbf{x}|y)$, which is an arbitrary positive definite matrix. The coordinate vector $\boldsymbol{\nu}(y) \in \mathbb{R}^d$ is an unknown function of y

that is assumed to have a positive definite covariance matrix and be centered at 0; that is $E(\boldsymbol{\nu}(y)) = 0$. The mean function $\boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\nu}(y)$ of model (2.2) is very flexible. Although model (2.2) assumes conditional normality, it is quite robust under modest deviations from normality. Cook and Forzani (2008) showed that the central subspace $\mathcal{S}_{y|\mathbf{x}}$ equals $\text{span}(\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma})$, the span of column vector(s) of $\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}$. The dimensionality of $\mathcal{S}_{y|\mathbf{x}}$ equals d. Through this thesis, we assume that $\text{cov}(\mathbf{x}, y) \neq 0$.

Proposition 1 *Suppose model (2.2) is true with* $d = 1$ *(*$\boldsymbol{\Gamma}$ *is a vector) and* $\text{span}(\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}) \neq \text{span}(\boldsymbol{\Gamma})$. *Let* c_1 *denote the population version of* $\hat{\mathbf{c}}_1$. *Then*

(i) $c_1 \neq \boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}/||\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}||_2$ *for* $\gamma > 0$, *and*

(ii) $\hat{\mathbf{c}}_1$ *converges to* $\mathbf{c}_1$ *in probability.*

In other words, $\text{span}(\hat{\mathbf{c}}_1)$ *is an inconsistent estimator of* $\mathcal{S}_{y|\mathbf{x}}$ *for* $\gamma > 0$.

The proof of Proposition 1 is given in the appendix. It tells us that there is no need of the "continuum" if we believe the data can be modeled by (2.2) with $d = 1$ and $\text{span}(\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}) \neq \text{span}(\boldsymbol{\Gamma})$ because only the ordinary least squares estimator ($\gamma = 0$) is consistent for the central subspace under this scenario. A small simulation study was conducted to illustrate the inconsistency. We generate 500 data sets from model (2.2) with $p = 10$, $\boldsymbol{\mu} = 0$, $\boldsymbol{\nu}(y) = y$, $\boldsymbol{\Gamma} = (1, \ldots, 1)^T/\sqrt{10}$, $\Delta_{ij} = 0.5^{|i-j|}$ for $1 \leq i, j \leq 10$ and y and ϵ following a standard normal distribution. Figure 2.1 shows boxplots of angles between the true direction $\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}$ and the continuum regression estimators versus n for $\gamma = 0.5$. As n increases, the median angle levels out and is always above 14 degrees. Unreported simulation studies show that continuum regression estimators with other $\gamma > 0$ and different $\boldsymbol{\nu}(y)$ have a similar pattern under the setting of Proposition 1.

The assumption of Proposition 1 that $\text{span}(\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}) \neq \text{span}(\boldsymbol{\Gamma})$ is usually satisfied for a general positive definite matrix $\boldsymbol{\Delta}$. Given $\boldsymbol{\Delta} = \sigma^2\mathbf{I}_p$, we can see that $\text{span}(\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}) = \text{span}(\boldsymbol{\Gamma})$. Under this condition, the central subspace $\mathcal{S}_{y|\mathbf{x}}$ equals $\text{span}(\boldsymbol{\Gamma})$. The following proposition shows that solutions of continuum regression for all γ under this condition are consistent estimators of $\mathcal{S}_{y|\mathbf{x}}$.

Since we are interested in the central subspace, we define $D(.,.)$ as the largest principal angle between two subspaces to measure the distance between them. Let $\hat{\mathcal{S}}_n = \text{span}(\hat{\mathbf{c}}_1 \ldots, \hat{\mathbf{c}}_\omega)$ and $\mathcal{S}_{\boldsymbol{\Gamma}} = \text{span}(\boldsymbol{\Gamma})$.

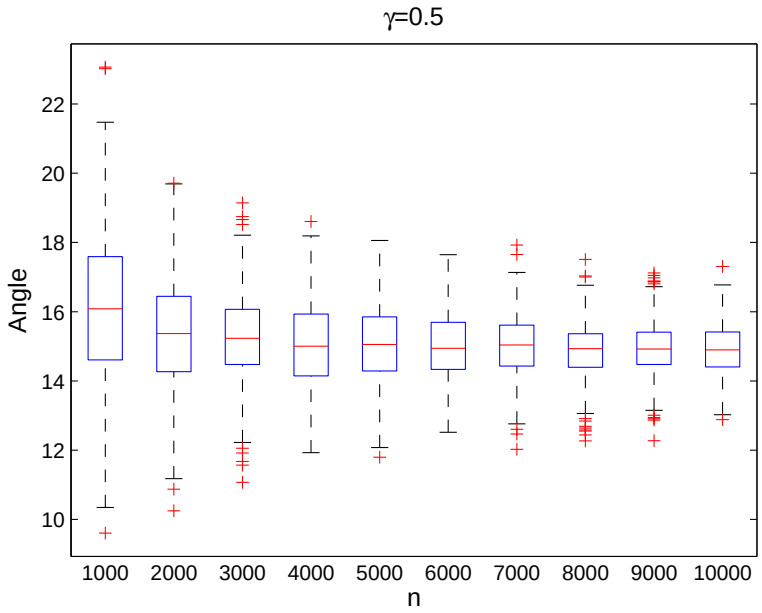

Figure 2.1: Boxplots of angles when $\boldsymbol{\nu}(y) = y$

Proposition 2 *Suppose that* $\boldsymbol{\Delta} = \sigma^2 \mathbf{I}_p$.

(i) We have $\mathbf{c}_1 \ldots, \mathbf{c}_\omega \in \mathcal{S}_{y|\mathbf{x}}$, *for all* $\gamma \geq 0$ *and* $\omega \leq d$, *where* $\mathbf{c}_i$ *denotes the population version of* $\hat{\mathbf{c}}_i$ *for* $i = 1, \ldots, \omega$.

(ii) If $\omega = d$, *then* $D(\hat{\mathcal{S}}_n, \mathcal{S}_{\boldsymbol{\Gamma}})$ *converges to 0 in probability as* $n \to \infty$, *for all* $\gamma \geq 0$.

Two more simulation studies were conducted as follows. We generated 500 data sets from model (2.2) with $p = 10$, $\boldsymbol{\mu} = 0$, $\boldsymbol{\Gamma} = (1, \ldots, 1)^T/\sqrt{10}$, $\boldsymbol{\Delta} = \mathbf{I}_p$ and y and the error ϵ following a standard normal distribution. Two different $\nu(y) = y$ and $\nu(y) = y + 1.5y^2$ were used. Figure 2.2 shows the average angles between the true direction $\boldsymbol{\Gamma}$ and the first direction of continuum regressions with $\gamma = 0, 0.25, 1, 3$ and ∞ versus n. Both figures indicate the consistency of estimating the central subspace in continuum regression for arbitrary γ when $\boldsymbol{\Delta} = \sigma^2 \mathbf{I}_p$. In Figure 2.2, lines marked with $\times$, $\triangle$, $\circ$, $\bullet$ and $\square$ represent estimators with $\gamma = 0, 0.25, 1, 3$ and ∞ respectively. The left panel of Figure 2.2 shows that partial least squares ($\gamma = 1$) performs best. This is not surprising because the first direction of continuum regression with $\gamma = 1$ is the same as the maximum likelihood

estimator when $d = 1$, $\mathbf{\Delta} = \sigma^2\mathbf{I}_p$ and $\nu(y) = y$ is considered as known in model (2.2). The right panel of Figure 2.2 shows that principal component regression ($\gamma = \infty$) is the best performer. In fact, Cook (2007) showed that principal component regression is the maximum likelihood estimator of the central subspace when $\nu(y)$ is unknown and $\mathbf{\Delta} = \sigma^2\mathbf{I}_p$ in model (2.2). Unreported simulation studies show that either partial least squares or principal component regression dominates in efficiency, or one of them is very close the best performer when $\mathbf{\Delta} = \sigma^2\mathbf{I}_p$.

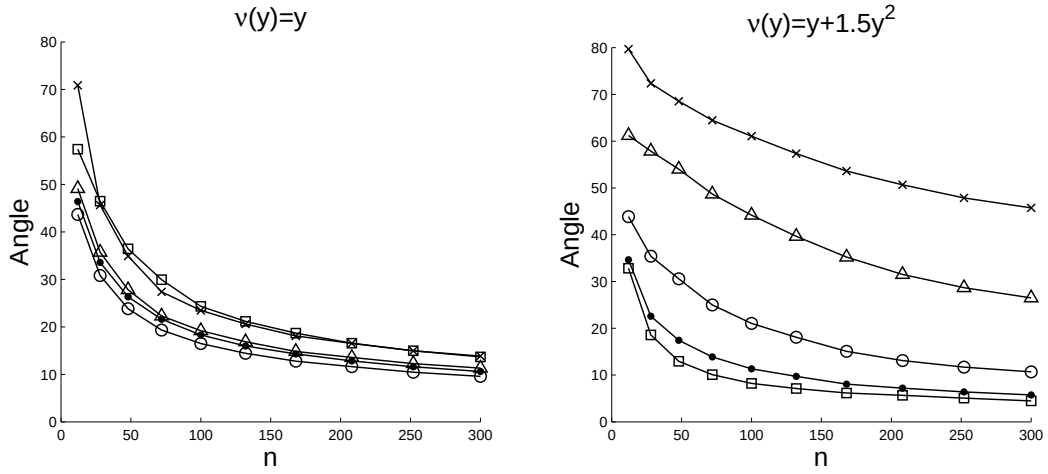

Figure 2.2: Average angles versus n when $\nu(y) = y$ and $\nu(y) = y + 1.5y^2$
Lines marked with $\times$, $\triangle$, $\circ$, $\bullet$ and $\square$ represent estimators with $\gamma = 0, 0.25, 1, 3$ and ∞ respectively.

2.2 Discussion

From Section 2.1 we have seen that if $\mathbf{x}$ given y has a normal distribution with the constant general conditional variance, the continuum regression solution with $\gamma = 0$ is the only consistent estimate of one direction in the central subspace. If the conditional variance is isotropic, continuum regression solutions for all γ are consistent estimators of directions in the central subspace. However, our simulation studies show that it seems enough to consider either partial least squares ($\gamma = 1$) or principal component regression ($\gamma = \infty$) in this scenario. The "continuum" seems unnecessary when the conditional distribution $\mathbf{x}$ given y is normal. When the conditional distribution deviates from

normality, the consistency theory of continuum regression is not clear and it is surely worth further research.

2.3 Appendix

Proof of Proposition 1

(i) Under model (2.2) and given $d = 1$, define $\text{cov}\{\nu(y), y\} = k$ and $\text{var}\{\nu(y)\} = l$ where k and l are both scalars. Then $\mathbf{\Sigma} = \text{var}(\mathbf{x}) = \text{var}\{E(\mathbf{x}|y)\} + E\{\text{var}(\mathbf{x}|y)\} = l\mathbf{\Gamma}\mathbf{\Gamma}^T + \mathbf{\Delta}$ and $\mathbf{s} = \text{cov}(\mathbf{x}, y) = k\mathbf{\Gamma}$, where $\mathbf{\Sigma}$ denotes the population marginal covariance matrix of $\mathbf{x}$ and $\mathbf{s}$ the population covariance vector of $\mathbf{x}$ and y.

Using the method of Lagrange multipliers, the first population continuum regression solution $\mathbf{c}_1$ can be found by searching a stationary point of the following formula defined on the vector $\mathbf{c}$:

$$\log\{T(\mathbf{c})\} + \lambda(\mathbf{c}^T\mathbf{c} - 1) = \log(\mathbf{c}^T\mathbf{s})^2 + (\gamma - 1)\log(\mathbf{c}^T\mathbf{\Sigma}\mathbf{c}) - \lambda(\mathbf{c}^T\mathbf{c} - 1), \tag{2.3}$$

where λ is the Lagrange multiplier. Taking derivative of 2.3 with respect to λ, we have $\mathbf{c}^T\mathbf{c} - 1 = 0$. Taking derivative of 2.3 with respect to $\mathbf{c}$, we have

$$\frac{\mathbf{s}}{\mathbf{c}^T\mathbf{s}} + (\gamma - 1)\frac{\mathbf{\Sigma}\mathbf{c}}{\mathbf{c}^T\mathbf{\Sigma}\mathbf{c}} - \lambda\mathbf{c} = 0.$$

Multiplying by $\mathbf{c}^T$ from the left of the formula above, we have $\lambda = \gamma$. Then replacing λ with γ, we found that $\mathbf{c}_1$ is a root of the following equation:

$$(\mathbf{c}^T\mathbf{s})\left(\frac{1-\gamma}{\mathbf{c}^T\mathbf{\Sigma}\mathbf{c}}\mathbf{\Sigma} + \gamma\mathbf{I}_p\right)\mathbf{c} - \mathbf{s} = 0, \tag{2.4}$$

subject to $\mathbf{c}^T\mathbf{c} = 1$. Assume that $\mathbf{c}_1 = \mathbf{\Delta}^{-1}\mathbf{\Gamma}/||\mathbf{\Delta}^{-1}\mathbf{\Gamma}||_2$ first, then $\mathbf{\Delta}^{-1}\mathbf{\Gamma}/||\mathbf{\Delta}^{-1}\mathbf{\Gamma}||_2$ has to be the root of 2.4. Replacing $\mathbf{c}$, $\mathbf{\Sigma}$ and $\mathbf{s}$ with $\mathbf{\Delta}^{-1}\mathbf{\Gamma}/||\mathbf{\Delta}^{-1}\mathbf{\Gamma}||_2$, $l\mathbf{\Gamma}\mathbf{\Gamma}^T + \mathbf{\Delta}$ and $k\mathbf{\Gamma}$ respectively in 2.4 and after simplification, we must have:

$$-k\gamma\mathbf{\Gamma} + k\gamma\frac{\mathbf{\Gamma}^T\mathbf{\Delta}^{-1}\mathbf{\Gamma}}{\mathbf{\Gamma}^T\mathbf{\Delta}^{-2}\mathbf{\Gamma}}(\mathbf{\Delta}^{-1}\mathbf{\Gamma}) = 0. \tag{2.5}$$

However, given $k \neq 0$, $\gamma > 0$ and $\text{span}(\mathbf{\Delta}^{-1}\mathbf{\Gamma}) \neq \text{span}(\mathbf{\Gamma})$, 2.5 can not equal 0. Thus the assumption $\mathbf{c}_1 = \mathbf{\Delta}^{-1}\mathbf{\Gamma}/||\mathbf{\Delta}^{-1}\mathbf{\Gamma}||_2$ leads to a contradiction, and consequently

$\mathbf{c}_1 \neq (\mathbf{\Delta}^{-1}\mathbf{\Gamma})/||\mathbf{\Delta}^{-1}\mathbf{\Gamma}||_2$. When $\gamma = 0$, 2.5 is always true.

(ii) Using the same argument as above, the first sample continuum regression solution $\hat{\mathbf{c}}_1$ can be found by searching a root of

$$(\mathbf{c}^T\hat{\mathbf{s}})\left(\frac{1-\gamma}{\mathbf{c}^T\hat{\mathbf{\Sigma}}\mathbf{c}}\hat{\mathbf{\Sigma}} + \gamma\mathbf{I}_p\right)\mathbf{c} - \hat{\mathbf{s}} = 0, \tag{2.6}$$

subject to $\mathbf{c}^T\mathbf{c} = 1$. Let $\rho = \mathbf{c}^T\hat{\mathbf{\Sigma}}\mathbf{c}$ and $\mathbf{Q}(\rho) = \{(1-\gamma)\rho^{-1}\hat{\mathbf{\Sigma}} + \gamma\mathbf{I}_p\}$. Suppose $\mathbf{Q}(\rho)$ is non-singular first and denote $\mathbf{H}(\rho) = \mathbf{Q}^{-1}(\rho)$, the inverse of $\mathbf{Q}(\rho)$. Following the discussion of Chan and Mak in the paper (Stone and Brooks, 1990), we have

$$\mathbf{c} = \frac{\pm\mathbf{H}(\rho)\hat{\mathbf{s}}}{\{\hat{\mathbf{s}}^T\mathbf{H}(\rho)\hat{\mathbf{s}}\}^{1/2}}, \tag{2.7}$$

and

$$\rho - \frac{\hat{\mathbf{s}}^T\mathbf{H}(\rho)\hat{\mathbf{\Sigma}}\mathbf{H}(\rho)\hat{\mathbf{s}}}{\hat{\mathbf{s}}^T\mathbf{H}(\rho)\hat{\mathbf{s}}} = 0. \tag{2.8}$$

Thus $\hat{\mathbf{c}}_1$ can be found by searching a root of 2.8 in ρ and substitute it into the expression of 2.7. Let $\hat{\boldsymbol{\theta}}^T\mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_i, \ldots, \hat{\lambda}_p)\hat{\boldsymbol{\theta}}$ be the spectral decomposition of $\hat{\mathbf{\Sigma}}$ where $\hat{\boldsymbol{\theta}}$ denotes the matrix of eigenvectors and $\hat{\lambda}_i$ the ith eigenvalue with descending orders. We have

$$\mathbf{H}(\rho) = \hat{\boldsymbol{\theta}}^T\mathrm{diag}\left(\frac{\rho}{(1-\gamma)\hat{\lambda}_1 + \rho\gamma}, \ldots, \frac{\rho}{(1-\gamma)\hat{\lambda}_i + \rho\gamma}, \ldots, \frac{\rho}{(1-\gamma)\hat{\lambda}_p + \rho\gamma}\right)\hat{\boldsymbol{\theta}}.$$

Denote $\hat{\boldsymbol{\theta}}\hat{\mathbf{s}} = (\hat{k}_1, \ldots, \hat{k}_p)^T$. Replacing $\mathbf{H}(\rho)$ in 2.8 with the formula above and after simplification, we have

$$\sum_{i=1}^{p}\frac{\hat{\lambda}_i\hat{k}_i^2}{\{(1-\gamma)\hat{\lambda}_i + \rho\gamma\}^2} - \sum_{i=1}^{p}\frac{\hat{k}_i^2}{\{(1-\gamma)\hat{\lambda}_i + \rho\gamma\}} = 0, \tag{2.9}$$

as an equation defined on ρ. Given $\mathbf{Q}(\rho)$ is non-singular, we know $(1-\gamma)\hat{\lambda}_i + \rho\gamma \neq 0$ for $i = 1, \ldots, p$. Thus 2.9 can be transformed to a polynomial equation with degree $2p-2$. It is well known that the root of a polynomial equation is a continuous function of its coefficients (Marden, 1949). The coefficients of the polynomial equation are algebraic functions of $\hat{\lambda}_i$ and $\hat{k}_i$ which are continuous functions of $(\hat{\mathbf{\Sigma}}, \hat{\mathbf{s}})$. Denote a root of 2.9 as $\hat{\rho}$. Substituting $\hat{\rho}$ into 2.7, we can conclude $\hat{\mathbf{c}}_1$ is a continuous function of $(\hat{\mathbf{\Sigma}}, \hat{\mathbf{s}})$.

If $\mathbf{Q}(\rho)$ is singular, then for some m, $(1-\gamma)\hat{\lambda}_m + \rho\gamma = 0$. We have $\rho = \mathbf{c}^T\hat{\mathbf{\Sigma}}\mathbf{c} = (\gamma-1)\hat{\lambda}_m/\gamma$, which induces that $\gamma > 1$ under this scenario. We rewrite 2.6 as

$$\begin{aligned}\frac{\hat{\mathbf{s}}}{\mathbf{c}^T\hat{\mathbf{s}}} &= \left(\frac{1-\gamma}{\mathbf{c}^T\hat{\mathbf{\Sigma}}\mathbf{c}}\hat{\mathbf{\Sigma}} + \gamma\mathbf{I}_p\right)\mathbf{c} = (-\gamma\hat{\mathbf{\Sigma}}/\hat{\lambda}_m + \gamma\mathbf{I}_p)\mathbf{c} \\ &= \gamma\hat{\boldsymbol{\theta}}^T \text{diag}\left(1-\frac{\hat{\lambda}_1}{\hat{\lambda}_m}, \ldots, 0, \ldots, 1-\frac{\hat{\lambda}_p}{\hat{\lambda}_m}\right)\hat{\boldsymbol{\theta}}\mathbf{c},\end{aligned}$$

Multiplying by $\hat{\boldsymbol{\theta}}$ from the left of the formula above, we have

$$\frac{\hat{\boldsymbol{\theta}}\hat{\mathbf{s}}}{\mathbf{c}^T\hat{\mathbf{s}}} = \gamma\text{diag}\left(1-\frac{\hat{\lambda}_1}{\hat{\lambda}_m}, \ldots, 0, \ldots, 1-\frac{\hat{\lambda}_p}{\hat{\lambda}_m}\right)\hat{\boldsymbol{\theta}}\mathbf{c}.$$

It tells us that the mth row of $\hat{\boldsymbol{\theta}}\hat{\mathbf{s}}$ is exactly 0. Given $\mathbf{x}$ and y has a joint distribution and a continuous density function, $\hat{\boldsymbol{\theta}}\hat{\mathbf{s}}$ has a continuous density too. Then we can conclude that $\mathbf{Q}(\rho)$ is singular with measure 0.

It is well known that $\hat{\mathbf{\Sigma}}$ converges to $\mathbf{\Sigma}$ in probability and $\hat{\mathbf{s}}$ converges to $\mathbf{s}$ in probability. By Slutsky theorem, we can conclude that $\hat{\mathbf{c}}_1$ converges to $\mathbf{c}_1$ in probability. In other words, span($\hat{\mathbf{c}}_1$) is an inconsistent estimator of $\mathcal{S}_{y|\mathbf{x}}$ for $\gamma > 0$.

Proof of Proposition 2

(i) Under model (2.2), define $\text{cov}\{\boldsymbol{\nu}(y), y\}\cdot\text{cov}\{\boldsymbol{\nu}(y), y\}^T = \mathbf{K}$ and $\text{var}\{\boldsymbol{\nu}(y)\}) = \mathbf{L}$. Then $\mathbf{\Sigma} = \text{var}(\mathbf{x}) = \text{var}(E(\mathbf{x}|y)) + E(\text{var}(\mathbf{x}|y)) = \mathbf{\Gamma}\mathbf{L}\mathbf{\Gamma}^T + \sigma^2 I$ and $\mathbf{s}\mathbf{s}^T = \text{cov}(\mathbf{x}, y)\text{cov}(\mathbf{x}, y)^T = \mathbf{\Gamma}\mathbf{K}\mathbf{\Gamma}^T$. Since we assume that $\text{cov}(\mathbf{x}, y) \neq 0$, then $\mathbf{K} \neq 0$. We have

$$\begin{aligned}\log\{T(\mathbf{c})\} &= \log\{\mathbf{c}^T(\mathbf{\Gamma}\mathbf{K}\mathbf{\Gamma}^T)\mathbf{c}\} + (\gamma-1)\log\{\mathbf{c}^T(\mathbf{\Gamma}\mathbf{L}\mathbf{\Gamma}^T + \sigma^2\mathbf{I}_p)\mathbf{c}\} \\ &= \log\{\mathbf{c}^T(\mathbf{\Gamma}\mathbf{K}\mathbf{\Gamma}^T)\mathbf{c}\} + (\gamma-1)\log\{\mathbf{c}^T(\mathbf{\Gamma}\mathbf{L}\mathbf{\Gamma}^T)\mathbf{c} + \sigma^2\} \\ &= \log\frac{\mathbf{c}^T(\mathbf{\Gamma}\mathbf{K}\mathbf{\Gamma}^T)\mathbf{c}}{\mathbf{c}^T(\mathbf{\Gamma}\mathbf{L}\mathbf{\Gamma}^T)\mathbf{c} + \sigma^2} + \gamma\log\{\mathbf{c}^T(\mathbf{\Gamma}\mathbf{L}\mathbf{\Gamma}^T)\mathbf{c} + \sigma^2\}.\end{aligned}$$

If $\mathbf{c} \in \mathcal{S}_{\mathbf{\Gamma}_0}$, where $\mathbf{\Gamma}_0$ denotes the orthogonal complement of $\mathbf{\Gamma}$, then $\log\{T(\mathbf{c})\} = -\infty$ as $\mathbf{c}^T(\mathbf{\Gamma}\mathbf{K}\mathbf{\Gamma}^T)\mathbf{c} = 0$. This case will not happen since we are maximizing $\log\{T(\mathbf{c})\}$. Decompose $\mathbf{c}_i = \mathbf{e}_i + \mathbf{e}_{i0}$ where $\mathbf{e}_i \in \mathcal{S}_{\mathbf{\Gamma}}$ and $\mathbf{e}_{i0} \in \mathcal{S}_{\mathbf{\Gamma}_0}$. Since $\|\mathbf{c}_i\| = 1$, it is clear that $\|\mathbf{c}_i\|^2 = \|\mathbf{e}_i\|^2 + \|\mathbf{e}_{i0}\|^2 = 1$.

Suppose $\mathbf{c}_1 \notin \mathcal{S}_{\mathbf{\Gamma}}$, then $\|\mathbf{e}_1\| < 1$. We will prove $\tilde{\mathbf{c}}_1 = \eta_1 \mathbf{e}_1$ outperforms $\mathbf{c}_1$ where $\eta_1 = 1/\|\mathbf{e}_1\| > 1$. It is straightforward to have

$$\begin{aligned}
\log\{T(\mathbf{c}_1)\} &= \log \frac{\mathbf{c}_1^T(\mathbf{\Gamma K \Gamma}^T)\mathbf{c}_1}{\mathbf{c}_1^T(\mathbf{\Gamma L \Gamma}^T)\mathbf{c}_1 + \sigma^2} + \gamma \log\{\mathbf{c}_1^T(\mathbf{\Gamma L \Gamma}^T)\mathbf{c}_1 + \sigma^2\} \\
&= \log \frac{\mathbf{e}_1^T(\mathbf{\Gamma K \Gamma}^T)\mathbf{e}_1}{\mathbf{e}_1^T(\mathbf{\Gamma L \Gamma}^T)\mathbf{e}_1 + \sigma^2} + \gamma \log\{\mathbf{e}_1^T(\mathbf{\Gamma L \Gamma}^T)\mathbf{e}_1 + \sigma^2\} \\
&< \log \frac{\eta_1^2\mathbf{e}_1^T(\mathbf{\Gamma K \Gamma}^T)\mathbf{e}_1}{\eta_1^2\mathbf{e}_1^T(\mathbf{\Gamma L \Gamma}^T)\mathbf{e}_1 + \sigma^2} + \gamma \log\{\eta_1^2\mathbf{e}_1^T(\mathbf{\Gamma L \Gamma}^T)\mathbf{e}_1 + \sigma^2\} \\
&= \log\{T(\tilde{\mathbf{c}}_1)\}.
\end{aligned}$$

The inequality in the above formula holds because the term on the right hand side is an increasing function with respect to η_1^2 and $\eta_1^2 > 1$. By the definition of $\tilde{\mathbf{c}}_1$ ($\|\tilde{\mathbf{c}}_1\| = 1$), the last equality holds. Thus $\mathbf{c}_1$ can not be the maximizer unless $\|e_{10}\| = 0$, and then $\mathbf{c}_1 \in \mathcal{S}_{\mathbf{\Gamma}}$. Assume $\mathbf{c}_j \in \mathcal{S}_{\mathbf{\Gamma}}$, for all $j = 1, \ldots, i-1$, now we want to show that $\mathbf{c}_i \in \mathcal{S}_{\mathbf{\Gamma}}$. We have

$$\mathbf{\Sigma} = \mathbf{\Gamma L \Gamma}^T + \sigma^2 \mathbf{I}_p = \mathbf{\Gamma}(\sigma^2 \mathbf{I}_d + \mathbf{L})\mathbf{\Gamma}^T + \sigma^2 \mathbf{\Gamma}_0 \mathbf{\Gamma}_0^T.$$

Moreover, the constraint ${\mathbf{c}_i}^T \mathbf{\Sigma} \mathbf{c}_j = 0$ is equivalent to ${\mathbf{c}_i}^T \mathbf{\Gamma}(\sigma^2 \mathbf{I}_d + \mathbf{L})\mathbf{\Gamma}^T \mathbf{c}_j = 0$ because $\mathbf{c}_j \in \mathcal{S}_{\mathbf{\Gamma}}$. Suppose $\mathbf{c}_i \notin \mathcal{S}_{\mathbf{\Gamma}}$, following the same argument, we can show that $\tilde{\mathbf{c}}_i = \eta_i \mathbf{e}_i$ outperforms $\mathbf{c}_i$ where $\eta_i = 1/\|\mathbf{e}_i\| > 1$. At the same time, $\tilde{\mathbf{c}}_i$ satisfies all the constraints. Then we can conclude $\mathbf{c}_i \in \mathcal{S}_{\mathbf{\Gamma}}$ for all $i = 1, \ldots, d$.

(ii) Denote $\boldsymbol{\psi}_0 = 0$, $\boldsymbol{\psi}_k = (\hat{\mathbf{c}}_1 \ldots, \hat{\mathbf{c}}_k)$ for $k \geq 1$. Using the method of Lagrange multipliers, $\hat{\mathbf{c}}_{k+1}$ can be reached by finding a root of the equation defined on the vector $\mathbf{c}$:

$$(\mathbf{c}^T \hat{\mathbf{s}}) \left(\frac{1-\gamma}{\mathbf{c}^T \hat{\mathbf{\Sigma}} \mathbf{c}} \hat{\mathbf{\Sigma}} + \gamma \mathbf{A} \right) \mathbf{c} - \mathbf{A}\hat{\mathbf{s}} = 0,$$

subject to $\mathbf{c}^T \mathbf{c} = 1$, where $\mathbf{A} = \mathbf{I}_p - \hat{\mathbf{\Sigma}} \boldsymbol{\psi}_k (\boldsymbol{\psi}_k^T \hat{\mathbf{\Sigma}} \boldsymbol{\psi}_k)^{-1} \boldsymbol{\psi}_k^T$.

Following the same argument as the proof of Proposition 1, it can be shown that $(\hat{\mathbf{c}}_1 \ldots, \hat{\mathbf{c}}_\omega)$ is a continuous function of $(\hat{\mathbf{\Sigma}}, \hat{\mathbf{s}})$. Since $\hat{\mathbf{\Sigma}}$ converges to $\mathbf{\Sigma}$ in probability and $\hat{s}$ converges to $\mathbf{s}$ in probability, by Slutsky theorem, we can conclude that $(\hat{\mathbf{c}}_1 \ldots, \hat{\mathbf{c}}_\omega)$ converges to $(\mathbf{c}_1, \ldots, \mathbf{c}_\omega)$ in probability.

Provided $\mathbf{\Sigma}$ is nonsingular, $\mathbf{c}_1, \ldots, \mathbf{c}_\omega$ are linearly independent for $\omega \leq d$. If $\omega = d$, by Proposition 2 (i), $(\mathbf{c}_1, \ldots, \mathbf{c}_\omega)$ forms a base of the central subspace. Since $(\hat{\mathbf{c}}_1 \ldots, \hat{\mathbf{c}}_\omega)$

converges to $(\mathbf{c}_1, \ldots, \mathbf{c}_\omega)$ in probability, we can conclude that $D(\hat{\mathcal{S}}_n, \mathcal{S}_{\mathbf{\Gamma}})$ converges to 0 in probability.

Chapter 3

Variable selection in sufficient dimension reduction

Most dimension reduction methods suffer because the estimated linear reductions usually involve all of the original predictors $\mathbf{x}$. As a consequence, the results can be hard to interpret, the important variables may be difficult to identify and the efficiency gain may be less than that possible with variable selection. These limitations can be overcome by screening irrelevant and redundant predictors while still estimating a few linear combinations of the active predictors. Some attempts have been made to address this problem in dimension reduction generally and SDR in particular. For example, Li et al. (2005) proposed a model-free variable selection method based on SDR. Zou et al. (2006) proposed a sparse principal component analysis. Ni et al. (2005) introduced a shrinkage version of SIR, while Li and Nachtsheim (2006) suggested a sparse version of SIR. Li (2007) studied sparse SDR by adapting the approach of Zou et al. (2006). Zhou and He (2008) proposed a constrained canonical correlation procedure (C^3) based on imposing the L_1-norm constraint on the effective dimension reduction estimates in CANCOR (Fung et al. 2002), followed by a simple variable filtering method. Their procedure is attractive because they showed that it has the oracle property (Donoho and Johnstone 1994; Fan and Li 2001). More recently, Leng and Wang (2009) proposed a general adaptive sparse principal component analysis and Johnstone and Lu (2009) studied the large p theory in sparse principal components analysis.

However, most existing sparse dimension reduction methods are conducted stepwise, estimating a sparse solution for a basis matrix of the central subspace column by column. Instead, in this thesis, we propose a unified one-step approach to reduce the number of variables appearing in the estimate of $\mathcal{S}_{y|\mathbf{x}}$. Our approach, which hinges operationally on Grassmann manifold optimization, is able to achieve dimension reduction and variable selection simultaneously. Additionally, our proposed method has the oracle property: under mild conditions the proposed estimator would perform asymptotically as well as if the true irrelevant predictors were known.

We start in Section 3.1.1 by reviewing the link between many SDR methods and a generalized eigenvalue problem disclosed by L. Li (2007). In Section 3.1.2 we describe a new SDR penalty function that is invariant under orthogonal transformations and targets the removal of row vectors from the basis matrix. Based on this penalty function, in Section 3.1.3, a coordinate-independent penalized procedure is proposed which enables us to incorporate many model-free and model-based SDR approaches into a simple and unified framework to implement variable selection within SDR. A fast algorithm, which combines a local quadratic approximation (Fan and Li 2001) and an eigensystem analysis in each iteration step, is suggested in Section 3.1.4 to handle our Grassmann manifold optimization problem with its non-differentiable penalty function. In Section 3.1.5 we describe the oracle property of our estimator. Its proof differs significantly from those in the context of variable selection in single-index models (e.g., Fan and Li 2001; Zou 2006) because the focus here is on subspaces rather than on coordinates. Results of simulation studies are reported in Section 3.2, and the Boston housing data, is analyzed in Section 3.3. We introduce a Matlab package for our method in Section 3.4 and a download link is provided. Concluding remarks about the proposed method can be found in Section 3.5. Technical details are given in the Appendix.

3.1 Theory and Methodology

3.1.1 Motivation: generalized eigenvalue problems revisited

L. Li (2007) showed that many moment based sufficient dimension reduction methods can be formulated as a generalized eigenvalue problem in the following form

$$\mathbf{M}_n\boldsymbol{\delta}_{ni} = \lambda_{ni}\mathbf{N}_n\boldsymbol{\delta}_{ni}, \text{ for } i = 1, \ldots, p, \tag{3.1}$$

where $\mathbf{M}_n \geq 0$ is a method-specific symmetric kernel matrix, $\mathbf{N}_n > 0$ is symmetric, often taking the form of the sample covariance matrix $\boldsymbol{\Sigma}_n$ of $\mathbf{x}$; $\boldsymbol{\delta}_{n1}, \ldots, \boldsymbol{\delta}_{np}$ are eigenvectors such that $\boldsymbol{\delta}_{ni}^T\mathbf{N}_n\boldsymbol{\delta}_{nj} = 1$ if $i = j$ and 0 if $i \neq j$ and $\lambda_{n1} \geq \ldots \geq \lambda_{np}$ are the corresponding eigenvalues. We use the subscript "n" to indicate that $\boldsymbol{\Sigma}_n$, $\mathbf{M}_n$, $\mathbf{N}_n$ and λ_{ni} are the sample versions of the corresponding population analogs $\boldsymbol{\Sigma}$, $\mathbf{M}$, $\mathbf{N}$ and λ_i. Under certain conditions that are usually imposed only on the marginal distribution of $\mathbf{x}$, the first d eigenvectors $\{\boldsymbol{\delta}_{n1}, \ldots, \boldsymbol{\delta}_{nd}\}$, which correspond to the nonzero eigenvalues $\lambda_{n1} > \ldots > \lambda_{nd}$ form a consistent estimator of a basis for the central subspace. Letting $\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}\{\mathbf{x} - E(\mathbf{x})\}$. Many commonly used moment based SDR methods are listed in Table 3.1 with the population versions of $\mathbf{M}_n$ and $\mathbf{N}_n$.

Table 3.1: The generalized eigenvalue formulations

Method	$\mathbf{M}$	$\mathbf{N}$
PCA	$\boldsymbol{\Sigma}$	$\mathbf{I}_p$
PFC	$\boldsymbol{\Sigma}_{\text{fit}}$	$\boldsymbol{\Sigma}$
SIR	$\text{Cov}[E\{\mathbf{x} - E(\mathbf{x})\|y\}]$	$\boldsymbol{\Sigma}$
SAVE	$\boldsymbol{\Sigma}^{1/2}E[\{\mathbf{I}_p - \text{Cov}(\mathbf{z}\|y)\}^2]\boldsymbol{\Sigma}^{1/2}$	$\boldsymbol{\Sigma}$
DR	$\boldsymbol{\Sigma}^{1/2}\{2E[E^2(\mathbf{z}\mathbf{z}^T\|y)] + 2E^2\big[E(\mathbf{z}\|y)E(\mathbf{z}^T\|y)\big]$ $+2E\big[E(\mathbf{z}\|y)E(\mathbf{z}\|y)\big]E\big[E(\mathbf{z}\|y)E(\mathbf{z}^T\|y)\big] - 2\mathbf{I}_p\}\boldsymbol{\Sigma}^{1/2}$	$\boldsymbol{\Sigma}$

Following Cook (2004), L. Li (2007) showed that the eigenvectors $\{\boldsymbol{\delta}_{n1}, \ldots, \boldsymbol{\delta}_{nd}\}$ from (3.1) can be obtained by minimizing a least square objective function. Let

$$\widehat{\boldsymbol{V}} = \arg\min_{\boldsymbol{V}} \sum_{i=1}^{p} \|\mathbf{N}_n^{-1}\mathbf{m}_i - \boldsymbol{V}\boldsymbol{V}^T\mathbf{m}_i\|_{\mathbf{N}_n}^2, \text{ subject to } \boldsymbol{V}^T\mathbf{N}_n\boldsymbol{V} = \mathbf{I}_d, \tag{3.2}$$

where $\mathbf{m}_i$ denotes the i-th column of $\mathbf{M}_n^{1/2}$, $i = 1, \ldots, p$, $\boldsymbol{V}$ is a $p \times d$ matrix and the norm here is with respect to the $\mathbf{N}_n$ inner product. Then $\widehat{\boldsymbol{V}}_j = \boldsymbol{\delta}_{nj}$, $j = 1, \ldots, d$, where

$\widehat{\boldsymbol{V}}_j$ stands for the j-th column of $\widehat{\boldsymbol{V}}$, so that span($\widehat{\boldsymbol{V}}$) is the estimator of the central subspace. To get a sparse solution, L. Li then added penalties to the objective function in (3.1.3), leading to the optimization problem

$$(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{V}}_s) = \min_{\boldsymbol{\alpha}, \boldsymbol{V}} \left\{ \sum_{i=1}^{p} \|\mathbf{N}_n^{-1}\mathbf{m}_i - \boldsymbol{\alpha}\boldsymbol{V}^T\mathbf{m}_i\|^2_{\mathbf{N}_n} + \tau_2 \mathrm{tr}(\boldsymbol{V}^T\mathbf{N}_n\boldsymbol{V}) + \sum_{i=1}^{d} \tau_{1,j}\|\boldsymbol{V}_j\|_1 \right\},$$

subject to $\boldsymbol{\alpha}^T\mathbf{N}_n\boldsymbol{\alpha} = \mathbf{I}_d$, where tr$(\cdot)$ stands for the trace operator, $\|.\|_r$ denotes the L_r norm, τ_2 is some positive constant and $\tau_{1,j} \geq 0$ for $j = 1, \ldots, d$ are the lasso shrinkage parameters that need to be determined by some method like cross validation (CV). The solution $\widehat{\boldsymbol{V}}_s$ is called the sparse sufficient dimension reduction estimator. As a result of the lasso constraint, $\widehat{\boldsymbol{V}}_s$ is expected to have some elements shrunk to zero.

We can see that L. Li's sparsity method is coordinate dependent because the L_1 penalty term is not invariant under the orthogonal transformation of the basis and it forces individual elements of the basis matrix $\widehat{\boldsymbol{V}}_s$ to zero. However variable screening requires that entire rows of $\widehat{\boldsymbol{V}}_s$ be zero, which is not the explicit goal of L. Li's method. To see this more clearly, partition $\mathbf{x}$ as $(\boldsymbol{x}_1^T, \boldsymbol{x}_2^T)^T$, where $\boldsymbol{x}_1$ corresponds to q elements of $\boldsymbol{x}$ and $\boldsymbol{x}_2$ to the remaining elements. If

$$y \perp\!\!\!\perp \boldsymbol{x}_2 | \boldsymbol{x}_1, \tag{3.3}$$

then $\boldsymbol{x}_2$ can be removed, as given $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ contains no further information about y. Let the $p \times d$ matrix $\boldsymbol{\eta}$ be a basis for $\mathcal{S}_{y|\mathbf{x}}$ and partition $\boldsymbol{\eta} = (\boldsymbol{\eta}_1^T, \boldsymbol{\eta}_2^T)^T$ in accordance with the partition of $\mathbf{x}$. Then the condition () is equivalent to $\boldsymbol{\eta}_2 = 0$ (Cook 2004), so the corresponding rows of the basis are zero vector.

In effect, L. Li's method is designed for element screening, not variable screening. Our experience reflects this limitation and reinforces the notion that $\widehat{\boldsymbol{V}}_s$ may not be sufficiently effective at variable screening. Inspired by L. Li's method, we propose a new variable screening method – called coordinate-independent sparse estimation (CISE) – in the next subsection. We will show that CISE is simpler and more effective than L. Li's method at variable screening.

CISE can be applied not only to moment based SDR approaches but also model based approaches. Cook (2007) and Cook and Forzani (2008) developed several powerful model-based dimension reduction approaches, collectively referred to as principal fitted

components (PFC). PFC-based SDR methods can also be formulated in the same way as (3.1), as summarized in the next proposition. In preparation, consider the following model for the conditional distribution of $\mathbf{x}$ given y,

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\xi}\mathbf{f}(y) + \boldsymbol{\Delta}^{1/2}\boldsymbol{\epsilon}, \tag{3.4}$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is a location vector, $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d}$, $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \mathbf{I}_d$, $\boldsymbol{\xi} \in \mathbb{R}^{d\times r}$ with rank d, $\mathbf{f} \in \mathbb{R}^r$ is a known vector-valued function of y, $\boldsymbol{\Delta} = \mathrm{Var}(\mathbf{x}|y) > 0$, and $\boldsymbol{\epsilon} \in \mathbb{R}^p$ is assumed to be independent of y and normally distributed with mean 0 and identity covariance matrix.

Proposition 3 *Suppose the conditional distribution of* $\mathbf{x}$ *given* y *can be described by (3.4). Then the maximum likelihood estimator (MLE) of* $\mathcal{S}_{y|\mathbf{x}}$ *can be obtained through the generalized eigenvalue problem of the form (3.1) with* $\mathbf{M}_n = \widehat{\boldsymbol{\Sigma}}_{\text{fit}}$ *and* $\mathbf{N}_n = \boldsymbol{\Sigma}_n$*, where* $\widehat{\boldsymbol{\Sigma}}_{\text{fit}}$ *is the sample covariance matrix of the fitted vectors from the linear regression of* $\mathbf{x}$ *on* $\mathbf{f}$*.*

A commonly used case in the PFC models is $\boldsymbol{\Delta} = \sigma^2\mathbf{I}_p$ for $\sigma > 0$, in which the MLE of $\mathcal{S}_{y|\mathbf{x}}$ can be obtained through (3.1) with $\mathbf{M}_n = \widehat{\boldsymbol{\Sigma}}_{\text{fit}}$ and $\mathbf{N}_n = \mathbf{I}_p$. The covariates $\mathbf{f}(y)$ in model (3.4) usually take form of polynomial, piecewise or Fourier basis functions. Thus the PFC models can effectively deal with the nonlinear relationship between the predictors and the response.

3.1.2 A coordinate-independent penalty function

Let $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_p)^T$ denote a $p \times d$ matrix with rows $\mathbf{v}_i^T$, $i = 1, \ldots, p$. In this section we introduce a coordinate-independent penalty, depending only on the subspace spanned by the columns of $\mathbf{V}$. Let $\mathbf{q}_i$ be the vector in $\mathbb{R}^p$ with the ith component one, else zero.

We define a general coordinate-independent penalty function as

$$\phi(\mathbf{V}) = \sum_i \theta_i h_i(\mathbf{q}_i^T\mathbf{V}\mathbf{V}^T\mathbf{q}_i)$$

where $\theta_i \geq 0$ serve as penalty parameters, and h_i are positive convex functions defined in $\mathbb{R}^d$. To achieve variable screening, the functions h_i must be non-differentiable at the zero vector. It is clear that the function ϕ is independent of the basis used to represent

the span of $\mathbf{V}$, since for any orthogonal matrix $\mathbf{O}$, $\phi(\mathbf{V}) = \phi(\mathbf{VO})$. In fact, any penalty function defined on $\mathbf{VV}^T$ meets our requirement.

Given $h_1 = \ldots = h_p = \sqrt{(\cdot)}$, we have a special coordinate-independent penalty function:

$$\rho(\mathbf{V}) = \sum_{i=1}^{p} \theta_i \|\mathbf{v}_i\|_2. \tag{3.5}$$

A method for selecting the tuning parameters will be discussed in Section 3.1.6. We can see that the penalty function ρ has the same form as the group lasso proposed by Yuan and Lin (2006) but their concepts and usages are essentially different. Through this thesis, we shall use only ρ in application and theory to demonstrate our ideas.

Penalty (3.5) is appealing for variable selection because it is independent of the basis used to represent the span of $\mathbf{V}$, $\rho(\mathbf{V}) = \rho(\mathbf{VO})$ for any orthogonal matrix $\mathbf{O}$, and because it groups the row vector coefficients of $\mathbf{V}$. This motivated us to consider the regularized function (3.5) that can shrink the corresponding row vectors of irrelevant variables to zero. Another appealing feature of using this penalty is its oracle property, which is discussed in Section 3.1.5.

3.1.3 Coordinate-independent sparse estimation

Recall the generalized eigenvalue problem (3.1) and the associated notation. Formally,

$$\sum_{i=1}^{p} ||\mathbf{N}_n^{-1}\mathbf{m}_i - \mathbf{VV}^T\mathbf{m}_i||^2_{\mathbf{N}_n} = \mathrm{tr}(\mathbf{G}_n) - \mathrm{tr}(\mathbf{V}^T\mathbf{M}_n\mathbf{V}),$$

where $\mathbf{G}_n = \mathbf{N}_n^{-1/2}\mathbf{M}_n\mathbf{N}_n^{-1/2}$ and we use $\mathbf{G}$ to denote its population analog in what follows. Hence, the ordinary sufficient dimension reduction estimation (OSDRE) given in () is

$$\widehat{\mathbf{V}} = \arg\min_{\mathbf{V}} -\mathrm{tr}(\mathbf{V}^T\mathbf{M}_n\mathbf{V}), \text{ subject to } \mathbf{V}^T\mathbf{N}_n\mathbf{V} = \mathbf{I}_d. \tag{3.6}$$

By using the coordinate independent penalty function given in last subsection, we propose the following coordinate-independent sparse sufficient dimension reduction estimator (CISE),

$$\tilde{\mathbf{V}} = \arg\min_{\mathbf{V}}\{-\mathrm{tr}(\mathbf{V}^T\mathbf{M}_n\mathbf{V}) + \rho(\mathbf{V})\}, \text{ subject to } \mathbf{V}^T\mathbf{N}_n\mathbf{V} = \mathbf{I}_d, \tag{3.7}$$

where $\rho(\mathbf{V})$ is defined in (3.5).

The solution $\tilde{\mathbf{V}}$ is not unique as $\tilde{\mathbf{V}}\mathbf{O}$ is also a solution for any orthogonal matrix $\mathbf{O}$. In a strict sense, we are minimizing (3.7) over the span of the columns of $\mathbf{V}$. Thus $\tilde{\mathbf{V}}$ denotes any basis of the solution of (3.7). Analogously, the solution $\widehat{\mathbf{V}}$ is one basis of the solution of (3.6). Before proceeding, we rewrite (3.6) and (3.7) as equivalent unitary constrained optimization problems which will facilitate our exposition. We summarize the result into the following proposition without giving its proof since it follows from some straightforward algebra.

Proposition 4 *The minimizer (3.6) is equivalent to* $\widehat{\mathbf{V}} = \mathbf{N}_n^{-1/2}\widehat{\mathbf{\Gamma}}$ *where*

$$\widehat{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}} -\mathrm{tr}(\mathbf{\Gamma}^T\mathbf{G}_n\mathbf{\Gamma}), \text{ subject to } \mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_d. \tag{3.8}$$

Furthermore, $\mathbf{G}_n\widehat{\mathbf{\Gamma}} = \widehat{\mathbf{\Gamma}}\mathbf{\Lambda}_{n1}$, *where* $\mathbf{\Lambda}_{n1} = \mathrm{diag}(\lambda_{n1}, \ldots, \lambda_{nd})$. *Correspondingly, the minimizer (3.7) is equivalent to* $\tilde{\mathbf{V}} = \mathbf{N}_n^{-1/2}\tilde{\mathbf{\Gamma}}$, *where*

$$\tilde{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}}\{-\mathrm{tr}(\mathbf{\Gamma}^T\mathbf{G}_n\mathbf{\Gamma}) + \rho(\mathbf{N}_n^{-\frac{1}{2}}\mathbf{\Gamma})\}, \text{ subject to } \mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_d. \tag{3.9}$$

The minimization of (3.8) and (3.9) is a Grassmann manifold optimization problem. A Grassmann manifold, which is defined as the set of all d-dimensional subspaces in $\mathbb{R}^p$, is the natural parameter space for the $\mathbf{\Gamma}$ parametrization in (3.8). For more background on Grassmann manifold optimization, see Edelman et al. (1998). The traditional Grassmann manifold optimization techniques can not be applied directly to (3.9) due to the non-differentiability of $\rho(\cdot)$. Nevertheless, we have devised a simple and fast algorithm to solve (3.9), as discussed in the next subsection.

3.1.4 Algorithm

To overcome the non-differentiability of $\rho(\cdot)$, we adopt the local quadratic approximation of Fan and Li (2001); that is, we approximate the penalty function locally with a quadratic function at every step of the iteration as follows.

Let $\tilde{\mathbf{V}}^{(0)} = (\tilde{\mathbf{v}}_1^{(0)}, \ldots, \tilde{\mathbf{v}}_p^{(0)})^T = \mathbf{N}_n^{-1/2}\tilde{\mathbf{\Gamma}}^{(0)}$ be the starting value. The unconstrained first derivative of $\rho(\mathbf{V})$ with respect to the $p \times d$ matrix $\mathbf{V}$ is given by

$$\frac{\partial \rho}{\partial \mathbf{V}} = \mathrm{diag}\left(\frac{\theta_1}{\|\mathbf{v}_1\|_2}, \ldots, \frac{\theta_i}{\|\mathbf{v}_i\|_2}, \ldots, \frac{\theta_p}{\|\mathbf{v}_p\|_2}\right)\mathbf{V}.$$

Following Fan and Li, the first derivative of $\rho(\mathbf{V})$ around $\tilde{\mathbf{V}}^{(0)}$ can be approximated by:

$$\frac{\partial \rho}{\partial \mathbf{V}} \approx \text{diag}\left(\frac{\theta_1}{\|\tilde{\mathbf{v}}_1^{(0)}\|_2}, \ldots, \frac{\theta_i}{\|\tilde{\mathbf{v}}_i^{(0)}\|_2}, \ldots, \frac{\theta_p}{\|\tilde{\mathbf{v}}_p^{(0)}\|_2}\right)\mathbf{V} := \mathbf{H}^{(0)}\mathbf{V}.$$

By using the second-order Taylor expansion and some algebraic manipulation, we have

$$\rho(\mathbf{V}) \approx \frac{1}{2}\text{tr}(\mathbf{V}^T\mathbf{H}^{(0)}\mathbf{V}) + C_0 = \frac{1}{2}\text{tr}(\mathbf{\Gamma}^T\mathbf{N}_n^{-\frac{1}{2}}\mathbf{H}^{(0)}\mathbf{N}_n^{-\frac{1}{2}}\mathbf{\Gamma}) + C_0,$$

where C_0 stands for a constant with respect to $\mathbf{V}$.

Then find $\tilde{\mathbf{\Gamma}}^{(1)}$ by minimizing:

$$-\text{tr}(\mathbf{\Gamma}^T\mathbf{G}_n\mathbf{\Gamma}) + \frac{1}{2}\text{tr}(\mathbf{\Gamma}^T\mathbf{N}_n^{-\frac{1}{2}}\mathbf{H}^{(0)}\mathbf{N}_n^{-\frac{1}{2}}\mathbf{\Gamma}) = \text{tr}\{\mathbf{\Gamma}^T(-\mathbf{G}_n + \frac{1}{2}\mathbf{N}_n^{-\frac{1}{2}}\mathbf{H}^{(0)}\mathbf{N}_n^{-\frac{1}{2}})\mathbf{\Gamma}\}.$$

This minimization problem can be easily solved by the eigensystem analysis of the matrix $\mathbf{G}_n - 2^{-1}\mathbf{N}_n^{-1/2}\mathbf{H}^{(0)}\mathbf{N}_n^{-1/2}$, i.e., the columns of $\tilde{\mathbf{\Gamma}}^{(1)}$ are the first d principal component directions of $\mathbf{G}_n - 2^{-1}\mathbf{N}_n^{-1/2}\mathbf{H}^{(0)}\mathbf{N}_n^{-1/2}$. Next, let $\tilde{\mathbf{V}}^{(1)} = \mathbf{N}_n^{-1/2}\tilde{\mathbf{\Gamma}}^{(1)}$ and start the second round of approximation of $\rho(\mathbf{V})$. The procedures repeat until it converges. During the iterations, if $\|\tilde{\mathbf{v}}_i^{(k)}\|_2 \approx 0$, say $\|\tilde{\mathbf{v}}_i^{(k)}\|_2 < \epsilon$ where ϵ is a pre-specified small positive number (e.g., $\epsilon = 10^{-6}$), then the variable x_i is removed.

With respect to the choice of the initial values $\tilde{\mathbf{\Gamma}}^{(0)}$, a simple but effective solution is to use $\tilde{\mathbf{\Gamma}}^{(0)} = \widehat{\mathbf{\Gamma}}$, the minimizer of (3.8). With $\widehat{\mathbf{\Gamma}}$ as the initial values, we found that the frequency of nonconvergence is negligible in all of our simulation studies and the convergence is quite fast, usually requiring a few dozen iterations.

3.1.5 Oracle property

In what follows, without loss of generality, we assume that only the first q predictors are relevant to the regression, where $d \leq q < p$. Given a $p \times d$ matrix $\mathbf{K}$, $\mathbf{K}_{(q)}$ and $\mathbf{K}_{(p-q)}$ indicate the sub-matrices consisting of its first q and remaining $p-q$ rows. If $\mathbf{K}$ is $p \times p$ then the notation indicates its first q and the last $p-q$ block sub-matrices. In the context of the single-index model, Fan and Li (2001) and Zou (2006) have shown that, with the proper choice of the penalty functions and regularization parameters, the penalized likelihood estimators have the oracle property. With continuous penalty functions, the coefficient estimates that correspond to insignificant predictors must shrink towards 0 as the penalty parameter increases, and these estimates will be exactly 0 if that parameter

is sufficiently large. In this section we present theorems which establish the oracle property of CISE.

Let $a_n = \max\{\theta_j, j \leq q\}$ and $b_n = \min\{\theta_j, j > q\}$, where the θ_j's are the penalty parameters defined in Section 3.1.2., let $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ denote the eigenvalues of $\mathbf{G}$, and define the matrix norm $||\mathbf{V}||_s = \sqrt{\text{tr}(\mathbf{V}^T\mathbf{V})}$. We also require a metric D in the set of all subspaces of $\mathbb{R}^p$ (Gohberg et al. 2006):

Definition 1 *The distance between the subspaces spanned by the columns of* $\mathbf{V}_n$ *and* $\mathbf{V}$*, denoted as* $D(\mathbf{V}_n, \mathbf{V})$*, is defined as the square root of the largest eigenvalue of*

$$(P_{\mathbf{V}_n} - P_{\mathbf{V}})^T(P_{\mathbf{V}_n} - P_{\mathbf{V}}).$$

We use the following assumptions to establish the oracle property.

ASSUMPTION 1: Let $\mathbf{V}_0$ denote the minimizer of (3.6) when the population matrices $\mathbf{M}$ and $\mathbf{N}$ are used in place of $\mathbf{M}_n$ and $\mathbf{N}_n$. Then $\mathbf{V}_{0(p-q)} = 0$.

ASSUMPTION 2: $\mathbf{M}_n = \mathbf{M} + O_p(n^{-1/2})$ and $\mathbf{N}_n = \mathbf{N} + O_p(n^{-1/2})$.

REMARK 1. Given some mild method-specified conditions, the minimizer of (3.6) $\widehat{\mathbf{V}}$ is a consistent estimator of a basis of the central subspace. For example, SIR provides the consistent estimate of the central subspace given that the linearity and coverage conditions hold (Cook 1998a; Chiaromonte et al. 2002). Consequently, the population version $\mathbf{V}_0$ will be a basis of the central subspace. Therefore, Assumption 1 is a reasonable one which facilitates our following presentations. Assumption 2 is mild and typically holds. These two assumptions suffice for our main results.

We state our theorems here, but their proofs are relegated to the Appendix. The constrained objective function in the minimization problem (3.7) is denoted as $Q(\mathbf{V}; \mathbf{M}_n) := f(\mathbf{V}; \mathbf{M}_n) + \rho(\mathbf{V})$ where $f(\mathbf{V}; \mathbf{M}_n) = -\text{tr}(\mathbf{V}^T\mathbf{M}_n\mathbf{V})$. The first theorem establishes existence of CISE.

Theorem 1 *If Assumptions 1-2 hold,* $\lambda_d > \lambda_{d+1}$ *and* $\sqrt{n}a_n \xrightarrow{p} 0$*, then there exists a local minimizer* $\tilde{\mathbf{V}}_n$ *of* $Q(\mathbf{V}; \mathbf{M}_n)$ *subject to* $\mathbf{V}^T\mathbf{N}_n\mathbf{V} = \mathbf{I}_d$*, so that*

$$D(\tilde{\mathbf{V}}_n, \mathbf{V}_0) = O_p(n^{-1/2}).$$

It is clear from Theorem 1 that by choosing the θ_i's properly, there exists a root-n consistent CISE. The next transition theorem states an oracle-like property of CISE.

Theorem 2 *If Assumptions 1-2 hold,* $\lambda_d > \lambda_{d+1}$, $\sqrt{n}a_n \xrightarrow{p} 0$ *and* $\sqrt{n}b_n \xrightarrow{p} \infty$, *then the root-n consistent local minimizer* $\tilde{\mathbf{V}}_n$ *in Theorem 1 must satisfy*

(i) $\Pr(\tilde{\mathbf{V}}_{n(p-q)} = \mathbf{0}) \to 1$,

(ii) $\sqrt{n}D(\tilde{\mathbf{V}}_{n(q)}, \widehat{\mathbf{V}}_{n(O)}) = o_p(1)$, *where* $\widehat{\mathbf{V}}_{n(O)}$ *is the minimizer of* $Q(\mathbf{V}; \mathbf{M}_{n(q)})$ *subject to* $\mathbf{V}^T\mathbf{N}_{n(q)}\mathbf{V} = \mathbf{I}_d$.

Theorem 2 (i) states that with probability tending to 1, all of the zero rows of $\mathbf{V}_0$ must be estimated as $\mathbf{0}$. Theorem 2 (ii) tells us that there exist a local minimizer $\tilde{\mathbf{V}}_n$ so that the difference between its non-zero submatrix $\tilde{\mathbf{V}}_{n(q)}$ and $\widehat{\mathbf{V}}_{n(O)}$ is of order $o_p(n^{-1/2})$. That is to say, we have the result that $\sqrt{n}D(\tilde{\mathbf{V}}_{n(q)}, \mathbf{V}_{0(q)})$ has the same asymptotic distribution as $\sqrt{n}D(\widehat{\mathbf{V}}_{n(O)}, \mathbf{V}_{0(q)})$. With respect to the asymptotic distribution of $\widehat{\mathbf{V}}_{n(O)}$, there seems to be no general result in the literature because different specifications on $\mathbf{M}_{n(q)}$ and $\mathbf{N}_{n(q)}$ yield different asymptotic distributions. This is not of great interest here and we refer to Zhu and Ng (1995), Li and Zhu (2007) and the references therein.

The second part of Theorem 2 is actually valid in a generalized sense. The OSDRE in the exact oracle property, denoted as $\dot{\mathbf{V}}_{n(O)}$, is obtained by using the $q \times q$ $\mathbf{M}_n$ and $\mathbf{N}_n$ formed with the first q variables (denoted as $\mathbf{M}_{n(O)}$ and $\mathbf{N}_{n(O)}$). Usually, $\mathbf{N}_{n(O)} = \mathbf{N}_{n(q)}$. From the definition, it is straightforward to see that $\mathbf{M}_{n(O)} = \mathbf{M}_{n(q)}$ for the PCA, SIR and PFC methods. Thus, in these cases, Theorem 2 establishes the exact oracle property. We conjecture that $\mathbf{M}_{n(O)}$ should be very close to $\mathbf{M}_{n(q)}$ for any SDR method that satisfies Assumptions 1 and 2. From the proof of Theorem 2 (ii), we can conclude that if

$$||\mathbf{M}_{n(O)} - \mathbf{M}_{n(q)}||_s = O_p(a_n), \tag{3.10}$$

the exact oracle property still holds. The next result establishes that the condition above holds for DR and SAVE under certain conditions.

Proposition 5 *Suppose the linearity and constant variance conditions (Li and Wang, 2007) hold and* $(na_n)^{-1} = O_p(1)$. *Then the condition (3.10) is satisfied for the DR and SAVE methods.*

By this proposition, Theorem 2 and the discussion above, we know that from asymptotic viewpoints CISE is effective for all of the commonly used SDR methods. We summarize this major result in the following theorem.

Theorem 3 *Assume that the conditions in Theorem 2 and Proposition 5 hold. Then the exact oracle property is achieved for the PCA, SIR, PFC, SAVE and DR methods. That is,* $\tilde{\mathbf{V}}_n$ *has the selection consistency and* $\sqrt{n}D(\tilde{\mathbf{V}}_{n(q)}, \mathbf{V}_{0(q)})$ *has the same asymptotic distribution as* $\sqrt{n}D(\dot{\mathbf{V}}_{n(O)}, \mathbf{V}_{0(q)})$.

In this paper, we make no attempt to further analysis general conditions for the validity of (3.10), but we think that such studies certainly warrant future research.

3.1.6 Choice of tuning parameters

We recommend using

$$\theta_i = \theta ||\widehat{\mathbf{v}}_i||_2^{-r}, \tag{3.11}$$

where $\widehat{\mathbf{v}}_i$ is the ith row vector of the OSDRE $\widehat{\mathbf{V}}$ defined in (3.6), and $r > 0$ is some pre-specified parameter. Following the suggestions of Zou (2006), $r = 0.5$ is used in both the simulation study and the illustration in Section 3.3. Such a strategy effectively transforms the original p-dimensional tuning parameter selection problem into a univariate one. By Lemma 2 in the Appendix, $\widehat{\mathbf{v}}_i$ is root-n consistent. Thus, it is easily to verify that the tuning parameter defined in (3.11) satisfies the conditions on a_n and b_n needed by Theorem 2 as long as $\sqrt{n}\theta \to 0$ and $n^{(1+r)/2}\theta \to \infty$. Hence, it suffices to select $\theta \in [0, +\infty)$ only.

To choose the tuning parameter θ, we use the following criterion which has a form similar to ones used by L. Li (2007) and Leng and Wang (2009),

$$-\mathrm{tr}(\tilde{\mathbf{V}}_\theta^T \mathbf{M}_n \tilde{\mathbf{V}}_\theta) + \gamma \cdot \mathrm{df}_\theta,$$

where $\tilde{\mathbf{V}}_\theta$ denotes the solution for $\mathbf{V}$ given θ, df_θ denotes the effective number of parameters, and $\gamma = 2/n$ for AIC-type and $\gamma = \log(n)/n$ for BIC-type criteria. Following the discussion of Li (2007), we estimate df_θ by $(p_\theta - d) \cdot d$ where p_θ denotes the number of non-zero rows of $\tilde{\mathbf{V}}_\theta$ because we need $(p_\theta - d) \cdot d$ parameters to describe a d-dimensional Grassmann manifold in $\mathbb{R}^{p_\theta}$ (Edelman et al. 1998).

3.2 Simulation studies

We report the results of four simulation studies in this section, three of which were conducted using forward regression models and one was conducted using an inverse regression model. We compared our method with the C^3 method (Zhou and He 2008) and the SSIR method (Ni et al. 2005). BIC and RIC (Shi and Tsai 2002) were used in SSIR to select the tuning parameters, and two α levels (0.01 and 0.005) were used in the C^3 method. We used SIR and PFC to generate $\mathbf{M}_n$ and $\mathbf{N}_n$ for CISE selection. For these methods, denoted CIS-SIR and CIS-PFC, we report only the results using the BIC criterion to select tuning parameters as we tend to believe that BIC has consistency property. Unreported simulations using the RIC criterion show slightly better performance in some cases though.

In each study, we generated 2500 datasets with the sample size $n = 60$ and $n = 120$. For the C^3 method, the quadratic spline with four internal knots was used, as suggested by Zhou and He (2008). Six slices were used for the SSIR method. We calculated $\mathbf{M}_n$ in the PFC model setting using $f(y) = (|y|, y, y^2)^T$ for all simulation studies.

We used three summary statistics – r_1, r_2 and r_3 – to assess how well the methods select variables: r_1 is the average fraction of non-zero rows of $\tilde{\mathbf{V}}$ associated with relevant predictors; r_2 is the average fraction of zero rows of $\tilde{\mathbf{V}}$ associated with irrelevant predictors; and r_3 is the fraction of runs in which the methods select both relevant and irrelevant predictors exactly right.

Study 1

$$y = x_1 + x_2 + x_3 + 0.5\epsilon,$$

where $\epsilon \sim N(0,1)$, $\mathbf{x} = (x_1, \ldots, x_{24})^T \sim N(0, \mathbf{\Sigma})$ with $\Sigma_{ij} = 0.5^{|i-j|}$ for $1 \leq i, j \leq 24$, and $\mathbf{x}$ and ϵ are independent. In this study, the central subspace is spanned by the direction $\boldsymbol{\beta}_1 = (1, 1, 1, 0, \ldots, 0)^T$ with twenty-one zero coefficients.

Study 2

$$y = x_1 + x_2 + x_3 + 2\epsilon,$$

where $\epsilon \sim N(0,1)$, $\mathbf{x} = (x_1, \ldots, x_{24})^T \sim N(0, \mathbf{\Sigma})$ with $\Sigma_{ij} = 0.5^{|i-j|}$ for $1 \leq i, j \leq 24$, and x and ϵ are independent. In this study, the central subspace is spanned by the

Table 3.2: Variable selection summary of Study 1

Method	CIS-SIR	CIS-PFC	C^3		SSIR	
Criterion	BIC	BIC	$\alpha = 0.01$	$\alpha = 0.005$	BIC	RIC
Sample size			$n = 60$			
r_1	0.991	1.000	1.000	1.000	0.993	0.974
r_2	0.999	1.000	0.999	0.999	0.997	0.999
r_3	0.970	1.000	0.978	0.991	0.939	0.914
Sample size			$n = 120$			
r_1	1.000	1.000	1.000	1.000	1.000	1.000
r_2	1.000	1.000	1.000	1.000	0.999	1.000
r_3	1.000	1.000	1.000	1.000	0.994	1.000

direction $\boldsymbol{\beta}_1 = (1, 1, 1, 0, \ldots, 0)^T$ with twenty-one zero coefficients. In short, this study was identical to the first, except the error was increased by a factor of 4.

Table 3.3: Variable selection summary of Study 2

Method	CIS-SIR	CIS-PFC	C^3		SSIR	
Criterion	BIC	BIC	$\alpha = 0.01$	$\alpha = 0.005$	BIC	RIC
Sample size			$n = 60$			
r_1	0.713	0.795	0.583	0.565	0.770	0.706
r_2	0.988	0.992	0.998	0.998	0.881	0.939
r_3	0.233	0.399	0.075	0.080	0.058	0.104
Sample size			$n = 120$			
r_1	0.909	0.951	0.669	0.615	0.973	0.930
r_2	0.998	0.998	1.000	1.000	0.928	0.981
r_3	0.694	0.827	0.209	0.131	0.244	0.554

STUDY 3

$$y = x_1/\{0.5 + (x_2 + 1.5)^2\} + 0.2\epsilon,$$

where $\epsilon \sim N(0, 1)$, $\mathbf{x} = (x_1, \ldots, x_{24})^T \sim N(0, \boldsymbol{\Sigma})$ with $\Sigma_{ij} = 0.5^{|i-j|}$ for $1 \leq i, j \leq 24$, and x and ϵ are independent. In this study, the central subspace is spanned by the directions $\boldsymbol{\beta}_1 = (1, 0, \ldots, 0)^T$ and $\boldsymbol{\beta}_2 = (0, 1, \ldots, 0)^T$.

Table 3.4: Variable selection summary of Study 3

Method	CIS-SIR	CIS-PFC	C^3		SSIR	
Criterion	BIC	BIC	$\alpha = 0.01$	$\alpha = 0.005$	BIC	RIC
Sample size			$n = 60$			
r_1	0.789	0.906	0.770	0.742	0.934	0.888
r_2	0.965	0.979	0.948	0.955	0.633	0.828
r_3	0.344	0.588	0.229	0.226	0.000	0.004
Sample size			$n = 120$			
r_1	0.948	0.995	0.839	0.781	0.994	0.983
r_2	0.992	0.998	0.956	0.963	0.664	0.865
r_3	0.838	0.973	0.309	0.245	0.001	0.027

STUDY 4

$$\mathbf{x} = \mathbf{\Gamma}(y, y^2)^T + \mathbf{\Delta}^{1/2}\boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}_{24})$, $y \sim N(0,1)$, $\Delta_{ij} = 0.5^{|i-j|}$ for $1 \leq i, j \leq 24$, and y and $\boldsymbol{\epsilon}$ are independent. The first column of $\mathbf{\Gamma}$ is $(0.5, 0.5, 0.5, 0.5, 0, \ldots, 0)^T$ and the second column of $\mathbf{\Gamma}$ is $(0.5, -0.5, 0.5, -0.5, 0, \ldots, 0)^T$. In this study, the central subspace is the column space of $\mathbf{\Delta}^{-1}\mathbf{\Gamma}$.

Table 3.5: Variable selection summary of Study 4

Method	CIS-SIR	CIS-PFC	C^3		SSIR	
Criterion	BIC	BIC	$\alpha = 0.01$	$\alpha = 0.005$	BIC	RIC
Sample size			$n = 60$			
r_1	0.676	0.817	0.670	0.643	0.871	0.776
r_2	0.968	0.989	0.956	0.958	0.641	0.832
r_3	0.069	0.327	0.022	0.029	0.000	0.000
Sample size			$n = 120$			
r_1	0.805	0.928	0.828	0.809	0.988	0.964
r_2	0.993	0.998	0.967	0.969	0.696	0.890
r_3	0.299	0.687	0.147	0.178	0.000	0.000

The simulation results from these four studies are summarized in Tables 3.2-3.5 respectively. The standard errors of the r_k's, $\sqrt{r_k(1-r_k)}/50$, are typically less than 0.01 throughout this section. In Study 1, the signal-to-noise ratio is close to 5 (the ratio of the standard deviation of $x_1 + x_2 + x_3$ to 0.5). Because of the large signal-to-noise

ratio, all the considered methods show very good performance, but CIS-SIR, CIS-PFC and C^3 perform slightly better than SSIR. In Study 2 we decreased the signal-to-noise ratio to about 1.2 and now CIS-SIR and CIS-PFC perform much better than C^3 and SSIR. In both Studies 3 and 4, CISE is generally superior to the other two methods, especially for CIS-PFC and the rate r_3. It should be pointed out that the superiority of CISE becomes more significant when n gets larger. When $n = 120$, C^3 still cannot perform exact identifications well, while SSIR rarely identifies all relevant and irrelevant variables correctly.

While both CISE and C^3 have the oracle property, they differ in many aspects. CISE is a unified method that can be applied to many popular sufficient dimension reduction methods, including PCA, PFC, SIR, SAVE and DR. On the other hand, C^3 is based on one specified sufficient dimension reduction method, canonical correlation (Fung et al. 2002). We regard r_3, the estimated probability all relevant and irrelevant variables are identified correctly, as the most important aspect of a method. On that measure CISE typically dominates C^3. There was only one case (Table 3.2, $n = 60$) in which C^3 did slightly better than CISE. Additionally, CISE seems conceptually simpler and is easily implemented.

3.3 Boston housing data

3.3.1 Variable screening

We applied our method to the Boston housing data, which has been widely studied in the literature. The Boston housing data contains 506 observations, and can be downloaded from the web site $http://lib.stat.cmu.edu/datasets/boston\_corrected.txt$. The response variable y is the median value of owner-occupied homes in each of the 506 census tracts in the Boston Standard Metropolitan Statistical Areas. The 13 predictor variables are per capita crime rate by town (x_1); proportion of residential land zoned for lots over 25,000 sq.ft (x_2); proportion of non retail business acres per town (x_3); Charles River dummy variable (x_4); nitric oxides concentration (x_5); average number of rooms per dwelling (x_6); proportion of owner-occupied units built prior to 1940 (x_7); weighted distances to five Boston employment centers (x_8); index of accessibility to radial highways (x_9); full-value property-tax rate (x_{10}); pupil-teacher ratio by town

(x_{11}); proportion of blacks by town (x_{12}); percentage of lower status of the population (x_{13}).

Table 3.6: Estimated bases of the central subspace in Boston housing data

Method	CIS-SIR		CIS-PFC		C^3		SSIR-BIC		SSIR-RIC	
x_1	0	0	0	0	0	0	-0.050	-0.131	-0.041	-0.123
x_2	-0.004	-0.047	0	0	0	0	-0.001	0.002	-0.001	-0.001
x_3	0	0	0	0	0	0	0.001	0.005	0	0
x_4	0	0	0	0	0	0	-0.033	0.020	0	0
x_5	0	0	0	0	0	0	0.719	-0.882	0.543	-0.765
x_6	-0.999	0.034	-0.999	0.034	0.962	-0.645	-0.684	-0.448	-0.834	-0.627
x_7	-0.008	-0.139	-0.003	-0.077	-0.174	-0.096	0.006	-0.001	0.005	-0.001
x_8	0	0	0	0	0	0	0.082	-0.012	0.060	-0.010
x_9	0	0	0	0	0	0	-0.019	0.035	-0.016	0.033
x_{10}	-0.001	-0.01	-0.002	-0.035	-0.166	0	0.001	-0.001	0.001	-0.001
x_{11}	0.021	-0.361	0.018	-0.280	-0.126	0	0.058	-0.033	0.055	-0.036
x_{12}	0.001	0.011	0.002	0.035	0	0	-0.000	0.000	0	0
x_{13}	-0.044	-0.920	-0.040	-0.955	0	-0.758	0.014	-0.043	0.017	-0.059

Previous studies suggested that we remove those observation with crime rate greater than 3.2, as a few predictors remain constant except for 3 observations in this case (Li 1991). So we used the 374 observations with crime rate smaller than 3.2 in this analysis. All the methods considered in Section 3.2 were applied to this dataset. Scatter-plotting of each predictor against y, we concluded that it would be sufficient to use $\mathbf{f} = (\sqrt{y}, y, y^2)^T$ in the PFC model. Since PFC is a scale-invariant method, we did not standardize the data as many other methods do. Similar to the previous studies in the literature, we pick up two directions to estimate the central subspace. The estimated bases of the central subspace for all the considered methods are summarized in Table 3.6.

The coefficients in Table 3.6 from CIS-SIR, CIS-PFC and SSIR are based on the original dataset, while the coefficients of C^3 is based on a data-specific weighted version (Zhou and He, 2008). As suggested by CIS-PFC, explanatory variables x_6, x_7, x_{10}, x_{11}, x_{12} and x_{13} would be important in explaining y.

3.3.2 Bootstrap study

In Table 3.7, we used the bootstrap to assess the accuracy of variable selection for all methods except C^3, as it is not clear how the weighting procedure used by Zhou and He should be automated. Without weighting we encountered serious convergence problems in the C^3 algorithm. This bootstrap study can be considered as another simulation study.

The bootstrap procedure was conducted as follows. Firstly, we randomly chose with replacement 374 observations for y jointly with x_6, x_7, x_{10}, x_{11}, x_{12} and x_{13}. Secondly, we separately randomly selected 374 observations for x_1, x_2, x_3, x_4, x_5, x_8 and x_9. Then we combine them to make one complete bootstrap dataset. In this way, we mimic the results of the analysis of original data, forcing x_1, x_2, x_3, x_4, x_5, x_8 and x_9 to be irrelevant. This procedure was repeated 2500 times. The resulting rates r_1, r_2 and r_3 are shown in Table 3.7. The results show a pattern similar to those in simulation studies and again CISE performed quite well.

Table 3.7: Variable selection in bootstrapping Boston housing data

Method	CIS-SIR	CIS-PFC	SSIR-BIC	SSIR-RIC
r_1	0.947	0.962	0.963	0.877
r_2	0.969	0.980	0.780	0.952
r_3	0.550	0.672	0.118	0.264

3.4 A Matlab package

A Matlab interface was used to implement the CISE algorithm described in Section 3.1.4. The programs can be obtained upon request or be downloaded through this link: http://www.stat.umn.edu/∼xchen/cise.zip. The detail of the usages can be founded in the package.

3.5 Discussion

The establishment of the oracle property in this paper takes advantage of the simple trace form of the objective function: $-\mathrm{tr}(\mathbf{V}^T\mathbf{M}_n\mathbf{V})$. However we believe that the proof

in the Appendix can be extended to more general objective functions. Moreover, it is also of great interest to see whether CISE and its oracle property are still valid in high-dimensional settings in which $p > n$.

We have seen that $\mathbf{N}_n$ usually takes the form of the marginal sample covariance matrix of $\mathbf{x}$, while $\mathbf{M}_n$ depends on the specific method. In practice, how to choose $\mathbf{M}_n$ for variable selection is an important issue and merits thorough investigation. In addition, it is well demonstrated that for the multiple regression model, the BIC criterion tends to identify the true sparse model well if the true model is included in the candidate set (Wang et al. 2007). The consistency of the BIC criterion proposed in Section 3.1.6 deserves further study as well.

3.6 Appendix

Throughout this section, we will use the following notation for ease of exposition. $Q(\mathbf{\Gamma}; \mathbf{G}_n, \mathbf{N}_n) := -\text{tr}(\mathbf{\Gamma}^T \mathbf{G}_n \mathbf{\Gamma}) + \rho(\mathbf{N}_n^{-1/2}\mathbf{\Gamma})$ denotes the constrained objective function in the minimization problem (3.9). Unless otherwise stated, we also use the generic notation $Q(\mathbf{\Gamma})$ or $Q(\mathbf{V})$ to represent the function $Q(\mathbf{\Gamma}; \mathbf{G}_n, \mathbf{N}_n)$ or $Q(\mathbf{V}; \mathbf{M}_n)$ for abbreviation, which should not cause any confusion. $\mathbf{1}_i$ denotes a row vector with one in the i-th position and zero in the others.

Proof of Proposition 3: Cook (2007) has shown that the maximum likelihood estimator of span$(\mathbf{\Delta}^{-1}\mathbf{\Gamma})$ in the general PFC model equals the span of $\{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$, where $\mathbf{e}_i = \mathbf{\Sigma}_n^{-1/2}\mathbf{r}_i$ and $\mathbf{r}_i$ is the ith eigenvector of $\mathbf{\Sigma}_n^{-1/2}\widehat{\mathbf{\Sigma}}_{\text{fit}}\mathbf{\Sigma}_n^{-1/2}$ corresponding to the eigenvalue k_i. Consequently, we have

$$\widehat{\mathbf{\Sigma}}_{\text{fit}}\mathbf{e}_i = k_i \mathbf{\Sigma}_n \mathbf{e}_i.$$

It follows that $\mathbf{M}_n = \widehat{\mathbf{\Sigma}}_{\text{fit}}$ and $\mathbf{N}_n = \mathbf{\Sigma}_n$. □

In order to prove the theorems, we firstly state a few necessary lemmas. For notation convenience, we need the following additional definitions. Define the Stiefel manifold $St(p, d)$ as

$$St(p, d) = \{\mathbf{\Gamma} \in \mathbb{R}^{p \times d} : \mathbf{\Gamma}^T \mathbf{\Gamma} = \mathbf{I}_d\}.$$

Denotes $\lfloor \mathbf{\Gamma} \rfloor$ as the subspace spanned by the columns of $\mathbf{\Gamma}$, then $\lfloor \mathbf{\Gamma} \rfloor \in Gr(p,d)$ where $Gr(p,d)$ stands for the Grassmann manifold. The projection operator $R : \mathbb{R}^{p\times d} \to St(p,d)$ onto the Stiefel manifold $St(p,d)$ is defined to be

$$R(\mathbf{\Gamma}) = \arg \min_{\mathbf{W}\in St(p,d)} ||\mathbf{\Gamma} - \mathbf{W}||_s^2.$$

The tangent space $T_{\mathbf{\Gamma}}(p,d)$ of $\mathbf{\Gamma} \in St(p,d)$ is defined by

$$T_{\mathbf{\Gamma}}(p,d) = \left\{ \mathbf{Z} \in \mathbb{R}^{p\times d} : \mathbf{Z} = \mathbf{\Gamma}\mathbf{A} + \mathbf{\Gamma}_\perp \mathbf{B}, \mathbf{A} \in \mathbb{R}^{d\times d}, \mathbf{A} + \mathbf{A}^T = 0, \mathbf{B} \in \mathbb{R}^{(p-d)\times d} \right\}, \tag{3.12}$$

where $\mathbf{\Gamma}_\perp \in \mathbb{R}^{p\times(p-d)}$ is the complement of $\mathbf{\Gamma}$ satisfies $[\mathbf{\Gamma}\ \mathbf{\Gamma}_\perp]^T[\mathbf{\Gamma}\ \mathbf{\Gamma}_\perp] = \mathbf{I}_p$.

Lemma 1 *If $\mathbf{Z} \in T_{\mathbf{\Gamma}}(p,d), \mathbf{\Gamma} \in St(p,d)$, we have*

(i) For any symmetric matrix $\mathbf{C} \in \mathbb{R}^{d\times d}$, $\mathrm{tr}(\mathbf{Z}^T\mathbf{\Gamma}\mathbf{C}) = 0$.

(ii) $R(\mathbf{\Gamma} + t\mathbf{Z}) = \mathbf{\Gamma} + t\mathbf{Z} - (1/2)t^2\mathbf{\Gamma}\mathbf{Z}^T\mathbf{Z} + O(t^3)$.

This lemma comes from Lemma 10 and Proposition 12 of Manton (2002).

Lemma 2 *Under conditions in Theorem 1, we have*

$$D(\widehat{\mathbf{\Gamma}}, \mathbf{\Gamma}_0) = O_p(n^{-1/2}),$$

where $\mathbf{\Gamma}_0$ denotes any minimizer of (3.8) when $\mathbf{G}_n$ is taken as the population matrix $\mathbf{G}$.

This lemma can be proved in a similar fashion to the proof of Theorem 1 and hence omitted here.

Proof of Theorem 1: Clearly, to prove this theorem is equivalent to show there exists a local minimizer $\tilde{\mathbf{\Gamma}}_n$ of $Q(\mathbf{\Gamma}; \mathbf{G}_n, \mathbf{N}_n)$ subject to $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_d$, so that

$$D(\tilde{\mathbf{\Gamma}}_n, \mathbf{\Gamma}_0) = O_p(n^{-1/2}),$$

Denote $\mathbf{\Gamma}_*$ as an orthonormal basis matrix of the subspace spanned by the columns of $\mathbf{N}_n^{1/2}\mathbf{V}_0$. Thus, there exists a positive-definite matrix $\mathbf{O} \in \mathbb{R}^{d\times d}$ so that $\mathbf{\Gamma}_* = \mathbf{N}_n^{1/2}\mathbf{V}_0\mathbf{O}$. By Assumption 2 and $\mathbf{V}_0^T\mathbf{N}\mathbf{V}_0 = \mathbf{I}_d$, we have

$$\mathbf{O}^T\mathbf{O} = \mathbf{I}_d + O_p(n^{-1/2}).$$

Note that $\mathbf{\Gamma}_0 = \mathbf{N}^{1/2}\mathbf{V}_0$, and thus it is equivalent to show that

$$D(\tilde{\mathbf{\Gamma}}_n, \mathbf{\Gamma}_*) = O_p(n^{-1/2}),$$

since $D(\mathbf{\Gamma}_*, \mathbf{\Gamma}_0) = O_p(n^{-1/2})$ and $D(\cdot, \cdot)$ satisfies the triangle inequality.

To ease demonstration, we need define the concept of the neighborhood of $\lfloor \mathbf{\Gamma}_* \rfloor$. For an arbitrary matrix $\mathbf{W} \in \mathbb{R}^{p\times d}$ and scaler $\delta \in \mathbb{R}$, the perturbed point around $\mathbf{\Gamma}_*$ in Stiefel manifold can be expressed by $R(\mathbf{\Gamma}_* + \delta\mathbf{W})$. The perturbed point around $\lfloor \mathbf{\Gamma}_* \rfloor$ in Grassmann manifold can be expressed by $\lfloor R(\mathbf{\Gamma}_* + \delta\mathbf{W}) \rfloor$. According to Lemma 8 of Manton (2002), $\mathbf{W}$ can be uniquely decomposed as

$$\mathbf{W} = \mathbf{\Gamma}_*\mathbf{A} + \mathbf{\Gamma}_{*\perp}\mathbf{B} + \mathbf{\Gamma}_*\mathbf{C},$$

where $\mathbf{A} \in \mathbb{R}^{d\times d}$ is a skew-symmetric matrix, $\mathbf{B} \in \mathbb{R}^{(p-d)\times d}$ is an arbitrary matrix, and $\mathbf{C} \in \mathbb{R}^{d\times d}$ is a symmetric matrix. Let $\mathbf{Z} = \mathbf{\Gamma}_*\mathbf{A} + \mathbf{\Gamma}_{*\perp}\mathbf{B}$. Obviously, $\mathbf{Z} \in T_{\mathbf{\Gamma}_*}(p, d)$. Henceforth, $\mathbf{Z}$ refers to the projection of an arbitrary matrix $\mathbf{W} \in \mathbb{R}^{p\times d}$ onto the tangent space $T_{\mathbf{\Gamma}_*}(p, d)$, unless otherwise stated.

From Proposition 20 of Manton (2002), it is straightforward to see

$$\begin{aligned}
\lfloor R(\mathbf{\Gamma}_* + \delta\mathbf{W}) \rfloor &= \lfloor R(\mathbf{\Gamma}_* + \delta(\mathbf{\Gamma}_*\mathbf{A} + \mathbf{\Gamma}_{*\perp}\mathbf{B} + \mathbf{\Gamma}_*\mathbf{C})) \rfloor \\
&= \lfloor R(\mathbf{\Gamma}_*(\mathbf{I}_d + \delta(\mathbf{A} + \mathbf{C})) + \delta\mathbf{\Gamma}_{*\perp}\mathbf{B}) \rfloor \\
&= \lfloor \mathbf{\Gamma}_*(\mathbf{I}_d + \delta(\mathbf{A} + \mathbf{C})) + \delta\mathbf{\Gamma}_{*\perp}\mathbf{B} \rfloor \\
&= \lfloor \mathbf{\Gamma}_* + \delta\mathbf{\Gamma}_{*\perp}\mathbf{B}(\mathbf{I}_d + \delta(\mathbf{A} + \mathbf{C}))^{-1} \rfloor = \lfloor R(\mathbf{\Gamma}_* + \delta\mathbf{\Gamma}_{*\perp}\mathbf{B}') \rfloor,
\end{aligned}$$

provided that δ is sufficiently small so that $\mathbf{I}_d + \delta(\mathbf{A} + \mathbf{C})$ is a full rank matrix, where $\mathbf{B}' = \mathbf{B}(\mathbf{I}_d + \delta(\mathbf{A} + \mathbf{C}))^{-1}$. Since $\mathbf{B} \in \mathbb{R}^{(p-d)\times d}$ is an arbitrary matrix and we don't need the specific form of $\mathbf{B}$ and $\mathbf{B}'$ in our proof, we only use $\mathbf{B}$ for notation convenience. This tells us that the movement from $\lfloor \mathbf{\Gamma}_* \rfloor$ in the near neighborhood only depends on the $\mathbf{\Gamma}_{*\perp}\mathbf{B}$. In other words, it suffices to only consider perturbed points like $R(\mathbf{\Gamma}_* + \delta\mathbf{Z})$ in the following proofs, where $||\mathbf{B}||_s = C$ for some given C. It is worth noting that though our problems essentially are Grassmann manifold optimization, we prove the theorem in a more general way, say in Stiefel manifold (using $\mathbf{Z} \in T_{\mathbf{\Gamma}_*}(p, d)$) since the latter has simpler matrix expressions and thus is more notationally convenient.

For any small ϵ, if we can show that there exits a sufficiently large constant C, such that

$$\lim_n \Pr\left(\inf_{\mathbf{Z}\in T_{\mathbf{\Gamma}_*}(p,d):||\mathbf{B}||_s=C} Q(R(\mathbf{\Gamma}_* + n^{-\frac{1}{2}}\mathbf{Z})) > Q(\mathbf{\Gamma}_*)\right) > 1-\varepsilon, \tag{3.13}$$

then we can conclude that there exists a local minimizer $\tilde{\mathbf{\Gamma}}_n$ of $Q(\mathbf{\Gamma})$ with arbitrarily large probabilities such that $||\tilde{\mathbf{\Gamma}}_n - \mathbf{\Gamma}_*||_s = O_p(n^{-1/2})$. This certainly implies that $D(\tilde{\mathbf{\Gamma}}_n, \mathbf{\Gamma}_*) = O_p(n^{-1/2})$ by Definition 1.

By using Lemma 1, for $\mathbf{Z} \in T_{\mathbf{\Gamma}_*}(p,d)$ we have

$$\begin{aligned}
&n\Big\{Q(R(\mathbf{\Gamma}_* + n^{-\frac{1}{2}}\mathbf{Z})) - Q(\mathbf{\Gamma}_*)\Big\} \\
&= \left[-\mathrm{tr}(\mathbf{Z}^T\mathbf{G}_n\mathbf{Z}) - 2\sqrt{n}\mathrm{tr}(\mathbf{Z}^T\mathbf{G}_n\mathbf{\Gamma}_*) + \mathrm{tr}(\mathbf{Z}^T\mathbf{Z}\mathbf{\Gamma}_*^T\mathbf{G}_n\mathbf{\Gamma}_*)\right](1+o_p(1)) \\
&\quad + n\sum_{j=1}^{p}\left[\theta_j||\mathbf{1}_j\mathbf{N}_n^{-\frac{1}{2}}(\mathbf{\Gamma}_* + n^{-\frac{1}{2}}\mathbf{Z} - \frac{1}{2}n^{-1}\mathbf{\Gamma}_*\mathbf{Z}^T\mathbf{Z})||_2 - \theta_j||\mathbf{1}_j\mathbf{N}_n^{-\frac{1}{2}}\mathbf{\Gamma}_*||_2\right](1+o_p(1)) \\
&\geq \left[-\mathrm{tr}(\mathbf{Z}^T\mathbf{G}_n\mathbf{Z}) - 2\sqrt{n}\mathrm{tr}(\mathbf{Z}^T\mathbf{G}_n\mathbf{\Gamma}_*) + \mathrm{tr}(\mathbf{Z}^T\mathbf{Z}\mathbf{\Gamma}_*^T\mathbf{G}_n\mathbf{\Gamma}_*)\right](1+o_p(1)) \\
&\quad + n\sum_{j=1}^{q}\left[\theta_j\left(||\mathbf{1}_j\mathbf{N}_n^{-\frac{1}{2}}(\mathbf{\Gamma}_* + n^{-\frac{1}{2}}\mathbf{Z} - \frac{1}{2}n^{-1}\mathbf{\Gamma}_*\mathbf{Z}^T\mathbf{Z})||_2 - ||\mathbf{1}_j\mathbf{N}_n^{-\frac{1}{2}}\mathbf{\Gamma}_*||_2\right)\right](1+o_p(1)) \\
&\geq \left[-\mathrm{tr}(\mathbf{Z}^T\mathbf{G}_n\mathbf{Z}) + \mathrm{tr}(\mathbf{Z}^T\mathbf{Z}\mathbf{\Gamma}_*^T\mathbf{G}_n\mathbf{\Gamma}_*) - 2\sqrt{n}\mathrm{tr}(\mathbf{Z}^T\mathbf{G}_n\mathbf{\Gamma}_*)\right](1+o_p(1)) \\
&\quad - \frac{1}{2}q(\sqrt{n}a_n)\max_j\{||\mathbf{1}_j\mathbf{N}_n^{-\frac{1}{2}}\mathbf{\Gamma}_*||_2^{-1}\cdot||\mathbf{1}_j\mathbf{N}_n^{-\frac{1}{2}}(\mathbf{Z} - (1/2)n^{-1/2}\mathbf{\Gamma}_*\mathbf{Z}^T\mathbf{Z})||_2\} \\
&:= (\Delta_1 + \Delta_2)(1+o_p(1)),
\end{aligned}$$

where the second inequality holds because $\mathbf{1}_j\mathbf{N}_n^{-1/2}\mathbf{\Gamma}_* = 0$ for any $j > q$ by Assumption 1, and the last inequality comes from first-order Taylor expansion and the definition of a_n. In addition, according to the theorem's condition $\sqrt{n}a_n \xrightarrow{p} 0$, we known that Δ_2 is $o_p(1)$. Furthermore, based on Lemma 1 and Assumption 2, we have

$$\begin{aligned}
\sqrt{n}\mathrm{tr}(\mathbf{Z}^T\mathbf{G}_n\mathbf{\Gamma}_*) =& \sqrt{n}\mathrm{tr}(\mathbf{Z}^T\mathbf{G}\mathbf{\Gamma}_0\mathbf{O}) + \sqrt{n}\mathrm{tr}(\mathbf{Z}^T(\mathbf{G}_n\mathbf{N}_n^{\frac{1}{2}}\mathbf{N}^{-\frac{1}{2}} - \mathbf{G})\mathbf{\Gamma}_0\mathbf{O}) \\
=& \sqrt{n}\mathrm{tr}(\mathbf{Z}^T\mathbf{\Gamma}_0\mathbf{\Lambda}_1\mathbf{O}) + \sqrt{n}\mathrm{tr}(\mathbf{Z}^T(\mathbf{G}_n - \mathbf{G})\mathbf{\Gamma}_0\mathbf{O}) \\
&+ \sqrt{n}\mathrm{tr}(\mathbf{Z}^T\mathbf{G}_n\mathbf{\Gamma}_0\mathbf{O})\cdot O_p(n^{-\frac{1}{2}}) \\
=& \sqrt{n}\mathrm{tr}(\mathbf{Z}^T(\mathbf{G}_n - \mathbf{G})\mathbf{\Gamma}_0\mathbf{O}) + O_p(n^{-\frac{1}{2}}) \\
=& \sqrt{n}\mathrm{tr}(\mathbf{A}^T\mathbf{\Gamma}_0^T(\mathbf{G}_n - \mathbf{G})\mathbf{\Gamma}_0\mathbf{O}) + \sqrt{n}\mathrm{tr}(\mathbf{B}^T\mathbf{\Gamma}_{0\perp}^T(\mathbf{G}_n - \mathbf{G})\mathbf{\Gamma}_0\mathbf{O}) + O_p(n^{-\frac{1}{2}}) \\
=& \sqrt{n}\mathrm{tr}(\mathbf{B}^T\mathbf{\Gamma}_{0\perp}^T(\mathbf{G}_n - \mathbf{G})\mathbf{\Gamma}_0)(1+O_p(n^{-\frac{1}{2}})),
\end{aligned}$$

where $\mathbf{\Lambda} = \text{diag}\{\mathbf{\Lambda}_1, \mathbf{\Lambda}_2\}$ is the diagonal eigenvalue matrix of $\mathbf{G}$ with the first $d \times d$ sub-matrix $\mathbf{\Lambda}_1$. By using the definition of $\mathbf{Z}$ in (3.12), we get

$$\begin{aligned}
\text{tr}(\mathbf{Z}^T\mathbf{Z}\mathbf{\Gamma}_*^T\mathbf{G}_n\mathbf{\Gamma}_*) - \text{tr}(\mathbf{Z}^T\mathbf{G}_n\mathbf{Z}) =& \text{tr}(\mathbf{Z}^T\mathbf{Z}\mathbf{O}\mathbf{\Gamma}_0^T\mathbf{G}\mathbf{\Gamma}_0\mathbf{O}) - \text{tr}(\mathbf{Z}^T\mathbf{G}\mathbf{Z}) + O_p(n^{-1/2}) \\
=& \text{tr}(\mathbf{Z}^T\mathbf{Z}\mathbf{\Lambda}_1) - \text{tr}(\mathbf{Z}^T\mathbf{G}\mathbf{Z}) + O_p(n^{-1/2}) \\
=& \text{tr}(\mathbf{A}^T\mathbf{A}\mathbf{\Lambda}_1) + \text{tr}(\mathbf{B}^T\mathbf{B}\mathbf{\Lambda}_1) - \text{tr}(\mathbf{B}\mathbf{B}^T\mathbf{\Lambda}_2) \\
& - \text{tr}(\mathbf{A}\mathbf{A}^T\mathbf{\Lambda}_1) + o_p(1) \\
\geq& (\lambda_d - \lambda_{d+1})||\mathbf{B}||_s^2,
\end{aligned}$$

where we use the fact $\text{tr}(\mathbf{A}^T\mathbf{A}\mathbf{\Lambda}_1) - \text{tr}(\mathbf{A}\mathbf{A}^T\mathbf{\Lambda}_1) = 0$ because $\mathbf{A}$ is skew-symmetric. Here the last inequality follows from basic properties of trace operator for semi-positive definite matrix. As a consequence, by the Cauchy-Schwarz inequality for trace operator, the third term in Δ_1 is uniformly bounded by $||\mathbf{B}||_s \times ||\sqrt{n}(\mathbf{G}_n - \mathbf{G})\mathbf{\Gamma}_0||_s$. Therefore, as long as the constant C is sufficiently large, the first two terms in Δ_1 will always dominate the third term and Δ_2 with arbitrarily large probabilities. This implies the inequality (3.13), and the proof is completed. □

Proof of Theorem 2: (i) To prove this part, we need represent (3.7) as vector forms. Define

$$\begin{aligned}
\boldsymbol{t} &= (\boldsymbol{t}_1^T, \ldots, \boldsymbol{t}_d^T)^T, \\
h_l(\boldsymbol{t}) &= \boldsymbol{t}^T\mathbf{C}_l\boldsymbol{t}, \quad l = 1, \ldots, d, \\
h_{kl}(\boldsymbol{t}) &= \boldsymbol{t}^T\mathbf{C}_{kl}\boldsymbol{t}, \ (k,l) \in \mathcal{J}, \\
\mathcal{J} &= \{(k,l)|k,l = 1, \ldots, d, \ k < l\},
\end{aligned}$$

where $\boldsymbol{t}_i$ denotes the i-th column vector of $\mathbf{V}$, $\mathbf{C}_l$'s are $pd \times pd$ block-diagonal matrices, $\mathbf{C}_{kl}$'s $pd \times pd$ block matrices, $\mathbf{C}_l$ and $\mathbf{C}_{kl}$ contain $\mathbf{N}_n$ in the l th diagonal block and in the (k,l) as well as (l,k) blocks, respectively. The $pd \times pd$ symmetric matrices $\mathbf{C}_{kl}$ are defined for all the pairs of different indices belonging to $\mathcal{J}$, given by the $d(d-1)/2$ combinations of the indices $1, \ldots, d$.

By these notation, we have

$$Q(\mathbf{\Gamma}) := Q^*(\boldsymbol{t}) = -\boldsymbol{t}^T\mathbf{A}\boldsymbol{t} + \sum_{i=1}^{p} \theta_i||\boldsymbol{v}_i||_2,$$

where $\mathbf{A}$ is a $pd \times pd$ block-diagonal matrix with all diagonal blocks $\mathbf{M}_n$. Of course, in the above equation each $\boldsymbol{v}_i$ is regarded as a function of $\boldsymbol{t}$.

By using the equality representation of the compact Stiefel manifolds $St(p,d)$, (3.7) is equivalent to

$$\begin{aligned} &\min_{\boldsymbol{t}} -\left\{ \boldsymbol{t}^T \mathbf{A} \boldsymbol{t} + \sum_{i=1}^{p} \theta_i ||\boldsymbol{v}_i||_2 \right\}, \qquad (3.14) \\ &\text{subject to} \quad h_l(\boldsymbol{t}) = 1,\ l = 1 \in [1,d], \quad \text{and} \quad h_{kl}(\boldsymbol{t}) = 0,\ (k,l) \in \mathcal{J}, \end{aligned}$$

As a consequence, this enables us to apply an improved global lagrange multiplier rule proposed by Rapcsák (1997).

We start by supposing that $\tilde{\boldsymbol{v}}_j \neq 0$ for all j. According to Theorem 15.2.1 in Rapcsák (1997) (or Theorem 3.1 in Rapcsák 2002), a necessary condition that $\tilde{\boldsymbol{t}}_n$ ($\tilde{\mathbf{V}}_n$) is a local minimum of (3.14) (Eq. (3.7)) is that, the geodesic gradient vector of the improved Lagrangian function of (3.14) evaluated at $\tilde{\boldsymbol{t}}_n$ equals to zero. That is,

$$\begin{aligned} \frac{\partial^g Q^*(\boldsymbol{t})}{\partial \boldsymbol{t}}\Big|_{\boldsymbol{t}=\tilde{\boldsymbol{t}}_n} &\equiv \left[\frac{\partial Q^*(\boldsymbol{t})}{\partial \boldsymbol{t}} - \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U} \frac{\partial Q^*(\boldsymbol{t})}{\partial \boldsymbol{t}} \right] \Big|_{\boldsymbol{t}=\tilde{\boldsymbol{t}}_n} \qquad (3.15) \\ &:= \frac{\partial^g f(\mathbf{V}_n)}{\partial \boldsymbol{t}}\Big|_{\boldsymbol{t}=\tilde{\boldsymbol{t}}_n} + \frac{\partial^g \rho(\mathbf{V}_n)}{\partial \boldsymbol{t}}\Big|_{\boldsymbol{t}=\tilde{\boldsymbol{t}}_n} = \mathbf{0}, \end{aligned}$$

where

$$\mathbf{U} = (\mathbf{C}_1\boldsymbol{t}, \ldots, \mathbf{C}_d\boldsymbol{t}, \mathbf{C}_{12}\boldsymbol{t}, \mathbf{C}_{13}\boldsymbol{t}, \ldots, \mathbf{C}_{d-1d}\boldsymbol{t}),$$

is a $(pd \times [d(d+1)/2])$-dimensional matrix, and $\partial^g f(\mathbf{V}_n)/\partial \boldsymbol{t}$ and $\partial^g \rho(\mathbf{V}_n)/\partial \boldsymbol{t}$ are defined in a similar form of $\partial^g Q^*(\boldsymbol{t})/\partial \boldsymbol{t}$ by replacing Q^* with f and ρ, respectively. By Theorem 1 and noting that $\partial f(\mathbf{V}_n)/\partial \boldsymbol{t}$ is linear in $\boldsymbol{t}$,

$$\frac{\partial^g f(\mathbf{V}_n)}{\partial \boldsymbol{t}}\Big|_{\boldsymbol{t}=\tilde{\boldsymbol{t}}_n} = \frac{\partial^g f(\mathbf{V}_n)}{\partial \boldsymbol{t}}\Big|_{\boldsymbol{t}=\widehat{\boldsymbol{t}}_n} + O_p(n^{-1/2}),$$

where $\widehat{\boldsymbol{t}}_n$ is the vector form of $\widehat{\mathbf{V}}_n$. Using Theorem 3.1 of Rapcsák (2002) we have $\partial^g f(\mathbf{V}_n)/\partial \boldsymbol{t}|_{\boldsymbol{t}=\widehat{\boldsymbol{t}}_n} = \mathbf{0}$, which yields that $\partial^g f(\mathbf{V}_n)/\partial \boldsymbol{t}|_{\boldsymbol{t}=\tilde{\boldsymbol{t}}_n} = O_p(n^{-1/2})$ and as a consequence

$$\partial^g \rho(\mathbf{V}_n)/\partial \boldsymbol{t}|_{\boldsymbol{t}=\tilde{\boldsymbol{t}}_n} = O_p(n^{-1/2}).$$

On the other hand,

$$\frac{\partial^g \rho(\mathbf{V}_n)}{\partial \boldsymbol{t}}\Big|_{\boldsymbol{t}=\tilde{\boldsymbol{t}}_n} = \left[\mathbf{I}_{pd} - \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}\right] \tilde{\boldsymbol{\theta}} \equiv \mathbf{H}\tilde{\boldsymbol{\theta}},$$

where

$$\tilde{\boldsymbol{\theta}} = \left(\frac{\theta_1 \tilde{t}_{n11}}{||\tilde{\boldsymbol{v}}_{n1}||_2}, \ldots, \frac{\theta_p \tilde{t}_{n1p}}{||\tilde{\boldsymbol{v}}_{np}||_2}, \ldots, \frac{\theta_1 \tilde{t}_{nd1}}{||\tilde{\boldsymbol{v}}_{n1}||_2}, \ldots, \frac{\theta_p \tilde{t}_{ndp}}{||\tilde{\boldsymbol{v}}_{np}||_2} \right)^T.$$

By using the fact that $\mathbf{U}$ has full column rank and $\mathbf{HU} = \mathbf{0}$, we know $\tilde{\boldsymbol{\theta}}$ can be expressed through a linear combination of the columns of $\mathbf{U}$ in probability, i.e.,

$$\begin{aligned} \tilde{\boldsymbol{\theta}} =& (\kappa_1 \mathbf{C}_1 + \ldots + \kappa_d \mathbf{C}_d + \kappa_{12} \mathbf{C}_{12} + \kappa_{13} \mathbf{C}_{13} \\ & + \ldots + \kappa_{d-1d} \mathbf{C}_{d-1d}) \frac{||\tilde{\boldsymbol{\theta}}||_2}{||\tilde{\boldsymbol{t}}_n||_2} \tilde{\boldsymbol{t}}_n + O_p(n^{-1/2}), \end{aligned}$$

where $\kappa_1, \ldots, \kappa_{d-1d}$ are a sequence of constants satisfy they are not all the zeros. Define a sequence of pd-dimensional vectors $\boldsymbol{z}_{ij}$'s,

$$\boldsymbol{z}_{ij} = (\mathbf{0}^T, \ldots, \tilde{\boldsymbol{t}}_{ni}^T, \ldots, \mathbf{0}^T, \ldots, \tilde{\boldsymbol{t}}_{ni}^T, \ldots, \mathbf{0}^T)^T,$$

for $j \geq i$, say, its $[(i-1)p+1]$-th to the $[(i-1)p+p]$-th elements and $[(j-1)p+1]$-th to the $[(j-1)p+p]$-th elements are both $\tilde{\boldsymbol{t}}_{ni}$. It is straightforward to see

$$\begin{aligned} \kappa_0 \kappa_i &= \boldsymbol{z}_{ii}^T \tilde{\boldsymbol{\theta}} + O_p(n^{-1/2}), \\ \kappa_0 (\kappa_i + \kappa_{ij}) &= \boldsymbol{z}_{ij}^T \tilde{\boldsymbol{\theta}} + O_p(n^{-1/2}), \quad \text{for} \quad j > i, \end{aligned} \tag{3.16}$$

where we denote $\kappa_0 = ||\tilde{\boldsymbol{\theta}}||_2 / ||\tilde{\boldsymbol{t}}_n||_2$. By Theorem 1, $\tilde{\boldsymbol{v}}_{nj} = O_p(n^{-1/2})$ for $j > q$. Thus, by recalling the theorem's condition on a_n and b_n, it can be easily verified that (3.16) leads to

$$\begin{aligned} \kappa_i + \kappa_{ij} &= \kappa_0^{-1} (\boldsymbol{z}_{ij}^T \tilde{\boldsymbol{\theta}} + O_p(n^{-1/2})) \\ &\leq O_p(b_n^{-1}) \cdot O_p(a_n + b_n n^{-1/2} + n^{-1/2}) = o_p(1). \end{aligned}$$

Similarly, $\kappa_i = o_p(1)$. Consequently, we can conclude all the κ_i and κ_{ij} equal to zero in probability which yields contradiction. As a result, with probability tending to 1 (w.p.1), (3.15) cannot hold, which implies there exists $j > q$ so that

$$\Pr(\tilde{\boldsymbol{v}}_{nj} = 0) \to 1.$$

Without loss of generality, we assume $\Pr(\tilde{\boldsymbol{v}}_{np} = 0) \to 1$. Let $\mathbf{M}_{n1}$ and $\mathbf{N}_{n1}$ be the first $(p-1) \times (p-1)$ sub-matrices of $\mathbf{M}_n$ and $\mathbf{N}_n$ respectively, and $\tilde{\mathbf{V}}_{n1}$ be the first $p-1$

rows of $\tilde{\mathbf{V}}_n$. As stated before, $\tilde{\mathbf{V}}_n$ is a local minimum of the objective function

$$Q(\mathbf{V};\mathbf{M}_n) = -\text{tr}(\mathbf{V}^T\mathbf{M}_n\mathbf{V}) + \sum_{i=1}^{p}\theta_i||\boldsymbol{v}_i||_2, \text{ subject to } \mathbf{V}^T\mathbf{N}_n\mathbf{V} = \mathbf{I}_d.$$

We will show that w.p.1 $\tilde{\mathbf{V}}_{n1}$ is also a local minimum of the objective function

$$Q(\mathbf{V}_1;\mathbf{M}_{n1}) = -\text{tr}({\mathbf{V}_1}^T\mathbf{M}_{n1}\mathbf{V}_1) + \sum_{i=1}^{p-1}\theta_i||\boldsymbol{v}_i||_2, \quad \text{subject to } {\mathbf{V}_1}^T\mathbf{N}_{n1}\mathbf{V}_1 = \mathbf{I}_d, \tag{3.17}$$

w.p.1. Denote the set $\mathcal{A}_1 = \{\mathbf{V}_1|\ ||\mathbf{V}_1 - \tilde{\mathbf{V}}_{n1}||_s < \delta;\ \mathbf{V}_1^T\mathbf{N}_{n1}\mathbf{V}_1 = \mathbf{I}_d\}$. For any $\mathbf{A}_1 \in \mathcal{A}_1$, denote $\mathbf{A} = (\mathbf{A}_1^T, \mathbf{0}^T)^T$. It is clear that $\mathbf{A}^T\mathbf{N}_n\mathbf{A} = \mathbf{I}_d$. Given δ small enough, we will have $Q(\mathbf{A};\mathbf{M}_n) \geq Q(\tilde{\mathbf{V}}_n;\mathbf{M}_n)$ since $\tilde{\mathbf{V}}_n$ is the local minimum. Note that $Q(\mathbf{A};\mathbf{M}_n) = Q(\mathbf{A}_1;\mathbf{M}_{n1})$ and $Q(\tilde{\mathbf{V}};\mathbf{M}_n) = Q(\tilde{\mathbf{V}}_{n1};\mathbf{M}_{n1})$ w.p.1. Consequently, we have

$$Q(\mathbf{A}_1;\mathbf{M}_{n1}) \geq Q(\tilde{\mathbf{V}}_{n1};\mathbf{M}_{n1}),\ \text{w.p.1},$$

for all $\mathbf{A}_1 \in \mathcal{A}$ provided that δ is sufficiently small. Hence, we can conclude that $\tilde{\mathbf{V}}_{n1}$ is also a local minimum of the objective function $Q(\mathbf{V}_1;\mathbf{M}_{n1})$ w.p.1.

Rewriting (3.17) as a similar form to (3.14) and following the same arguments above in proving $\Pr(\tilde{\boldsymbol{v}}_{np} = 0) \to 1$, we can show that there exists $q < j < p$ so that $\Pr(\tilde{\boldsymbol{v}}_{nj} = 0) \to 1$. The remaining proofs can be completed by deduction.

(ii) For convenience purposes, first decompose the matrix $\mathbf{M}_n$ and $\mathbf{N}_n$ into the following block form:

$$\mathbf{M}_n = \begin{bmatrix} \mathbf{M}_{n(q)} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{n(p-q)} \end{bmatrix} \quad \mathbf{N}_n = \begin{bmatrix} \mathbf{N}_{n(q)} & \mathbf{N}_{12} \\ \mathbf{N}_{21} & \mathbf{N}_{n(p-q)} \end{bmatrix},$$

where $\mathbf{M}_{n(q)}$ and $\mathbf{N}_{n(q)}$ are the first $q \times q$ sub-matrices. It then follows

$$f(\mathbf{V};\mathbf{M}_n) = -\text{tr}(\mathbf{V}_{(q)}^T\mathbf{M}_{n(q)}\mathbf{V}_{(q)}) - \text{tr}(\mathbf{V}_{(p-q)}^T\mathbf{M}_{n(p-q)}\mathbf{V}_{(p-q)}).$$

Next we will show $\tilde{\mathbf{V}}_{n(q)} = \widehat{\mathbf{V}}_{n(O)}(1 + o_p(n^{-1/2}))$. Similar to the proof of Theorem 1, since $\tilde{\mathbf{V}}_{n(p-q)} = 0$ w.p.1, it suffices to show, for any arbitrarily small $\varepsilon > 0$, there exits

a sufficiently large constant C, such that

$$\liminf_n \Pr\Bigg(\inf_{\mathbf{Z}\in T_{\widehat{\mathbf{\Gamma}}_{n(O)}}} \inf_{(q,d):||\mathbf{B}||_s=C} Q(R(\widehat{\mathbf{\Gamma}}_{n(O)}+a_n\mathbf{Z});\mathbf{G}_{n(q)},\mathbf{N}_{n(q)}) \\ > Q(\widehat{\mathbf{\Gamma}}_{n(O)};\mathbf{G}_{n(q)},\mathbf{N}_{n(q)})\Bigg) > 1-\varepsilon, \tag{3.18}$$

where

$$\widehat{\mathbf{\Gamma}}_{n(O)} = \arg\min_{\mathbf{\Gamma}\in\mathbb{R}^{q\times d}} -\mathrm{tr}(\mathbf{\Gamma}^T\mathbf{G}_{n(q)}\mathbf{\Gamma}),\ \text{subject to}\ \mathbf{\Gamma}^T\mathbf{\Gamma}=\mathbf{I}_d,$$

and $\mathbf{G}_{n(q)} = \mathbf{N}_{n(q)}^{-1/2}\mathbf{M}_{n(q)}\mathbf{N}_{n(q)}^{-1/2}$. Note that

$$\begin{aligned} &a_n^{-2}\Big\{Q(R(\widehat{\mathbf{\Gamma}}_{n(O)}+a_n\mathbf{Z});\mathbf{G}_{n(q)},\mathbf{N}_{n(q)}) - Q(\widehat{\mathbf{\Gamma}}_{n(O)};\mathbf{G}_{n(q)},\mathbf{N}_{n(q)})\Big\} \\ &\quad\geq \Big[-\mathrm{tr}(\mathbf{Z}^T\mathbf{G}_{n(q)}\mathbf{Z}) - 2a_n^{-1}\mathrm{tr}(\mathbf{Z}^T\mathbf{G}_{n(q)}\widehat{\mathbf{\Gamma}}_{n(O)}) + \mathrm{tr}(\mathbf{Z}^T\mathbf{Z}\widehat{\mathbf{\Gamma}}_{n(O)}^T\mathbf{G}_{n(q)}\widehat{\mathbf{\Gamma}}_{n(O)})\Big](1+o_p(1)) \\ &\qquad - q||\mathbf{1}_j\mathbf{N}_{n(q)}^{-\frac{1}{2}}(\mathbf{Z}-(1/2)a_n\widehat{\mathbf{\Gamma}}_{n(O)}\mathbf{Z}^T\mathbf{Z})||_2, \end{aligned}$$

where $2a_n^{-1}\mathrm{tr}(\mathbf{Z}^T\mathbf{G}_{n(q)}\widehat{\mathbf{\Gamma}}_{n(O)}) = 0$ by using Lemma 2, and

$$-\mathrm{tr}(\mathbf{Z}^T\mathbf{G}_{n(q)}\mathbf{Z}) + \mathrm{tr}(\mathbf{Z}^T\mathbf{Z}\widehat{\mathbf{\Gamma}}_{n(O)}^T\mathbf{G}_{n(q)}\widehat{\mathbf{\Gamma}}_{n(O)}) > 0.$$

Using the similar arguments in the proof of Theorem 1, we can show (3.18) holds. This implies that $\sqrt{n}\tilde{\mathbf{\Gamma}}_{n(q)}$ is asymptotically equivalent to $\sqrt{n}\widehat{\mathbf{\Gamma}}_{n(O)}$ where

$$\tilde{\mathbf{\Gamma}}_{n(q)} = \arg\min_{\mathbf{\Gamma}\in\mathbb{R}^{q\times d}} Q(\mathbf{\Gamma};\mathbf{G}_{n(q)},\mathbf{N}_{n(q)}),\ \text{subject to}\ \mathbf{\Gamma}^T\mathbf{\Gamma}=\mathbf{I}_d,$$

and thus it follows that $\sqrt{n}D(\mathbf{N}_{n(q)}^{1/2}\tilde{\mathbf{V}}_{n(q)},\mathbf{N}_{n(q)}^{1/2}\widehat{\mathbf{V}}_{n(O)}) = o_p(1)$ which completes the proof. □

Proof of Proposition 5: To illustrate the idea, we elaborate on verifying the condition (3.10) for DR. In this case, by Eq. (5) in Li and Wang (2007), $\mathbf{M}_n$ can be reexpressed as

$$\begin{aligned} \mathbf{M}_n =& 2\Big\{\mathbf{\Sigma}_n^{1/2}\widehat{E}[\widehat{\mathrm{Var}}(\mathbf{z}|\tilde{y}) - \mathbf{I}_p]^2\mathbf{\Sigma}_n^{1/2} + \mathbf{\Sigma}_n^{1/2}\widehat{E}\big[(\widehat{\mathrm{Var}}(\mathbf{z}|\tilde{y}) - \mathbf{I}_p)\widehat{E}(\mathbf{z}|\tilde{y})\widehat{E}(\mathbf{z}^T|\tilde{y})\big]\mathbf{\Sigma}_n^{1/2} \\ &+ \mathbf{\Sigma}_n^{1/2}\widehat{E}\big[\widehat{E}(\mathbf{z}|\tilde{y})\widehat{E}(\mathbf{z}^T|\tilde{y})(\widehat{\mathrm{Var}}(\mathbf{z}|\tilde{y}) - \mathbf{I}_p)\big]\mathbf{\Sigma}_n^{1/2} + \mathbf{\Sigma}_n^{1/2}\widehat{E}\big[\widehat{E}(\mathbf{z}|\tilde{y})\widehat{E}(\mathbf{z}^T|\tilde{y})\big]^2\mathbf{\Sigma}_n^{1/2} \\ &+ \mathbf{\Sigma}_n^{1/2}\widehat{E}^2\big[\widehat{E}(\mathbf{z}|\tilde{y})\widehat{E}(\mathbf{z}^T|\tilde{y})\big]\mathbf{\Sigma}_n^{1/2} + \mathbf{\Sigma}_n^{1/2}\widehat{E}\big[\widehat{E}(\mathbf{z}^T|\tilde{y})\widehat{E}(\mathbf{z}|\tilde{y})\big]\widehat{E}\big[\widehat{E}(\mathbf{z}|\tilde{y})\widehat{E}(\mathbf{z}^T|\tilde{y})\big]\mathbf{\Sigma}_n^{1/2}\Big\} \\ :=& 2(\mathbf{M}_{n1}+\ldots,+\mathbf{M}_{n6}). \end{aligned}$$

Here $\tilde{y}$ is the discretized y over a collection of slices, $\widehat{\mathrm{Var}}(\mathbf{z}|\tilde{y})$ denotes the sample covariance matrix of $\mathbf{z}$ within a slice, $\widehat{E}(\cdot)$ denotes the weighted average across slices. Next we will show $\mathbf{M}_{n(O)i} = \mathbf{M}_{n(q)i} + O_p(n^{-1})$ for $i = 1, \ldots, 6$.

Now we first deal with $\mathbf{M}_{n1}$. Rewrite it as

$$\mathbf{M}_{n1} = \widehat{E}\Big\{ \big[\widehat{\mathrm{Var}}(\mathbf{x}|\tilde{y}) - \boldsymbol{\Sigma}_n\big]\boldsymbol{\Sigma}_n^{-1}\big[\widehat{\mathrm{Var}}(\mathbf{x}|\tilde{y}) - \boldsymbol{\Sigma}_n\big]\Big\}.$$

We assume that the collection of slices is fixed; that is, it does not vary with n. This implies that the sample conditional moments such as $\widehat{\mathrm{Var}}(\mathbf{x}|\tilde{y})$ are $\sqrt{n}$-consistent estimates of their population-level counterparts, such as $\mathrm{Var}(\mathbf{x}|\tilde{y})$. Let $\boldsymbol{\Omega}$ be the matrix consisting of the first q columns of the matrix $\mathbf{I}_p$. Then, by definition,

$$\begin{aligned}
\mathbf{M}_{n(O)1} &= \boldsymbol{\Omega}^T\widehat{E}\Big\{ \big[\widehat{\mathrm{Var}}(\mathbf{x}|\tilde{y}) - \boldsymbol{\Sigma}_n\big]\boldsymbol{\Omega}\big(\boldsymbol{\Omega}^T\boldsymbol{\Sigma}_n\boldsymbol{\Omega}\big)^{-1}\boldsymbol{\Omega}^T\big[\widehat{\mathrm{Var}}(\mathbf{x}|\tilde{y}) - \boldsymbol{\Sigma}_n\big]\Big\}\boldsymbol{\Omega},\\
\mathbf{M}_{n(q)1} &= \boldsymbol{\Omega}^T\widehat{E}\Big\{ \big[\widehat{\mathrm{Var}}(\mathbf{x}|\tilde{y}) - \boldsymbol{\Sigma}_n\big]\boldsymbol{\Sigma}_n^{-1}\big[\widehat{\mathrm{Var}}(\mathbf{x}|\tilde{y}) - \boldsymbol{\Sigma}_n\big]\Big\}\boldsymbol{\Omega}.
\end{aligned}$$

Let $\mathbf{P}_{\boldsymbol{\Omega}}(\boldsymbol{\Sigma}_n) = \boldsymbol{\Omega}\big(\boldsymbol{\Omega}^T\boldsymbol{\Sigma}_n\boldsymbol{\Omega}\big)^{-1}\boldsymbol{\Omega}^T\boldsymbol{\Sigma}_n$ and let $\mathbf{Q}_{\boldsymbol{\Omega}}(\boldsymbol{\Sigma}_n) = \mathbf{I}_p - \mathbf{P}_{\boldsymbol{\Omega}}(\boldsymbol{\Sigma}_n)$. Then

$$\begin{aligned}
\mathbf{M}_{n(q)1} =& \boldsymbol{\Omega}^T\widehat{E}\Big\{ \big[\widehat{\mathrm{Var}}(\mathbf{x}|\tilde{y}) - \boldsymbol{\Sigma}_n\big]\big[\mathbf{P}_{\boldsymbol{\Omega}}(\boldsymbol{\Sigma}_n) + \mathbf{Q}_{\boldsymbol{\Omega}}(\boldsymbol{\Sigma}_n)\big]\boldsymbol{\Sigma}_n^{-1}\\
&\big[\mathbf{P}_{\boldsymbol{\Omega}}(\boldsymbol{\Sigma}_n) + \mathbf{Q}_{\boldsymbol{\Omega}}(\boldsymbol{\Sigma}_n)\big]^T\big[\widehat{\mathrm{Var}}(\mathbf{x}|\tilde{y}) - \boldsymbol{\Sigma}_n\big]\Big\}\boldsymbol{\Omega} := \widehat{E}(\mathbf{M}_{1I} + \mathbf{M}_{1II} + \mathbf{M}_{1III} + \mathbf{M}_{1IV}),
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{M}_{1I} &= \boldsymbol{\Omega}^T\big[\widehat{\mathrm{Var}}(\mathbf{x}|\tilde{y}) - \boldsymbol{\Sigma}_n\big]\mathbf{P}_{\boldsymbol{\Omega}}(\boldsymbol{\Sigma}_n)\boldsymbol{\Sigma}_n^{-1}\mathbf{P}_{\boldsymbol{\Omega}}^T(\boldsymbol{\Sigma}_n)\big[\widehat{\mathrm{Var}}(\mathbf{x}|\tilde{y}) - \boldsymbol{\Sigma}_n\big]\boldsymbol{\Omega},\\
\mathbf{M}_{1II} &= \boldsymbol{\Omega}^T\big[\widehat{\mathrm{Var}}(\mathbf{x}|\tilde{y}) - \boldsymbol{\Sigma}_n\big]\mathbf{Q}_{\boldsymbol{\Omega}}(\boldsymbol{\Sigma}_n)\boldsymbol{\Sigma}_n^{-1}\mathbf{P}_{\boldsymbol{\Omega}}^T(\boldsymbol{\Sigma}_n)\big[\widehat{\mathrm{Var}}(\mathbf{x}|\tilde{y}) - \boldsymbol{\Sigma}_n\big]\boldsymbol{\Omega},\\
\mathbf{M}_{1III} &= \boldsymbol{\Omega}^T\big[\widehat{\mathrm{Var}}(\mathbf{x}|\tilde{y}) - \boldsymbol{\Sigma}_n\big]\mathbf{P}_{\boldsymbol{\Omega}}(\boldsymbol{\Sigma}_n)\boldsymbol{\Sigma}_n^{-1}\mathbf{Q}_{\boldsymbol{\Omega}}^T(\boldsymbol{\Sigma}_n)\big[\widehat{\mathrm{Var}}(\mathbf{x}|\tilde{y}) - \boldsymbol{\Sigma}_n\big]\boldsymbol{\Omega},\\
\mathbf{M}_{1IV} &= \boldsymbol{\Omega}^T\big[\widehat{\mathrm{Var}}(\mathbf{x}|\tilde{y}) - \boldsymbol{\Sigma}_n\big]\mathbf{Q}_{\boldsymbol{\Omega}}(\boldsymbol{\Sigma}_n)\boldsymbol{\Sigma}_n^{-1}\mathbf{Q}_{\boldsymbol{\Omega}}^T(\boldsymbol{\Sigma}_n)\big[\widehat{\mathrm{Var}}(\mathbf{x}|\tilde{y}) - \boldsymbol{\Sigma}_n\big]\boldsymbol{\Omega}.
\end{aligned}$$

It can be easily seen that $\widehat{E}(\mathbf{M}_{1I})$ is exactly $\mathbf{M}_{n(O)1}$. We will show that $\mathbf{M}_{1II}$, $\mathbf{M}_{1III}$, and $\mathbf{M}_{1IV}$ are of the order $O_p(n^{-1})$. Note that

$$\begin{aligned}
&\mathbf{Q}_{\boldsymbol{\Omega}}^T(\boldsymbol{\Sigma}_n)\big[\widehat{\mathrm{Var}}(\mathbf{x}|\tilde{y}) - \boldsymbol{\Sigma}_n\big]\\
=&\big[\mathbf{Q}_{\boldsymbol{\Omega}}^T(\boldsymbol{\Sigma}) + O_p(n^{-1/2})\big]\big[\widehat{\mathrm{Var}}(\mathbf{x}|\tilde{y}) - \boldsymbol{\Sigma} + O_p(n^{-1/2})\big]\\
=&\mathbf{Q}_{\boldsymbol{\Omega}}^T(\boldsymbol{\Sigma})\big[\widehat{\mathrm{Var}}(\mathbf{x}|\tilde{y}) - \boldsymbol{\Sigma}\big] + O_p(n^{-1/2}).
\end{aligned}$$

By construction, $\mathcal{S}_{y|\mathbf{x}} \subseteq \text{span}(\mathbf{\Omega})$. Under certain conditions (Cook 1998a), we know $\text{span}\{\mathbf{\Sigma}^{-1}[\mathbf{\Sigma} - \text{Var}(\mathbf{x}|y)]\} \subseteq \mathcal{S}_{y|\mathbf{x}}$. Hence

$$\text{span}\{\mathbf{\Sigma}^{-1}[\mathbf{\Sigma} - \text{Var}(\mathbf{x}|y)]\} \subseteq \text{span}(\mathbf{\Omega}).$$

It then follows that

$$\mathbf{Q}_{\mathbf{\Omega}}(\mathbf{\Sigma})\mathbf{\Sigma}^{-1}\big[\text{Var}(\mathbf{x}|\tilde{y}) - \mathbf{\Sigma}\big] = \mathbf{\Sigma}^{-1}\mathbf{Q}_{\mathbf{\Omega}}^{T}(\mathbf{\Sigma})\big[\text{Var}(\mathbf{x}|\tilde{y}) - \mathbf{\Sigma}\big] = \mathbf{0}. \tag{3.19}$$

Thus, we have $\mathbf{M}_{1IV} = O_p(n^{-1/2}) \cdot O_p(n^{-1/2}) = O_p(n^{-1})$.

Substituting $\mathbf{P}_{\mathbf{\Omega}}(\mathbf{\Sigma}_n) = \mathbf{I}_p - \mathbf{Q}_{\mathbf{\Omega}}(\mathbf{\Sigma}_n)$ into $\mathbf{M}_{1II}$ and using $\mathbf{Q}_{\mathbf{\Omega}}(\mathbf{\Sigma}_n)$'s idempotency, we have

$$\mathbf{M}_{1II} = \mathbf{\Omega}^T\big[\widehat{\text{Var}}(\mathbf{x}|\tilde{y}) - \mathbf{\Sigma}_n\big]\mathbf{Q}_{\mathbf{\Omega}}(\mathbf{\Sigma}_n)\mathbf{Q}_{\mathbf{\Omega}}(\mathbf{\Sigma}_n)\mathbf{\Sigma}_n^{-1}\big[\widehat{\text{Var}}(\mathbf{x}|\tilde{y}) - \mathbf{\Sigma}_n\big]\mathbf{\Omega} - \mathbf{M}_{1IV}.$$

By using (3.19) again, we know that $\mathbf{M}_{1II} = O_p(n^{-1})$. Similarly, $\mathbf{M}_{1III} = O_p(n^{-1})$. From these we deduce that $\mathbf{M}_{1II}$, $\mathbf{M}_{1III}$, $\mathbf{M}_{1IV}$ are all of order $O_p(n^{-1})$. Since $\widehat{E}(\mathbf{M}_{1II} + \mathbf{M}_{1III} + \mathbf{M}_{1IV})$ is the sum of finite number of terms each of the order $O_p(n^{-1})$, it is itself of this order. It follows that $\mathbf{M}_{n(O)1} = \mathbf{M}_{n(q)1} + O_p(n^{-1})$.

Next, let us deal with $\mathbf{M}_{n2}$. Similar to $\mathbf{M}_{n(q)1}$, $\mathbf{M}_{n(q)2}$ can be divided into four terms $\mathbf{M}_{n(q)2} = \mathbf{M}_{n(O)2} + \mathbf{M}_{2II} + \mathbf{M}_{2III} + \mathbf{M}_{2IV}$, where

$$\begin{aligned}
\mathbf{M}_{2II} &= \mathbf{\Omega}^T\big[\widehat{\text{Var}}(\mathbf{x}|\tilde{y}) - \mathbf{\Sigma}_n\big]\mathbf{Q}_{\mathbf{\Omega}}(\mathbf{\Sigma}_n)\mathbf{\Sigma}_n^{-1}\mathbf{P}_{\mathbf{\Omega}}^{T}(\mathbf{\Sigma}_n)\big\{[\widehat{E}(\mathbf{x}|\tilde{y}) - \widehat{E}(\mathbf{x})][\widehat{E}(\mathbf{x}^T|\tilde{y}) - \widehat{E}(\mathbf{x}^T)]\big\}\mathbf{\Omega},\\
\mathbf{M}_{2III} &= \mathbf{\Omega}^T\big[\widehat{\text{Var}}(\mathbf{x}|\tilde{y}) - \mathbf{\Sigma}_n\big]\mathbf{P}_{\mathbf{\Omega}}(\mathbf{\Sigma}_n)\mathbf{\Sigma}_n^{-1}\mathbf{Q}_{\mathbf{\Omega}}^{T}(\mathbf{\Sigma}_n)\big\{[\widehat{E}(\mathbf{x}|\tilde{y}) - \widehat{E}(\mathbf{x})][\widehat{E}(\mathbf{x}^T|\tilde{y}) - \widehat{E}(\mathbf{x}^T)]\big\}\mathbf{\Omega},\\
\mathbf{M}_{2IV} &= \mathbf{\Omega}^T\big[\widehat{\text{Var}}(\mathbf{x}|\tilde{y}) - \mathbf{\Sigma}_n\big]\mathbf{Q}_{\mathbf{\Omega}}(\mathbf{\Sigma}_n)\mathbf{\Sigma}_n^{-1}\mathbf{Q}_{\mathbf{\Omega}}^{T}(\mathbf{\Sigma}_n)\big\{[\widehat{E}(\mathbf{x}|\tilde{y}) - \widehat{E}(\mathbf{x})][\widehat{E}(\mathbf{x}^T|\tilde{y}) - \widehat{E}(\mathbf{x}^T)]\big\}\mathbf{\Omega}.
\end{aligned}$$

Under the linearity condition, we know $\text{span}\{[E(\mathbf{x}|\tilde{y}) - E(\mathbf{x})]\} \subseteq \mathcal{S}_{y|\mathbf{x}}$ (Cook 1998a). Hence

$$\text{span}\{[E(\mathbf{x}|\tilde{y}) - E(\mathbf{x})]\} \subseteq \text{span}(\mathbf{\Omega}).$$

It then follows that

$$\mathbf{Q}_{\mathbf{\Omega}}(\mathbf{\Sigma})\mathbf{\Sigma}^{-1}[E(\mathbf{x}|\tilde{y}) - E(\mathbf{x})] = \mathbf{\Sigma}^{-1}\mathbf{Q}_{\mathbf{\Omega}}^{T}(\mathbf{\Sigma})[E(\mathbf{x}|\tilde{y}) - E(\mathbf{x})] = \mathbf{0}. \tag{3.20}$$

By using (A.9) and the similar arguments for $\mathbf{M}_{n(q)1}$, we can show that $\mathbf{M}_{2II}$, $\mathbf{M}_{2III}$, and $\mathbf{M}_{2IV}$ are all of order $O_p(n^{-1})$. Thus, we can conclude that $\mathbf{M}_{n(O)2} = \mathbf{M}_{n(q)2} + O_p(n^{-1})$.

By (3.19) and (3.20), $\mathbf{M}_{n(O)i} = \mathbf{M}_{n(q)i} + O_p(n^{-1})$ for $i = 3, \ldots, 6$, can be proved in a similar fashion to the foregoing proofs. We omit the details here for saving some space. It follows that for the DR method,

$$\mathbf{M}_{n(O)} = \mathbf{M}_{n(q)} + O_p(n^{-1}).$$

Thus, the condition (3.10) is satisfied as long as $(na_n)^{-1} = O_p(1)$.

Note that for SAVE, $\mathbf{M}_n$ takes the form of $\mathbf{M}_{n1}$ for DR. Thus, the condition (3.10) is also satisfied for SAVE. □

Chapter 4

Principal envelope model

Principal component analysis (PCA) is a popular data processing and dimension reduction technique. First introduced by Pearson (1901), PCA has a long history and is now widely used in many areas, including agriculture, ecology, genetics and economics. PCA seeks uncorrelated linear combinations of the original variables that capture maximal variance. Suppose we have n observations on p features $x_1, \cdots, x_p$. Let $\mathbf{x}^{(i)} = (x_1^{(i)}, \cdots, x_p^{(i)})$ denote the i-th observation, $i = 1, \ldots, n$, and $\mathbf{x} = (x_1, \cdots, x_p)^T$ be the vector variable. Let $\tilde{\mathbf{x}}^{(i)}$ denote the centered observation vectors, $i = 1, \ldots, n$, and $\mathbb{X}$ be the $n \times p$ centered data matrix with row $\tilde{\mathbf{x}}^{(i)}$ and rank $r \leq \min(n, p)$. As there is no response involved, this chapter is about unsupervised multivariate dimension methods.

Let $\hat{\mathbf{g}}_1, \hat{\mathbf{g}}_2, \cdots, \hat{\mathbf{g}}_r$ be the eigenvectors of the sample covariance matrix $\hat{\boldsymbol{\Sigma}} = \mathbb{X}^T\mathbb{X}/n$ corresponding to its non-zero eigenvalues. Without loss of generality, $\hat{\mathbf{g}}_1, \hat{\mathbf{g}}_2, \cdots, \hat{\mathbf{g}}_r$ are ordered in the order of descending eigenvalues. Let $\hat{\lambda}_1, \hat{\lambda}_2, \cdots, \hat{\lambda}_r$ be the non-zero eigenvalues with descending order. The principal component directions $\hat{\mathbf{g}}_k$, $k = 1, \ldots, r$, can also be obtained by maximizing $\boldsymbol{\alpha}_k^T(\mathbb{X}^T\mathbb{X})\boldsymbol{\alpha}_k$ successively subject to $\boldsymbol{\alpha}_k^T\boldsymbol{\alpha}_k = 1$ and $\boldsymbol{\alpha}_h^T\boldsymbol{\alpha}_k = 0$, $\forall h < k$. This demonstrates that PCA pursues the linear combinations of the original variables such that the derived variables capture maximal variance. The sample variance of the ith principal component (PC) equals $\hat{\lambda}_i$. There are many methods for selecting the number of principal components, depending on application specific requirements (Jolliffe, 2002).

Despite its popularity, PCA is not based on a probability model. Tipping and Bishop

(1999) introduced probabilistic principal component analysis (PPCA) in which the first few principal component directions can be obtained through maximum likelihood estimation. However, the assumption of an isotropic error in the PPCA model is quite limiting. By assuming a general error structure and incorporating the novel "envelope" idea of Cook et al. (2010), we establish principal envelope models that encompass PPCA and demonstrate the possibility that any subset of principal components could retain most of the sample's information.

We revisit PPCA in Section 4.1.1 to study the link between PPCA and principal envelope models. In Section 4.1.2 we describe the concept of an envelope and demonstrate the possibility that any subset of principal components could retain most of the sample's information. We build some intermediate models in Section 4.1.3. The log-likelihood function of one specific principal envelope model has the same form as probabilistic extreme components analysis (PXCA, Welling and et. al, 2003) if the dimension of the envelope is the same as the minimum dimension reduction subspace. However, the concepts and statistical meanings of these two approaches are very different. In Section 4.2.1, we describe a special penalty function that depends only on subspaces. Based on this penalty function, in Section 4.2.2, a method is proposed to select the features in principal envelope models. Results of simulation studies are presented in Section 4.1.5 and Section 4.2.4. An illustrative data analysis is given in Section 4.3. Concluding remarks about the proposed methods can be found in Section 5. Technical details are given in the Appendix.

4.1 Principal envelope model

4.1.1 Probabilistic principal component analysis revisited

Tipping and Bishop (1999) proposed a probabilistic principal component model as follows:

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\beta}\boldsymbol{\nu} + \sigma\boldsymbol{\epsilon}, \tag{4.1}$$

where $\boldsymbol{\mu}$ permits $\mathbf{x}$ to have non-zero mean and the $p \times d$ matrix $\boldsymbol{\beta}$ relates the observable variable $\mathbf{x}$ and the latent variable $\boldsymbol{\nu}$, which is assumed to be normally distributed with mean 0 and identity covariance matrix. The error $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_p)$ is assumed to be

independent of $\boldsymbol{\nu}$ and d is assumed to be known. The parameter $\boldsymbol{\beta}$ is not identified since $\boldsymbol{\beta}\boldsymbol{\nu} = (\boldsymbol{\beta}\mathbf{O})(\mathbf{O}^T\boldsymbol{\nu})$ for any orthogonal matrix $\mathbf{O}$, resulting in an equivalent model. However, the subspace $\mathcal{B} = \text{span}(\boldsymbol{\beta})$ is identified and estimable. Tipping and Bishop showed that the maximum likelihood estimator of $\mathcal{B}$ is the span of the first d eigenvectors of $\hat{\boldsymbol{\Sigma}}$. A Grassmann manifold, which is defined as the set of all d-dimensional subspaces in $\mathbb{R}^p$, is the natural parameter space for the $\mathcal{B}$ parametrization. For more background on Grassmann manifold optimization, see Edelman, Arias and Smith (1998).

We reformulate model (4.1) as

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\delta}\boldsymbol{\nu} + \sigma\boldsymbol{\epsilon}, \tag{4.2}$$

where $\boldsymbol{\Gamma}$ is a $p \times d$ semi-orthogonal matrix ($\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \mathbf{I}_d$), $\boldsymbol{\delta}$ is a full rank $d \times d$ coordinate matrix, $\boldsymbol{\nu}$ and $\boldsymbol{\epsilon}$ are defined previously. Let $\mathcal{S}_{\boldsymbol{\Gamma}}$ denote the column space of $\boldsymbol{\Gamma}$. The population covariance variance of var($\mathbf{x}$) is

$$\boldsymbol{\Gamma}\boldsymbol{\delta}\boldsymbol{\delta}^T\boldsymbol{\Gamma}^T + \sigma^2\mathbf{I}_p = \boldsymbol{\Gamma}(\boldsymbol{\delta}\boldsymbol{\delta}^T + \sigma^2\mathbf{I}_d)\boldsymbol{\Gamma}^T + \sigma^2\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T = \boldsymbol{\Gamma}\mathbf{V}\boldsymbol{\Gamma}^T + \sigma^2\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T$$

where $\mathbf{V} = \boldsymbol{\delta}\boldsymbol{\delta}^T + \sigma^2\mathbf{I}_d$. The full log-likelihood function $L_0(\mathcal{S}_{\boldsymbol{\Gamma}}, \mathbf{V}, \sigma^2)$ can be calculated straightforwardly:

$$\begin{aligned} & -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Gamma}\mathbf{V}\boldsymbol{\Gamma}^T + \sigma^2\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T| - \frac{1}{2}\sum_{i=1}^{n}\tilde{x}_i^T(\boldsymbol{\Gamma}\mathbf{V}\boldsymbol{\Gamma}^T + \sigma^2\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T)^{-1}\tilde{x}_i \\ = \quad & -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{V}| - \frac{n}{2}\text{trace}(\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}\mathbf{V}^{-1}) - \frac{n}{2}(p-d)\log(\sigma^2) - \frac{n}{2\sigma^2}\text{trace}(\boldsymbol{\Gamma}_0^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}_0). \end{aligned}$$

In the above expression, $|A|$ denotes the determinant of the matrix A. In later sections, we will ignore the common constant $-(np/2)\log(2\pi)$ in the log-likelihood functions.

If we maximize over $\mathbf{V}$ and σ^2 separatively, we arrive at the same partially maximized likelihood function as Tipping and Bishop (1999). However, the parameters $\mathbf{V}$ and σ^2 are not in proper product spaces because the eigenvalues of $\mathbf{V}$ are bounded below by σ^2. Thus it seems inappropriate to maximize over $\mathbf{V}$ and σ^2 separatively. The result of Proposition 6 is the same as Tipping and Bishop (1999), but we present a totally different proof in the appendix.

Proposition 6 *The maximum likelihood estimator $\hat{\mathcal{S}}_{\boldsymbol{\Gamma}}$ in $L_0(\mathcal{S}_{\boldsymbol{\Gamma}}, \mathbf{V}, \sigma^2)$ is the subspace spanned by the first d principal component directions and can be obtained by maximizing* $\text{trace}(\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma})$ *subject to* $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \mathbf{I}_d$.

4.1.2 Motivation: general error structure

Instead of assuming an isotropic error structure, which is very limiting, we assume

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\beta}\boldsymbol{\nu} + \boldsymbol{\Phi}^{1/2}\boldsymbol{\epsilon}, \tag{4.3}$$

where $\boldsymbol{\Phi}$ is a general positive definite matrix. The latent variable $\boldsymbol{\nu}$ represents extrinsic variation in $\mathbf{x}$, while the error $\boldsymbol{\epsilon}$ represents intrinsic variation. Traditional PCA reduces dimensionality while keeping most of its total variation. Our goal is different and we reduce the dimension of $\mathbf{x}$ accounting for its extrinsic variation. Under this case, we can show that $\mathbf{x} \perp\!\!\!\perp \boldsymbol{\nu}|\boldsymbol{\beta}^T\boldsymbol{\Phi}^{-1}\mathbf{x}$ (Cook, 2007). Thus $R = \boldsymbol{\beta}^T\boldsymbol{\Phi}^{-1}\mathbf{x}$ is the reduction we would like to estimate because $\mathbf{x}$ contains no further information about $\boldsymbol{\nu}$ given $\boldsymbol{\beta}^T\boldsymbol{\Phi}^{-1}\mathbf{x}$. Let $\mathcal{T} = \text{span}(\boldsymbol{\Phi}^{-1}\boldsymbol{\beta})$. Since any full rank linear transformation $\mathbf{A}$ of $\mathbf{R}$ results in an equivalent reduction; $\mathbf{x} \perp\!\!\!\perp \boldsymbol{\nu}|\mathbf{R}$ if and only if $\mathbf{x} \perp\!\!\!\perp \boldsymbol{\nu}|\mathbf{AR}$, so it is sufficient to estimate the $\mathcal{T}$. Additionally, if $\mathcal{T}$ is minimal and if $\mathbf{x} \perp\!\!\!\perp \boldsymbol{\nu}|\mathbf{B}^T\mathbf{x}$, then $\mathcal{T} \subseteq \text{span}(\mathbf{B})$.

Under model (4.3), we see that $\mathbf{x}$ is normal with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma} = \boldsymbol{\Phi} + \boldsymbol{\beta}\boldsymbol{\beta}^T$. The maximum likelihood estimator of $\boldsymbol{\mu}$ is simply the sample mean of $\mathbf{x}$, however $\boldsymbol{\Phi}$ and $\boldsymbol{\beta}$ are confounded, thus $\mathcal{T}$ can not be estimated without assuming additional structure. The principal envelope idea is to estimate an upper bound on $\mathcal{T}$. By doing so, we don't lose any information on its extrinsic variation. Before we explain the concept of an envelope, we review the concept of reducing subspace.

Definition 2 *A subspace $\mathscr{R}$ is a reducing subspace of $\mathbf{M} \in \mathbb{R}^{p\times p}$ if $\mathbf{M}\mathscr{R} \subseteq \mathscr{R}$ and $\mathbf{M}\mathscr{R}^{\perp} \subseteq \mathscr{R}^{\perp}$ where $\mathscr{R}^{\perp}$ stands for the complement of $\mathscr{R}$ in the usual inner product.*

Definition 3 *(Cook, Li and Chiaromonte, 2010) Suppose that the symmetric matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$ and let the subspace $\mathcal{K} \subseteq \text{span}(\mathbf{M})$. The $\mathbf{M}$-envelope of $\mathcal{K}$, to be written as $\mathcal{E}_{\mathbf{M}}(\mathcal{K})$, is the intersection of all reducing subspaces of $\mathbf{M}$ that contain of $\mathcal{K}$.*

Let $\boldsymbol{\Gamma}$ be an orthonormal basis of $\mathcal{E}_{\boldsymbol{\Phi}}(\mathcal{T})$ and $\boldsymbol{\Gamma}_0$ be the orthogonal complement of $\boldsymbol{\Gamma}$ where $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times u}$, $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{p\times(p-u)}$ and $u \geq d$. By definition, we have $\mathcal{T} \subseteq \mathcal{E}_{\boldsymbol{\Phi}}(\mathcal{T})$ and $\boldsymbol{\Gamma}^T\mathbf{x} \perp\!\!\!\perp \boldsymbol{\Gamma}_0^T\mathbf{x}|\boldsymbol{\nu}$. Then model (4.3) can be re-written as

$$\begin{aligned} \mathbf{x} &= \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{\nu} + \boldsymbol{\Phi}^{1/2}\boldsymbol{\epsilon}, \\ \boldsymbol{\Phi} &= \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T, \end{aligned} \tag{4.4}$$

where $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$ and $\boldsymbol{\eta}$ is a $u \times d$ matrix with rank d. This model is referred as principal envelope model (PEM). We note that $\mathcal{S}_{\boldsymbol{\Gamma}}$ is the same as $\mathcal{E}_{\boldsymbol{\Phi}}(\mathcal{T})$. In the likelihood function, we prefer to use $\mathcal{S}_{\boldsymbol{\Gamma}}$ so that it is clear what parameter we are going to estimate. The estimate of $\mathcal{S}_{\boldsymbol{\Gamma}}$ provides an upper bound on the estimate of $\mathcal{T}$. The parameter d is not estimable under this model. When $u = d$, $\boldsymbol{\Omega} = \sigma^2\mathbf{I}_d$ and $\boldsymbol{\Omega}_0 = \sigma^2\mathbf{I}_{p-d}$, model (4.4) reduces to PPCA.

The population marginal covariance of $\mathbf{x}$ can be calculated straightforwardly: $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}(\boldsymbol{\Omega} + \boldsymbol{\eta}\boldsymbol{\eta}^T)\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$. Let $\boldsymbol{\Psi} = \boldsymbol{\Omega} + \boldsymbol{\eta}\boldsymbol{\eta}^T$. The parameter $\boldsymbol{\eta}$ is confounded with $\boldsymbol{\Omega}$ and can not be estimated here and in later sections, but Ψ can be estimated. After some algebra, we have the log-likelihood function

$$\begin{aligned}
& -\frac{n}{2}\log|\boldsymbol{\Gamma}\boldsymbol{\Psi}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T| - \frac{1}{2}\sum_{i=1}^n \tilde{x}_i^T(\boldsymbol{\Gamma}\boldsymbol{\Psi}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T)^{-1}\tilde{x}_i \\
= & -\frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{n}{2}\log|\boldsymbol{\Omega}_0| - \frac{1}{2}\sum_{i=1}^n \tilde{x}_i^T(\boldsymbol{\Gamma}\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma}^T)\tilde{x}_i - \frac{1}{2}\sum_{i=1}^n \tilde{x}_i^T(\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T)\tilde{x}_i \\
= & -\frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{n}{2}\mathrm{trace}(\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}\boldsymbol{\Psi}^{-1}) - \frac{n}{2}\log|\boldsymbol{\Omega}_0| - \frac{n}{2}\mathrm{trace}(\boldsymbol{\Gamma}_0^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}).
\end{aligned}$$

Maximizing over $\boldsymbol{\Psi}$ and $\boldsymbol{\Omega}_0$, we have the partially maximized log-likelihood function

$$\begin{aligned}
L_1(\mathcal{S}_{\boldsymbol{\Gamma}}) &= -\frac{n}{2}\log|\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}| - \frac{n}{2}\log|\boldsymbol{\Gamma}_0^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}_0| - \frac{n}{2}p \\
&= -\frac{n}{2}\log|\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}| - \frac{n}{2}\log|\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Gamma}| - \frac{n}{2}p - \frac{n}{2}\log|\hat{\boldsymbol{\Sigma}}|
\end{aligned}$$

The function $L_1(\mathcal{S}_{\boldsymbol{\Gamma}})$ requires $n > p$ as $\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\Gamma}$ must not be singular.

Proposition 7 *(i) We have* $L_1(\mathcal{S}_{\boldsymbol{\Gamma}}) \le -(np)/2 - (n/2)\log|\hat{\boldsymbol{\Sigma}}|$ *for all* $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = \mathbf{I}_u$.

(ii) Let $\mathcal{J}$ *be a subset of the index set* $\{1, \ldots, p\}$ *with* u *elements. Define* $\hat{\mathcal{S}}_{\boldsymbol{\Gamma}} = \mathrm{span}(\mathbf{g}_{\mathcal{J}_1}, \ldots, \mathbf{g}_{\mathcal{J}_u})$ *where* $\mathbf{g}_{\mathcal{J}_1}, \ldots, \mathbf{g}_{\mathcal{J}_u}$ *denotes any* u *principal component directions, then* $L_1(\hat{\mathcal{S}}_{\boldsymbol{\Gamma}}) = -(np)/2 - (n/2)\log|\hat{\boldsymbol{\Sigma}}|$.

From Proposition 7, we see that the span of any u principal component directions is the maximum likelihood estimator of $\mathcal{S}_{\boldsymbol{\Gamma}}$. In other words, any subset of principal component directions is equally supported by the likelihood function. It also tells us that we need extra information to tell which subset is useful.

4.1.3 Specific principal envelope models

Assuming that we can model $\mathbf{\Psi} = \sigma^2\mathbf{I}_u$ and $\mathbf{\Omega}_0 = \sigma_0^2\mathbf{I}_{p-u}$, we have the log-likelihood function:

$$\begin{aligned} & -\frac{n}{2}\log|\sigma^2\mathbf{\Gamma}\mathbf{\Gamma}^T + \sigma_0^2\mathbf{\Gamma}_0\mathbf{\Gamma}_0^T| - \frac{1}{2}\sum_{i=1}^n \tilde{x}_i^T(\sigma^2\mathbf{\Gamma}\mathbf{\Gamma}^T + \sigma_0^2\mathbf{\Gamma}_0\mathbf{\Gamma}_0^T)^{-1}\tilde{x}_i \\ = & -\frac{n}{2}u\log(\sigma^2) - \frac{n}{2}(p-u)\log(\sigma_0^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n \tilde{x}_i^T(\mathbf{\Gamma}\mathbf{\Gamma}^T)\tilde{x}_i - \frac{1}{2\sigma_0^2}\sum_{i=1}^n \tilde{x}_i^T(\mathbf{\Gamma}_0\mathbf{\Gamma}_0^T)\tilde{x}_i \\ = & -\frac{n}{2}u\log(\sigma^2) - \frac{n}{2\sigma^2}\text{trace}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}) - \frac{n}{2}(p-u)\log(\sigma_0^2) - \frac{n}{2\sigma_0^2}\text{trace}(\mathbf{\Gamma}_0^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}_0). \end{aligned}$$

Maximizing over σ^2 and σ_0^2, we have the partially maximized log-likelihood function

$$\begin{aligned} L_2(\mathcal{S}_{\mathbf{\Gamma}}) = & -\frac{n}{2}u\log(\text{trace}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma})) - \frac{n}{2}(p-u)\log(\text{trace}(\mathbf{\Gamma}_0^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}_0)) \\ & -\frac{n}{2}p + \frac{n}{2}u\log(u) + \frac{n}{2}(p-u)\log(p-u) \\ = & -\frac{n}{2}u\log(\text{trace}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma})) - \frac{n}{2}(p-u)\log(\text{trace}(\hat{\mathbf{\Sigma}}) - \text{trace}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma})) \\ & -\frac{n}{2}p + \frac{n}{2}u\log(u) + \frac{n}{2}(p-u)\log(p-u) \end{aligned}$$

The function $L_2(\mathcal{S}_{\mathbf{\Gamma}})$ requires $n > p - u + 1$ to ensure $\text{trace}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}) > 0$.

Proposition 8 *When $\mathbf{\Psi} = \sigma^2\mathbf{I}_u$ and $\mathbf{\Omega}_0 = \sigma_0^2\mathbf{I}_{p-u}$, the maximum likelihood estimator $\hat{\mathcal{S}}_{\mathbf{\Gamma}}$ is the span of either the first u principal component directions or the last u principal component directions.*

It is quite restrictive that $\mathbf{\Psi}$ is modeled as isotropic (Cook, 2007), however we demonstrate a situation that the last a few principal component directions can retain most of the sample's information.

Assuming that only $\mathbf{\Omega}_0 = \sigma_0^2\mathbf{I}_{p-u}$, we have the model

$$\begin{aligned} \mathbf{x} &= \boldsymbol{\mu} + \mathbf{\Gamma}\boldsymbol{\eta}\boldsymbol{\nu} + \mathbf{\Phi}^{1/2}\boldsymbol{\epsilon} \\ \mathbf{\Phi} &= \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \sigma_0^2\mathbf{\Gamma}_0\mathbf{\Gamma}_0^T. \end{aligned} \tag{4.5}$$

Then we have the log-likelihood function of model (4.5):

$$\begin{aligned}
& -\frac{n}{2}\log|\mathbf{\Gamma\Psi\Gamma}^T+\sigma_0^2\mathbf{\Gamma}_0\mathbf{\Gamma}_0^T|-\frac{1}{2}\sum_{i=1}^{n}\tilde{x}_i^T(\mathbf{\Gamma\Psi\Gamma}^T+\sigma_0^2\mathbf{\Gamma}_0\mathbf{\Gamma}_0^T)^{-1}\tilde{x}_i \\
= & -\frac{n}{2}\log|\mathbf{\Psi}|-\frac{n}{2}(p-u)\log(\sigma_0^2)-\frac{1}{2}\sum_{i=1}^{n}\tilde{x}_i^T(\mathbf{\Gamma\Psi}^{-1}\mathbf{\Gamma}^T)\tilde{x}_i-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}\tilde{x}_i^T(\mathbf{\Gamma}_0\mathbf{\Gamma}_0^T)\tilde{x}_i \\
= & -\frac{n}{2}\log|\mathbf{\Psi}|-\frac{n}{2}\mathrm{trace}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma\Psi}^{-1})-\frac{n}{2}(p-u)\log(\sigma_0^2)-\frac{n}{2\sigma_0^2}\mathrm{trace}(\mathbf{\Gamma}_0^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}_0).
\end{aligned}$$

Maximizing over $\mathbf{\Psi}$ and σ_0^2, we have the partially log-likelihood function

$$\begin{aligned}
L_3(\mathcal{S}_{\mathbf{\Gamma}}) = & -\frac{n}{2}\log|\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}|-\frac{n}{2}(p-u)\log(\mathrm{trace}(\mathbf{\Gamma}_0^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}_0)) \\
& -\frac{n}{2}p+\frac{n}{2}(p-u)\log(p-u) \\
= & -\frac{n}{2}\log|\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}|-\frac{n}{2}(p-u)\log(\mathrm{trace}(\hat{\mathbf{\Sigma}})-\mathrm{trace}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma})) \\
& -\frac{n}{2}p+\frac{n}{2}(p-u)\log(p-u).
\end{aligned}$$

The function $L_3(\mathcal{S}_{\mathbf{\Gamma}})$ requires $n > p$ as $\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}$ must not be singular.

Proposition 9 *Under model (4.5), the maximum likelihood estimator $\hat{\mathcal{S}}_{\mathbf{\Gamma}}$ is the span of the first k and last $u-k$ principal component directions that maximizes $L_3(\mathcal{S}_{\mathbf{\Gamma}})$ subject to $\mathbf{\Gamma}^T\mathbf{\Gamma}=\mathbf{I}_u$ where k needs to be determined.*

Let σ_i, $i=1,2,\ldots,u$ be the population eigenvalues of $\mathbf{\Psi}$, $\sigma_1<\sigma_2<\ldots<\sigma_u$. The setting of model (4.5) basically says that the signals can have different scales but the noises have the same magnitude. If $\sigma_i>\sigma_0$ for all $i=1,2,\ldots,u$, then we have the same solution as the usual principal component analysis. It is equivalent to say that if the signal is strong enough, the usual principal component analysis is doing a sensible thing. If $\sigma_0>\sigma_i$ for all $i=1,2,\ldots,u$, then we have the last u principal component directions as the solution. If σ_0 lies among σ_i for $i=1,2,\ldots,u$, then the solution is the span of the first k principal component directions and $u-k$ last principal component directions where k ranges from 1 to u. This provides a fast algorithm to search the maximizer of $L_3(\mathcal{S}_{\mathbf{\Gamma}})$ which was also addressed by Welling et. al (2003). However, it is worth to point out that the log-likelihood function of model (4.5) has the same form as PXCA only when $u=d$. The main difference between these two approaches lies in

the fact that model (4.5) aims to estimate an upper bound on the minimum dimension reduction subspaces.

If we assume $\mathbf{\Omega}_0$ in model (4.4) to be a diagonal matrix, it is equivalent to model (4.4) because $\mathbf{\Gamma}_0\mathbf{\Omega}\mathbf{\Gamma}_0^T$ can be always re-parameterized as $\mathbf{\Gamma}_0'\Lambda{\mathbf{\Gamma}_0'}^T$. The positive definite matrix $\mathbf{\Omega}_0$ in model (4.4) can be considered as the covariance matrix for $\mathbf{\Gamma}_0^T\mathbf{x}$ and it may not be always a diagonal matrix given $\mathbf{\Gamma}_0$. Suppose we can model $\mathbf{\Omega}_0$ as

$$\sigma_0^2\begin{pmatrix} 1 & c & c & \dots & c \\ c & 1 & c & \dots & c \\ c & c & 1 & \dots & c \\ \vdots & & & \ddots & \vdots \\ c & c & c & \dots & 1 \end{pmatrix}. \tag{4.6}$$

This means that the correlation coefficients for $\mathbf{\Gamma}_0^T\mathbf{x}$ are modeled as constant c where $-1/(p-u-1) < c < 1$. We can represent $\mathbf{\Omega}_0$ as $\sigma_0^2\{(1-c)\mathbf{Q}_1 + (1+(p-u-1)c)\mathbf{P}_1\}$ where $\mathbf{P}_1$ is the projection matrix onto the $(p-u)\times 1$ vector of ones and $\mathbf{Q}_1 = \mathbf{I}_{p-u} - \mathbf{P}_1$. The log-likelihood function can be calculated as

$$\begin{aligned} & -\frac{n}{2}\log|\mathbf{\Psi}| - \frac{n}{2}\mathrm{trace}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}\mathbf{\Psi}^{-1}) - \frac{n}{2}\log|\mathbf{\Omega}_0| - \frac{n}{2}\mathrm{trace}(\mathbf{\Gamma}_0^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}_0\mathbf{\Omega}_0^{-1}) \\ = \; & -\frac{n}{2}\log|\mathbf{\Psi}| - \frac{n}{2}\mathrm{trace}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}\mathbf{\Psi}^{-1}) - \frac{n}{2}(p-u)\log\sigma_0^2 - \frac{n}{2}(p-u-1)\log(1-c) \\ & -\frac{n}{2}\log(1+(p-u-1)c) - \frac{n}{2\sigma_0^2}\mathrm{trace}\{\mathbf{\Gamma}_0^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}_0\left(\frac{\mathbf{Q}_1}{1-c} + \frac{\mathbf{P}_1}{1+(p-u-1)c}\right)\}. \end{aligned}$$

Maximizing over $\mathbf{\Psi}$ and σ_0^2 first, we have

$$\begin{aligned} & -\frac{n}{2}\log|\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}| - \frac{n}{2}(p-u-1)\log(1-c) - \frac{n}{2}\log(1+(p-u-1)c) \\ & -\frac{n}{2}\log\{\mathrm{trace}(\mathbf{\Gamma}_0^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}_0\left(\frac{\mathbf{Q}_1}{1-c} + \frac{\mathbf{P}_1}{1+(p-u-1)c}\right))\} - \frac{n}{2}p + \frac{n(p-u)}{2}\log(p-u). \end{aligned}$$

Then maximizing above over c, we have the partially maximized log-likelihood

$$\begin{aligned} L_5(\mathcal{S}_{\mathbf{\Gamma}}) \;=\; & -\frac{n}{2}\log|\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}| - \frac{n(p-u-1)}{2}\log\mathrm{trace}(\mathbf{\Gamma}_0^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}_0\mathbf{Q}_1) \\ & -\frac{n}{2}\log\mathrm{trace}(\mathbf{\Gamma}_0^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}_0\mathbf{P}_1) - \frac{n}{2}p + \frac{n(p-u-1)}{2}\log(p-u-1), \end{aligned}$$

where the maximum likelihood estimators $\hat{\sigma}_0^2 = \mathrm{trace}(\mathbf{\Gamma}_0^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}_0)/(p-u)$ and

$$\hat{c} = 1 - \frac{(p-u)\mathrm{trace}(\mathbf{\Gamma}_0^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}_0\mathbf{Q}_1)}{(p-u-1)\mathrm{trace}(\mathbf{\Gamma}_0^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}_0)}.$$

Proposition 10 *When* $\mathbf{\Psi} > 0$ *and* $\mathbf{\Omega}_0 = (4.6)$*, the maximum likelihood estimator* $\hat{\mathcal{S}}_{\mathbf{\Gamma}}$ *is the span of one subset of* u *principal component directions that maximizes* $L_5(\mathcal{S}_{\mathbf{\Gamma}})$ *subject to* $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_u$.

In this setting, the coordinate matrix $\mathbf{\Omega}_0$ has two different eigenvalues, one with $p-u-1$ replicates. From Proposition 10, we have a simple algorithm to find the maximum likelihood estimator $\hat{\mathcal{S}}_{\mathbf{\Gamma}}$. Among the first k and last $u+1-k$ principal component directions where k ranges from 0 to $u+1$, we record any subset with dimension u as a possible candidate. The total number of candidates is less than $(u+2)(u+1)$. Then we search among all candidates and find the one that maximizes L_5.

4.1.4 Selection of the dimension u

We use the likelihood ratio test to determine the dimensionality of the envelope denoted by u. For example, let us consider model (4.5). The hypothesis $u = u_0$ can be tested by using the likelihood ratio statistic $\Lambda(u_0) = 2(\hat{L}_{fm} - \hat{L}^{(u_0)})$, where $\hat{L}_{fm}$ denotes the maximum value of the log likelihood for the full model ($u = p$), and $\hat{L}^{(u_0)}$ the maximum value of the log likelihood when $u = u_0$. In fact, $\hat{L}_{fm} = -(np)/2 - (n/2)\log|\hat{\mathbf{\Sigma}}|$. The total number of parameters needed to estimate (4.5) is

$$df(u) = p + u(p-u) + \frac{u(u+1)}{2} + 1.$$

The first term on the right hand side corresponds to the grand mean $\boldsymbol{\mu}$. The second term corresponds to the unconstrained symmetric matrix $\mathbf{\Psi}$. The third term corresponds the number of parameters needed to describe the subspace $\mathcal{S}_{\mathbf{\Gamma}}$ (Edelman et al, 1998). The last term corresponds to σ_0^2. Following standard likelihood theory, under the null hypothesis, $\Lambda(u_0)$ is distributed asymptotically as a chi-squared random variable with $(p-u_0+2)(p-u_0-1)/2$ degrees of freedom.

4.1.5 Simulation studies

Two small simulation studies were conducted in this section. We generated data from model (4.5) with

$$\mathbf{\Gamma} = \begin{pmatrix} 1 & 1 & \dots & 1 & 1 \\ 1 & -1 & \dots & 1 & -1 \end{pmatrix}^T / \sqrt{20} \in \mathbb{R}^{20\times 2},$$

$\boldsymbol{\eta} = (1, 5)^T$, $\boldsymbol{\Omega} = \mathbf{I}_2$ and $\sigma_0 = 2$. We have $p = 20$, $u = 2$ and $d = 1$.

Since $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$ is a vector in this setting, $\mathcal{S}$ has dimension one. There is no maximum likelihood estimator for $\boldsymbol{\beta}$, however there we can estimate $\mathcal{E}_{\boldsymbol{\Phi}}(\mathcal{S})$, an upper bound of $\mathcal{S}$. It can be shown that the population maximum likelihood estimator, denoted as $\mathcal{S}_{\boldsymbol{\Gamma}}$, is the span of the first and the last eigenvector of $\boldsymbol{\Sigma}$. It is clear to see that the usual principal component analysis fails. On each of 100 replications with fixed n we computed the maximum angle between $\hat{\mathcal{S}}_{\boldsymbol{\Gamma}}$ and span($\boldsymbol{\beta}$). Figure 4.1 summarizes the average maximum angles versus n. We see that the span of the principal envelope solution is very efficient at estimating an upper bound of span($\boldsymbol{\beta}$).

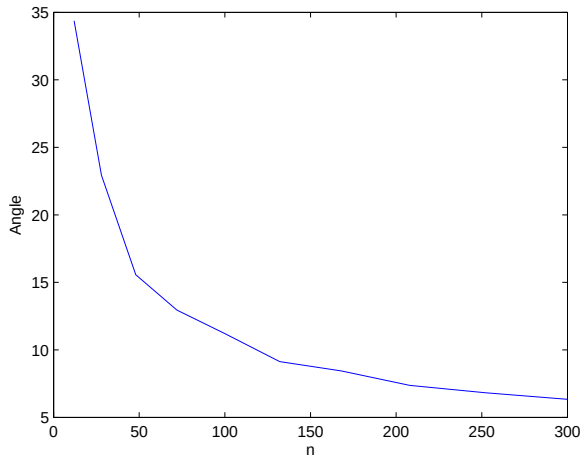

Figure 4.1: Average angles versus n when $p = 20$

In the second simulation study, we generated data from

$$\mathbf{x} = \boldsymbol{\Gamma}\boldsymbol{\nu} + \boldsymbol{\epsilon},\ \boldsymbol{\Phi} = 0.1\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T + 10\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T,$$

where $\boldsymbol{\Gamma}$ was set the same as the first study and $\boldsymbol{\nu}$ was generated as points on a 2×2 square instead of a normal distribution to visualize the effectiveness of principal envelope model (4.5). We have $p = 20$, $u = 2$ and $d = 2$. Figure 4.2 shows the scatter plot of $\mathbf{PE}_2 = \mathbb{X}\hat{\mathbf{e}}_2$ versus $\mathbf{PE}_1 = \mathbb{X}\hat{\mathbf{e}}_1$, and the scatter plot of $\mathbf{PC}_2 = \mathbb{X}\hat{\mathbf{g}}_2$ versus $\mathbf{PC}_1 = \mathbb{X}\hat{\mathbf{g}}_1$ for two different sample size $n = 200$ and $n = 400$ with one replication where $\hat{\mathbf{e}}_1$ and $\hat{\mathbf{e}}_2$ represent the two directions of $\hat{\mathcal{S}}_{\boldsymbol{\Gamma}}$. From Figure 4.2, we can see that the solution of principal envelope model (4.5) can recover the square well while the solution of principal

component analysis fails. We notice that in this setting $\boldsymbol{\nu}$ does not follow a normal distribution which is required by model (4.5), however the solution is still reasonable. This tells us that principal envelope model can be robust.

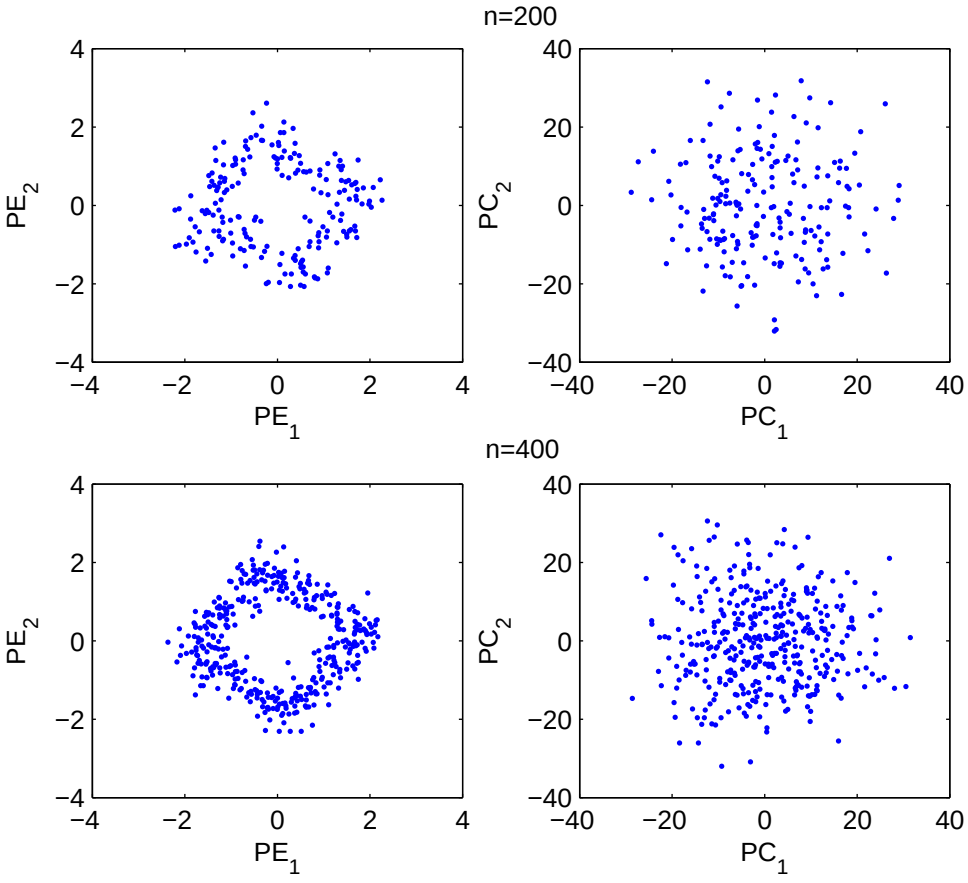

Figure 4.2: Recover the square using PEM and PCA methods

4.2 Feature selection

There is a growing number of applications in which the data collected has a large number of features while the number of samples is limited. It is desirable to screen out irrelevant and redundant features from the data and keep the most of the sample's information in order to improve the performance of classification or prediction. In this section, we

propose a new method based on the maximum likelihood functions derived in previous sections to select useful features.

4.2.1 A simple coordinate-independent penalty function

Following Section 3.2.1, we introduce here a simple coordinate-independent penalty function, depending only on the subspace $\mathcal{S}_{\mathbf{\Gamma}}$ where $\mathbf{\Gamma}$ stands for an orthonormal basis of the subspace. Let $\mathbf{\Gamma} = (\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_p)^T$ where $\boldsymbol{\gamma}_i^T$ stands for the i-th row vector of $\mathbf{\Gamma}$, $i = 1, \ldots, p$.

Given $\theta_1 = \ldots \theta_p = \pi$ in the penalty function 3.5, a simple coordinate-independent penalty function can be defined as

$$\rho_0(\mathbf{\Gamma}) = \sum_i \pi \|\boldsymbol{\gamma}_i\|_2.$$

This special case $\rho_0(\mathbf{\Gamma})$ has the same format as the group lasso proposed by M. Yuan and Y. Lin (2006) but their concepts and usages are essentially different. We use $\rho_0(\mathbf{\Gamma})$ to select features in the following sections.

4.2.2 Methodology

To show how we select features, partition $\mathbf{x}$ as $(\mathbf{x}_1^T, \mathbf{x}_2^T)^T$, where $\mathbf{x}_1$ corresponds to q elements of $\mathbf{x}$ and $\mathbf{x}_2$ to the remaining elements. If

$$\boldsymbol{\nu} \perp\!\!\!\perp \mathbf{x}_2 | \mathbf{x}_1, \tag{4.7}$$

then $\mathbf{x}_2$ can be removed, as given $\mathbf{x}_1$, $\mathbf{x}_2$ contains no further information about $\boldsymbol{\nu}$. Let $\boldsymbol{\varphi}$ be an orthonormal basis for the minimum dimension reduction subspace $\mathcal{T}$ and partition $\boldsymbol{\varphi} = (\boldsymbol{\varphi}_1^T, \boldsymbol{\varphi}_2^T)^T$ in accordance with the partition of $\mathbf{x}$. Then the condition (4.7) is equivalent to $\boldsymbol{\varphi}_2 = 0$, so the corresponding rows of the basis are zero vector. This motivates us to shrink the rows of $\boldsymbol{\varphi}$ corresponding irrelevant features to 0.

If we believe in model (4.1), then $\boldsymbol{\varphi}$ can be estimated by the first d principal component directions that can be obtained by maximizing $\text{trace}(\mathbf{\Gamma}^T \hat{\mathbf{\Sigma}} \mathbf{\Gamma})$ subject to $\mathbf{\Gamma}^T \mathbf{\Gamma} = \mathbf{I}_d$. The regularized solution $\tilde{\mathcal{S}}_{\mathbf{\Gamma}}$ can obtained by minimizing the following

$$-\text{trace}(\mathbf{\Gamma}^T \hat{\mathbf{\Sigma}} \mathbf{\Gamma}) + \rho_0(\mathbf{\Gamma}), \tag{4.8}$$

subject to $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_d$.

When $\mathcal{T}$ is not estimable, we use principal envelope model to estimate an upper bound $\mathcal{E}_{\mathbf{\Phi}}(\mathcal{T})$. Let $\mathbf{\Gamma}$ be an orthonormal basis of $\mathcal{E}_{\mathbf{\Phi}}(\mathcal{T})$. Partition $\mathbf{\Gamma} = (\mathbf{\Gamma}_1^T, \mathbf{\Gamma}_2^T)^T$ in accordance with the partition of $\mathbf{x}$. If $\mathbf{\Gamma}_2 = 0$, then $\boldsymbol{\varphi}_2 = 0$. However $\boldsymbol{\varphi}_2 = 0$ does not imply $\mathbf{\Gamma}_2 = 0$. This tells us that it is more conservative to select feature from the principal envelope model.

If we believe in model (4.5), then the regularized estimator $\tilde{\mathcal{S}}_{\mathbf{\Gamma}}$ can be obtained by minimizing the following

$$\log|\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}| + (p-u)\log(\text{trace}(\hat{\mathbf{\Sigma}}) - \text{trace}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma})) + \rho_0(\mathbf{\Gamma}). \tag{4.9}$$

subject to $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_u$. The tuning parameter π in $\rho_0(\mathbf{\Gamma})$ can be chosen using cross validation or an information criterion like AIC or BIC. For convenience, we use the sparse basis matrix $\tilde{\mathbf{\Gamma}}$ to denote the regularized solution $\tilde{\mathcal{S}}_{\mathbf{\Gamma}}$.

4.2.3 Algorithm

To overcome the non-differentiability of the ρ_0 function, we adopted the local quadratic approximation of Fan, et al. (2001). Suppose we want to minimize $f(\mathbf{\Gamma}) + \rho_0(\mathbf{\Gamma})$ subject to $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_u$. First assume $\forall i$, $\|\boldsymbol{\gamma}_i\|_2 \neq 0$ where $\boldsymbol{\gamma}_i^T$ denotes the ith row vector of $\mathbf{\Gamma}$. Taking the unconstrainted derivative of $\rho_0(\mathbf{\Gamma})$ with respect to the $p \times u$ matrix $\mathbf{\Gamma}$, we have

$$\frac{\partial\rho_0}{\partial\mathbf{\Gamma}} = \text{diag}\left(\frac{\pi}{\|\boldsymbol{\gamma}_1\|_2}, \cdots, \frac{\pi}{\|\boldsymbol{\gamma}_i\|_2}, \cdots, \frac{\pi}{\|\boldsymbol{\gamma}_p\|_2}\right)\mathbf{\Gamma}.$$

An iterative algorithm can be constructed as follows. Let $\tilde{\mathbf{\Gamma}}^0 = (\tilde{\boldsymbol{\gamma}}_1^0, \cdots, \tilde{\boldsymbol{\gamma}}_p^0)^T$ denote a starting value for $\mathbf{\Gamma}$, usually the solution without the penalty. The first derivative of ρ_0 can be approximated by

$$\frac{\partial\rho_0}{\partial\mathbf{\Gamma}} = \text{diag}\left(\frac{\pi}{\|\tilde{\boldsymbol{\gamma}}_1^0\|_2}, \cdots, \frac{\pi}{\|\tilde{\boldsymbol{\gamma}}_i^0\|_2}, \cdots, \frac{\pi}{\|\tilde{\boldsymbol{\gamma}}_p^0\|_2}\right)\mathbf{\Gamma} = \mathbf{H}^0\mathbf{\Gamma}$$

where diag denotes the usual diagonal operator. Expanding ρ_0 to the second order unconstrained Taylor series, we have

$$\rho_0(\mathbf{\Gamma}) \approx \frac{1}{2}\text{trace}(\mathbf{\Gamma}^T\mathbf{H}^0\mathbf{\Gamma}) + C_0,$$

where C_0 stands for a constant with respect to $\mathbf{\Gamma}$. Then minimizing

$$f(\mathbf{\Gamma}) + \frac{1}{2}\text{trace}(\mathbf{\Gamma}^T\mathbf{H}^0\mathbf{\Gamma})$$

subject to $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_u$ gives a first approximation $\tilde{\mathbf{\Gamma}}^1$ to $\tilde{\mathbf{\Gamma}}$. The starting value $\tilde{\mathbf{\Gamma}}^1 = (\tilde{\boldsymbol{\gamma}}_1^1, \cdots, \tilde{\boldsymbol{\gamma}}_p^1)^T$ for the next iteration can be usually solved via Grassmann manifolds optimization and the approximation of $\rho_0(\mathbf{\Gamma})$ is updated by

$$\rho_0(\mathbf{\Gamma}) \approx \frac{1}{2}\text{trace}(\mathbf{\Gamma}^T\mathbf{H}^1\mathbf{\Gamma}) + C_1,$$

where C_1 stands for a constant function with respect to $\mathbf{\Gamma}$. Then minimizing

$$f(\mathbf{\Gamma}) + \frac{1}{2}\text{trace}(\mathbf{\Gamma}^T\mathbf{H}^1\mathbf{\Gamma}),$$

subject to $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_u$ with $\mathbf{H}^1 = \text{diag}(\pi/\|\tilde{\boldsymbol{\gamma}}_1^1\|_2, \cdots, \pi/\|\tilde{\boldsymbol{\gamma}}_i^1\|_2, \cdots, \pi/\|\tilde{\boldsymbol{\gamma}}_p^1\|_2)$, gives a second approximation to $\tilde{\mathbf{\Gamma}}$. The process repeats until it converges. During the process, if $\|\tilde{\boldsymbol{\gamma}}_i^k\|_2 \approx 0$, then the feature x_i is removed. With respect to the choice of the initial values $\tilde{\mathbf{\Gamma}}^0$, a simple but effective solution is to set it to be the minimizer of $f(\mathbf{\Gamma})$ subject to $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_u$.

4.2.4 Simulation studies

We define three rates r_1, r_2 and r_3 to assess how well the methods select features. The rate r_1 is defined as the fraction of the average number of non-zero row vectors of $\tilde{\mathbf{\Gamma}}$ associated with active features to the true number of active features. The rate r_2 is defined as the fraction of the average number of zero row vectors of $\tilde{\mathbf{\Gamma}}$ associated with inactive features to the true number of inactive features. The rate r_3 is defined as the fraction of replicates in which the regularized method selects both active and inactive features exactly right. The tuning parameter π is chosen using BIC following the discussion in Section 3.16.

Two simulation studies were conducted as follows. We first generated 100 replications from model (4.2) with

$$\mathbf{\Gamma} = \begin{pmatrix} 0.5 & 0.5 & 0.5 & 0.5 & 0 & \dots & 0 \\ 0.5 & -0.5 & 0.5 & -0.5 & 0 & \dots & 0 \end{pmatrix}^T \in \mathbb{R}^{20\times 2}, \boldsymbol{\eta} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix},$$

for each $\sigma = 1$, 2 and 3 with sample size $n = 30$. We have $p = 20$ and $d = 2$. The regularized solution $\tilde{\mathbf{\Gamma}}$ was obtained through minimizing formula (4.8). Table 4.1 shows the average rates for three different levels of noise. It is not surprising that when the noise gets larger, all rates go down.

Table 4.1: Rates of feature selection from a PPCA model

Rate	r_1	r_2	r_3
Sample size		$n = 30$	
$\sigma = 1$	0.998	1.000	0.990
$\sigma = 2$	0.985	0.976	0.630
$\sigma = 3$	0.975	0.839	0.260

In the second simulations study, we generated data from model (4.5) with the same $\mathbf{\Gamma}$ and $\boldsymbol{\eta}$ as above. We set the parameter $\mathbf{\Omega} = \mathbf{I}_2$, the sample size $n = 50$ and $\sigma_0 = 3$, 5, 10. The regularized solution $\tilde{\mathbf{\Gamma}}$ was obtained through minimizing formula (4.9). Table 4.2 shows the average rates for 100 replications. We see that the regularized method performs better when σ_0 increases. This is because the principal envelope model (4.5) is more efficient when the eigenvalues of $\mathbf{\Psi}$ differ a lot from σ_0.

We do not compare our new regularized method with existing sparse principal component methods, for example, sparse principal component analysis (SPCA) introduced by Zou, et al. (2006). Existing methods start from the point of view that they try to keep most of the variance of the data while our method starts from a very different point of view of the sufficient dimension reduction concept based on the latent variable model (4.3).

Table 4.2: Rates of feature selection from a principal envelope model

Rate	r_1	r_2	r_3
Sample size		$n = 50$	
$\sigma_0 = 3$	0.945	0.301	0
$\sigma_0 = 5$	1	0.682	0.400
$\sigma_0 = 10$	1	0.867	0.540

4.3 Data analysis

We applied our method to the man hours data, a good example of multicollinearity data structure. This data contains 25 observations, one response y and 7 explanatory variables $x_1, \ldots, x_7$. The response variable y is monthly man hours needed to operate an establishment of U.S. Navy bachelor officers' quarters. The 7 explanatory variables are average daily occupancy (x_1); monthly average number of checkins (x_2); weekly hours of service desk operation (x_3); common use area (in square feet) (x_4); number of building wings (x_5); operational berthing capacity (x_6); number of rooms (x_7). The correlation coefficients between x_6 and x_7, x_2 and x_7, and x_2 and x_6 are 0.98, 0.86, 0.85 respectively. As a usual treatment in PCA, we standardize $x_1, \ldots, x_7$ to $z_1, \ldots, z_7$ with mean 0 and standard deviation 1. Let $p_1, \ldots, p_7$ be the eigenvectors of the correlation matrix with descending eigenvalues 4.672, 0.742, 0.676, 0.451, 0.298, 0.152, 0.010. The data can be downloaded from the web site $http://www.stat.uconn.edu/\ nalini/fcilmweb/example14.html$.

We use model (4.5) to fit the data and the likelihood ratio tests suggest $u = 2$. The corresponding principal envelope solution consists the first and the seventh eigenvector. The p-value for the hypothesis test of $u = 2$ against the full model is 0.19 while the p-value for the hypothesis test of $u = 3$ against the full model is 0.73. As the sample size is only 25 and the power is not big enough, we might choose $u = 3$ in a conservative way. Given $u = 3$, it turns out that the maximum likelihood estimator is the span of the first, the sixth and the seventh eigenvectors.

Let $\mathbf{m}$ be the regression coefficient vector of y on $z_1, \ldots, z_7$. The first 5 principal components explain 98% total variance while the largest principal angle (Knyazev and Argentati, 2002) between $\mathbf{m}$ and the subspace spanned by the first 5 principal component directions is about 75 degrees. The principal envelope solution, the first and the seventh principal components explain 67% total variance while the largest principal angle between m and the subspace spanned by principal envelope solution is about 11 degrees. The maximum likelihood estimator with $u = 3$ explain 69% total variance while the largest principal angle between $\mathbf{m}$ and the subspace spanned by the first, the sixth and the seventh principal component directions is about 9 degrees. Even though model (4.5) does not use any information on y, it detects useful principal components

that relate to the regression of y on z_i. This demonstrates that the principal envelope model can be much more efficient than PCA.

4.4 Discussion

We have seen that if the error structure deviates from the isotropic error in the model (4.1), the usual PCA may not work anymore. Motivated by a general error structure, we establish probabilistic models named as principal envelope model that show any combination of principal component directions could contain most of the sample's information. Under more specific principal envelope models, we are able to discern which combination is useful via maximum likelihood estimators. Hence we provide an alternative to PCA in multivariate analysis when PCA fails. We also studied several different structures of $\mathbf{\Omega}_0$ in model (4.4).

4.5 Appendix

A few Lemmas

Lemma 3 *Let* $\mathbf{\Gamma} = [\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_d]$ *where* $\boldsymbol{\gamma}_i$ *stands for the ith column of* $\mathbf{\Gamma}$ *and* $\hat{\mathbf{\Gamma}} = [\hat{\mathbf{g}}_1, \cdots, \hat{\mathbf{g}}_d]$ *where* $\hat{\mathbf{g}}_i$ *denotes the ith principal component direction. Then* $\hat{\mathcal{S}}_{\mathbf{\Gamma}}$ *maximizes the objective function* $\text{trace}(\mathbf{\Gamma}^T \hat{\mathbf{\Sigma}} \mathbf{\Gamma})$ *subject to* $\mathbf{\Gamma}^T \mathbf{\Gamma} = \mathbf{I}_d$.

PROOF: Let $\boldsymbol{\gamma}_i = \sum_{j=1}^{p} c_{ij} \hat{\mathbf{g}}_j$. Then

$$\text{trace}(\mathbf{\Gamma}^T \hat{\mathbf{\Sigma}} \mathbf{\Gamma}) = \sum_{i=1}^{d} \boldsymbol{\gamma}_i^T \hat{\mathbf{\Sigma}} \boldsymbol{\gamma}_i = \sum_{i=1}^{d} \sum_{j=1}^{p} \lambda_j c_{ij}^2 = \sum_{j=1}^{p} \lambda_j \left(\sum_{i=1}^{d} c_{ij}^2 \right).$$

where $\hat{\lambda}_j$ denotes the jth eigenvalue.

Since $\sum_{i=1}^{d} c_{ij}^2 \leq 1$ and $\sum_{j=1}^{p} \sum_{i=1}^{d} c_{ij}^2 = d$, to maximize $\text{trace}(\mathbf{\Gamma}^T \hat{\mathbf{\Sigma}} \mathbf{\Gamma})$, the optimum situation arrives when $\sum_{i=1}^{d} c_{i1}^2 = \sum_{i=1}^{d} c_{i2}^2 = \cdots = \sum_{i=1}^{d} c_{id}^2 = 1$ and $\sum_{i=1}^{d} c_{i(d+1)}^2 = \cdots = \sum_{i=1}^{d} c_{ip}^2 = 0$. It is clear to see that $\hat{\boldsymbol{\gamma}}_i = \hat{\mathbf{g}}_i$ for $i = 1, \cdots, d$ satisfying the optimum condition. The global maximum value of $\text{trace}(\mathbf{\Gamma}^T \hat{\mathbf{\Sigma}} \mathbf{\Gamma})$ equals $\sum_{i=1}^{d} \hat{\lambda}_i$.

Lemma 4 *Let* $\mathbf{\Gamma} = [\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_d]$ *where* $\boldsymbol{\gamma}_i$ *stands for the* i*th column of* $\mathbf{\Gamma}$ *and* $\hat{\mathbf{\Gamma}} = [\hat{\mathbf{g}}_1, \cdots, \hat{\mathbf{g}}_d]$ *where* $\hat{\mathbf{g}}_i$ *denotes the* i*th principal component direction. Then* $\hat{\mathcal{S}}_{\mathbf{\Gamma}}$ *maximizes the objective function*

$$\text{trace}\left(\mathbf{\Gamma}^T \hat{\mathbf{\Sigma}} \mathbf{\Gamma} \text{diag}(k_1, k_2, \cdots, k_d)\right)$$

subject to $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_d$ *where* $k_1 \geq k_2 \geq \cdots \geq k_d$ *are positive real numbers.*

PROOF: Let $\mathbf{\Gamma}_i = [\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_{d-i}]$ for $i = 1, \cdots, d-1$. It is clear to see that

$$\begin{aligned} & \text{trace}\left(\mathbf{\Gamma}^T \hat{\mathbf{\Sigma}} \mathbf{\Gamma} \text{diag}(k_1, k_2, \cdots, k_d)\right) \\ = \quad & k_d \text{trace}(\mathbf{\Gamma}^T \hat{\mathbf{\Sigma}} \mathbf{\Gamma}) + (k_{d-1} - k_d)\text{trace}(\mathbf{\Gamma}_1^T \hat{\mathbf{\Sigma}} \mathbf{\Gamma}_1) + \cdots + (k_1 - k_2)\text{trace}(\mathbf{\Gamma}_{d-1}^T \hat{\mathbf{\Sigma}} \mathbf{\Gamma}_{d-1}). \end{aligned}$$

From Lemma 3, we know $\hat{\mathbf{\Gamma}}$ maximizes every term on the right side of the formula above. The global maximum value of

$$\text{trace}\left(\mathbf{\Gamma}^T \hat{\mathbf{\Sigma}} \mathbf{\Gamma} \text{diag}(k_1, k_2, \cdots, k_d)\right)$$

equals $\sum_{i=1}^{d} \lambda_i k_i$. If $k_1, k_2, \cdots, k_d$ are not in descending order, then we need permutate columns of $\hat{\mathbf{\Gamma}}$ such that it corresponds the order of k_i. However the subspace spanned by $\hat{\mathbf{\Gamma}}$ does not change.

Lemma 5 *Let* $\mathbf{\Gamma} = [\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_d]$ *where* $\boldsymbol{\gamma}_i$ *stands for the* i*th column of* $\mathbf{\Gamma}$ *and* $\hat{\mathbf{\Gamma}} = [\hat{\mathbf{g}}_1, \cdots, \hat{\mathbf{g}}_d]$ *where* $\hat{\mathbf{g}}_i$ *denotes the* i*th principal component direction. Then* $\hat{\mathcal{S}}_{\mathbf{\Gamma}}$ *maximizes the objective function* $\text{trace}(\mathbf{\Gamma}^T \hat{\mathbf{\Sigma}} \mathbf{\Gamma} \mathbf{\Phi})$ *subject to* $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_d$ *where* $\mathbf{\Phi}$ *is a* $d \times d$ *positive definite matrix.*

PROOF: Let $A\text{diag}(m_1, \cdots, m_d)A^T$ be the spectral decomposition of $\mathbf{\Phi}$ where $m_1 \geq m_2 \geq \cdots \geq m_d$ are positive real numbers. Then

$$\text{trace}(\mathbf{\Gamma}^T \hat{\mathbf{\Sigma}} \mathbf{\Gamma} \mathbf{\Phi}) = \text{trace}\left(A^T \mathbf{\Gamma}^T \hat{\mathbf{\Sigma}} \mathbf{\Gamma} A \text{diag}(m_1, \cdots, m_d)\right).$$

Since $(\mathbf{\Gamma}A)^T \mathbf{\Gamma} A = \mathbf{I}_d$, from Lemma 4, $\text{span}(\hat{\mathbf{\Gamma}})$ is the maximum likelihood estimator of $\text{span}(\mathbf{\Gamma}A)$ in the objective function above while $\text{span}(\mathbf{\Gamma}A) = \text{span}(\mathbf{\Gamma})$.

Lemma 6 *Let a real function* $f(x) = \log(x) + C\log(K - x)$ *defined on the interval* $[a, b]$, $0 < a < K/(1 + C) < b < K$, *then* $f(x)$ *reaches its maximum at* $K/(1 + C)$ *and reaches its minimum at either* a *or* b.

It is easy to calculate the first derivative of $f(x)$

$$f^{'}(x) = \frac{K-(1+C)x}{x(K-x)},$$

and the second derivative

$$f^{''}(x) = -\frac{1}{x^2} - \frac{C}{(K-x)^2} < 0.$$

We see that $f(x)$ is concave with the only stationary point $K/(1+C)$. So we can conclude that $f(x)$ reaches its maximum at $K/(1+C)$ and reaches its minimum at the boundary point, either a or b.

Proof of Proposition 6

We can rewrite $L_0(\mathcal{S}_{\mathbf{\Gamma}}, \mathbf{V}, \sigma^2)$ as

$$-\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{V}| - \frac{n}{2}(p-d)\log(\sigma^2) - \frac{n}{2\sigma^2}\text{trace}(\hat{\mathbf{\Sigma}}) + \frac{n}{2}\text{trace}[\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}\{(1/\sigma^2)\mathbf{I}_d - \mathbf{V}^{-1}\}].$$

It is clear to see that

$$\frac{1}{\sigma^2}\mathbf{I}_d - \mathbf{V}^{-1}$$

is a positive definite matrix by the definition of $\mathbf{V}$, from Lemma 5, we can conclude that the subspace spanned by the first d principal component directions is the maximum likelihood estimator of span($\mathbf{\Gamma}$) in this case.

Proof of Proposition 7

Since

$$\log|\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}| + \log|\mathbf{\Gamma}_0^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}_0| \quad \geq \quad \log|\hat{\mathbf{\Sigma}}|,$$

we have $L_1(\mathbf{\Gamma}) \leq -(np)/2 - (n/2)\log|\hat{\mathbf{\Sigma}}|$ for all $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_u$. Also

$$\begin{aligned}
& \log|\hat{\mathbf{\Gamma}}_{\mathcal{J}}^T\hat{\mathbf{\Sigma}}\hat{\mathbf{\Gamma}}_{\mathcal{J}}| + \log|\hat{\mathbf{\Gamma}}_{\mathcal{J}_0}^T\hat{\mathbf{\Sigma}}\hat{\mathbf{\Gamma}}_{\mathcal{J}_0}| \\
= \quad & \log|\hat{\mathbf{\Gamma}}_{\mathcal{J}}^T\hat{\mathbf{\Sigma}}\hat{\mathbf{\Gamma}}_{\mathcal{J}}| + \log|\hat{\mathbf{\Gamma}}_{\mathcal{J}}^T\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{\Gamma}}_{\mathcal{J}}| + \log|\hat{\mathbf{\Sigma}}| \\
= \quad & \log(\hat{\lambda}_{\mathcal{J}_1}\ldots\hat{\lambda}_{\mathcal{J}_u}) + \log(\hat{\lambda}_{\mathcal{J}_1}^{-1}\ldots\hat{\lambda}_{\mathcal{J}_u}^{-1}) + \log|\hat{\mathbf{\Sigma}}| \\
= \quad & \log|\hat{\mathbf{\Sigma}}|.
\end{aligned}$$

Then we know $L_1(\hat{\mathbf{\Gamma}}_{\mathcal{J}}) = -(np)/2 - n/2\log|\hat{\mathbf{\Sigma}}|$.

Proof of Proposition 8

Maximizing $L_2(\mathcal{S}_{\mathbf{\Gamma}})$ is equivalent to minimizing

$$\log(\text{trace}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma})) + \frac{p-u}{u}\log(\text{trace}(\hat{\mathbf{\Sigma}}) - \text{trace}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma})).$$

Let $x = \text{trace}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma})$, $C = (p-u)/u$ and $K = \text{trace}(\hat{\mathbf{\Sigma}})$. With probability one, we have

$$\min\{\text{trace}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma})\} < \frac{K}{1+C} = \frac{u}{p}\text{trace}(\hat{\mathbf{\Sigma}}) < \max\{\text{trace}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma})\}$$

subject to $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_d$.

Following Lemma 6, we know that the maximum value of $L_2(\mathbf{\Gamma})$ is reached at either $\max\{\text{trace}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma})\}$ or $\min\{\text{trace}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma})\}$. That is to say, the maximum likelihood estimator of $\mathbf{\Gamma}$ is either the first u principal component directions or the last u principal component directions.

Proof of Proposition 9

Maximizing $L_3(\mathcal{S}_{\mathbf{\Gamma}})$ is equivalent to minimize

$$\log|\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}| + (p-u)\log(\text{trace}(\hat{\mathbf{\Sigma}}) - \text{trace}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma})) \tag{4.10}$$

subject to $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_u$. Using the Lagrange multiplier rule, the solution $\hat{\mathbf{\Gamma}}$ can be gotten by finding the stationary points of the following unconstrained function:

$$\log|\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}| + (p-u)\log(\text{trace}(\hat{\mathbf{\Sigma}}) - \text{trace}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma})) + \text{trace}(\mathbf{U}(\mathbf{\Gamma}^T\mathbf{\Gamma} - \mathbf{I}_u))$$

where $\mathbf{U}$ is a $u \times u$ matrix of the Lagrange multipliers. Taking derivatives with respect to $\mathbf{\Gamma}$ and $\mathbf{U}$, we have the condition that the stationary points must satisfy

$$2\hat{\mathbf{\Sigma}}\mathbf{\Gamma}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma})^{-1} - \frac{2(p-u)\hat{\mathbf{\Sigma}}\mathbf{\Gamma}}{\text{trace}(\hat{\mathbf{\Sigma}}) - \text{trace}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma})} + \mathbf{\Gamma}(\mathbf{U} + \mathbf{U}^T) = 0$$

subject to $\mathbf{\Gamma}^T\mathbf{\Gamma} - \mathbf{I}_u = 0$. Let $w = (\text{trace}(\hat{\mathbf{\Sigma}}) - \text{trace}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}))/(p-u)$. Then $\mathbf{U} + \mathbf{U}^T = (2/w)\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma} - 2\mathbf{I}_u$. Substituting $\mathbf{U} + \mathbf{U}^T$ into the condition above, we have

$$\hat{\mathbf{\Sigma}}\mathbf{\Gamma}\{(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma})^{-1} - \frac{1}{w}\mathbf{I}_u\} = \mathbf{\Gamma}\{\mathbf{I}_u - \frac{1}{w}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma})\} \tag{4.11}$$

subject to $\mathbf{\Gamma}^T\mathbf{\Gamma} - \mathbf{I}_u = 0$.

If w is not equal to any eigenvalue of the $u\times u$ matrix $\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}$, the matrices $(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma})^{-1}-(1/w)\mathbf{I}_u$ and $\mathbf{I}_u-(1/w)(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma})$ are of full rank. Then span$(\hat{\mathbf{\Sigma}}\mathbf{\Gamma})$ must equal span$(\mathbf{\Gamma})$, implying that $\hat{\mathcal{S}}_{\mathbf{\Gamma}}$ has to be the span of one subset of u principal component directions.

If w equals an eigenvalue of $\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}$, we will show that $\hat{\mathbf{\Gamma}}$ can not be the maximizer of $L_3(\mathcal{S}_{\mathbf{\Gamma}})$. Then the matrices $(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma})^{-1}-(1/w)\mathbf{I}_u$ and $\mathbf{I}_u-(1/w)(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma})$ are singular with rank $u-1$. Let the spectral decomposition of $\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}$ be $\boldsymbol{\theta}\text{diag}(\kappa_1,\ldots,\kappa_u)\boldsymbol{\theta}^T$ where $\boldsymbol{\theta}$ is a $u\times u$ orthogonal matrix. With probability one, $\kappa_1,\ldots,\kappa_u$ are positive and distinct. Let $\mathbf{\Gamma}\boldsymbol{\theta}=(\boldsymbol{\tau}_1,\ldots,\boldsymbol{\tau}_u)$. Then

$$(\boldsymbol{\tau}_1,\ldots,\boldsymbol{\tau}_u)^T\hat{\mathbf{\Sigma}}(\boldsymbol{\tau}_1,\ldots,\boldsymbol{\tau}_u)=\text{diag}(\kappa_1,\ldots,\kappa_u).$$

Without loss of generality, assume $w=\kappa_u$ as κ_i are not ordered here.

The condition 4.11 is equivalent to

$$\hat{\mathbf{\Sigma}}(\boldsymbol{\tau}_1,\ldots,\boldsymbol{\tau}_u)\text{diag}(\frac{1}{\kappa_1}-\frac{1}{\kappa_u},\ldots,\frac{1}{\kappa_{u-1}}-\frac{1}{\kappa_u},0)=(\boldsymbol{\tau}_1,\ldots,\boldsymbol{\tau}_u)\text{diag}(1-\frac{\kappa_1}{\kappa_u},\ldots,1-\frac{\kappa_{u-1}}{\kappa_u},0),$$

subject to $\mathbf{\Gamma}^T\mathbf{\Gamma}-\mathbf{I}_u=0$. We have $\hat{\mathbf{\Sigma}}\boldsymbol{\tau}_i=\kappa_i\boldsymbol{\tau}_i$ for $i=1,\ldots,u-1$. In another word, $\boldsymbol{\tau}_i$ are eigenvectors of $\hat{\mathbf{\Sigma}}$ for $i=1,\ldots,u-1$. The formula (4.10) equals

$$\sum_1^u\log(\kappa_i)+(p-u)\log\{(\text{trace}(\hat{\mathbf{\Sigma}})-(\kappa_1+\ldots+\kappa_u))\} \tag{4.12}$$

$$\begin{aligned}=&\sum_1^{u-1}\log(\kappa_i)+\log(\kappa_u)+(p-u)\log\{(\text{trace}(\hat{\mathbf{\Sigma}})\\ -&(\kappa_1+\ldots+\kappa_{u-1}))-\kappa_u\}.\end{aligned} \tag{4.13}$$

Since

$$\begin{aligned}w&=\kappa_u=(\text{trace}(\hat{\mathbf{\Sigma}})-\text{trace}(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}))/(p-u)\\&=\text{trace}(\hat{\mathbf{\Sigma}})-(\kappa_1+\ldots+\kappa_u)/(p-u),\end{aligned}$$

we have $\kappa_u=(\text{trace}(\hat{\mathbf{\Sigma}})-(\kappa_1+\ldots+\kappa_{u-1}))/(p-u+1)$.

Fixing $\kappa_1,\ldots,\kappa_{u-1}$, by Lemma 6, κ_u reaches the maximum value for (4.13). Replacing κ_u with any other eigenvalues of $\hat{\mathbf{\Sigma}}$ that is different to $\kappa_1,\ldots,\kappa_{u-1}$ would make (4.13) smaller. This is to say, if w equals to one eigenvalue of $\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}$, $\hat{\mathbf{\Gamma}}$ can't reach the minimum of (4.10), i.e. the maximum of $L_3(\mathcal{S}_{\mathbf{\Gamma}})$. From the discussion, we can

conclude that the maximum likelihood estimator is one subset of u principal component directions.

Now assume $\kappa_1, \ldots, \kappa_u$ is one subset of u eigenvalues of $\hat{\mathbf{\Sigma}}$ that minimizes (4.12). Let $\kappa_{u+1}, \ldots, \kappa_p$ denote the complement of $\kappa_1, \ldots, \kappa_u$. Suppose there exists $\kappa_i < \kappa_l < \kappa_j$ where $1 \leq l \leq u$ and $u+1 \leq i, j \leq p$. Fixing $\kappa_1, \kappa_{l-1}, \kappa_{l+1} \ldots, \kappa_u$, by Lemma 6, the formula (4.12) can be reduced by replacing κ_l with either κ_i or κ_j. This tells us that $\kappa_{u+1}, \ldots, \kappa_p$ must form a "continuum block" of the eigenvalues. In other words, the maximum likelihood estimator $\hat{\mathcal{S}}_{\mathbf{\Gamma}}$ is the span of the first k and last $u-k$ principal component directions where k needs to be determined by the maximization of $L_3(\mathcal{S}_{\mathbf{\Gamma}})$ subject to $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_u$.

Proof of Proposition 10

Using the Lagrange multiplier rule, the solution $\hat{\mathbf{\Gamma}}$ can be gotten by finding the stationary points of the following unconstrained function:

$$-\frac{2}{n}L_5(\mathcal{S}_{\mathbf{\Gamma}}) + \text{trace}(\mathbf{U}(\mathbf{\Gamma}_0^T\mathbf{\Gamma}_0 - \mathbf{I}_u))$$

where $\mathbf{U}$ is a $(p-u)\times(p-u)$ matrix that stands for the Lagrange multipliers. Taking derivatives with respect to $\mathbf{\Gamma}_0$ and $\mathbf{U}$, we have the condition that the stationary points must satisfy

$$2\hat{\mathbf{\Sigma}}^{-1}\mathbf{\Gamma}_0(\mathbf{\Gamma}_0^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{\Gamma}_0)^{-1} + \frac{2(p-u-1)\hat{\mathbf{\Sigma}}\mathbf{\Gamma}_0\mathbf{Q}_1}{\text{trace}(\mathbf{\Gamma}_0^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}_0\mathbf{Q}_1)} + \frac{2\hat{\mathbf{\Sigma}}\mathbf{\Gamma}_0\mathbf{P}_1}{\text{trace}(\mathbf{\Gamma}_0^T\hat{\mathbf{\Sigma}}\mathbf{\Gamma}_0\mathbf{P}_1)} + \mathbf{\Gamma}_0(\mathbf{U}+\mathbf{U}^T) = 0$$

subject to $\mathbf{\Gamma}_0^T\mathbf{\Gamma}_0 - \mathbf{I}_{p-u} = 0$. It is straightforward to get the expression of $\mathbf{U}+\mathbf{U}^T$. Substituting the expression $\mathbf{U}+\mathbf{U}^T$ into the condition above and after simplification, we will finally find out that

$$\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{\Gamma}_0 = 0.$$

Then we can conclude that the maximum likelihood estimator $\hat{\mathcal{S}}_{\mathbf{\Gamma}}$ in $L_5(\mathcal{S}_{\mathbf{\Gamma}})$ is the span of one subset of u principal component directions.

References

[1] Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*. **34** 122-148.

[2] Brooks R.J. and Stone M. (1994). Joint continuum regression for multiple predictands. *J. Amer. Statist. Assoc.*. **89** 1374–1377.

[3] Chiaromonte, F., Cook, R. D. and Li, B. (2002). Sufficient dimension reduction in regressions with categorical predictors. *Ann. Statist.* **30** 475–97.

[4] Cook, R. D. and Weisberg, S. (1991). Discussion of Li (1991). *J. Amer. Statist. Assoc.* **86** 328–332.

[5] Cook, R. D. (1994). On the interpretation of regression plots. *J. Amer. Statist. Assoc.* **89** 177–190.

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

[7] Cook, R. D. (1998b). Principal Hessian directions revisited (with discussion). *J. Amer. Statist. Assoc.* **93** 84–100.

[8] Cook, R. D. and Li, B. (2002). Dimension reduction for the conditional mean in regression. *Ann. Statist.* **30** 455–474.

[9] Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *Ann. Statist.* **32** 1062–1092.

[10] Cook, R.D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *J. Amer. Statist. Assoc.*. **100** 410-428.

[11] Cook, R. D. (2007). Fisher lecture: dimension reduction in regression (with discussion). *Statist. Sci.* **22** 1–26.

[12] Cook, R. D. and Forzani, L. (2008). Principal fitted components for dimension reduction in regression. *Statist. Sci.* **23** 485–501.

[13] Cook, R. D. and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *J. Amer. Statist. Assoc.* **104** 197–208.

[14] Cook, R. D., Li, B. and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statist. Sinica* **20** 927–1010.

[15] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.

[16] Eaton, M. (1972). *Multivariate Statistical Analysis.* Institute of Mathematical Statistics. University of Copenhagen.

[17] Edelman, A., Arias, T. A. and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM J. Math. Anal.* **20** 303–353.

[18] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360.

[19] Fung, W. K., He, X., Liu, L. and Shi, P. (2002). Dimension reduction based on canonical correlation. *Statist. Sinica* **12** 1093–1113.

[20] Gohberg, I., Lancaster, P. and Rodman, L. (2006). *Invariant Subspaces of Matrices with Applications, 2nd edition.* SIAM.

[21] Helland I. S. (1988). On the structure of partial least squares regression. *Commun. Statist.*. **17** 581–607.

[22] Helland I. S. (1990). Partial least squares regression and statistical models. *Scand. J. Statist.*. **17** 97–114.

[23] Johnson, R. A.and Wichern, D. W. (2003). Applied Multivariate Statistical Analysis (5th Ed.). Pearson Education.

[24] Johnstone, I. M. and Lu, Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693.

[25] Jolliffe, I. (2002). *Principal Components Analysis.* New York: Springer, 2nd edition.

[26] Knyazev, A. V. and Argentati M. E. 2002, Principal Angles between Subspaces in an A-Based Scalar Product: Algorithms and Perturbation Estimates. *SIAM Journal on Scientific Computing*, **23**, no. 6, 2009-2041.

[27] Krzanowski, W. J. and Marriott, F. H. C. (1994). *Multivariate Analysis: Classification, Covariance Structures and Repeated Measurements.* London: Arnold.

[28] Leng, C. and Wang, H. (2009). On general adaptive sparse principal component analysis. *J. Comput. Graph. Statist.* **18** 201–215.

[29] Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc.* **102** 997–1008.

[30] Li, B., Zha, H. and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *Ann. Statist.* **33** 1580–1616.

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

[32] Li, L., Cook, R. D. and Nachtsheim, C. J. (2005). Model-free variable selection. *J. Roy. Statist. Soc. Ser. B* **67** 285–299.

[33] Li, L. and Nachtsheim, C. J. (2006). Sparse sliced inverse regression. *Technometrics* **48** 503–510.

[34] Li, L, and Yin, X. (2008). Sliced Inverse Regression with Regularization. *Biometrics.* **64** 124-131.

[35] Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86** 316–327.

[36] Li, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *J. Amer. Statist. Assoc.* **87** 1025–1039.

[37] Li, Y. and Zhu, L.-X. (2007). Asymptotics for sliced average variance estimation. *Ann. Statist.* **35** 41–69.

[38] Manton, J. H. (2002). Optimization algorithms exploiting unitary constraints. *IEEE Trans. Signal Process.* **50** 635–650.

[39] Marden, M. (1949). The geometry of the zeros of a polynomial in a complex variable. *Mathematical Surveys.* **No. 3** New York: American Mathematical Society.

[40] Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory.* New York: Wiley.

[41] Naik, P. and Tsai, C.L. (2000). Partial least squares estimator for single-index models. *J. Roy. Statist. Soc. Ser. B.* **62** 763–771.

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

[43] Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine* . **2** 559-572.

[44] Rapcsák, T. (1997). *Smoothed nonlinear optimization in* R^n. Boston: Kluwer Academic Publishers.

[45] Rapcsák, T. (2002). On minimization on Stiefel manifolds. *Eur. J. Oper. Res.* **143** 365–376.

[46] Shi, P. and Tsai, C.-L. (2002). Regression model selectiona residual likelihood approach. *J. Roy. Statist. Soc. Ser. B* **64** 237–252.

[47] Stone, M. and Brooks, R.J. (1990). Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *J. R. Statist. Soc. B.* **52** 237–269.

[48] Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *J. Roy. Statist. Soc. Ser. B.* **61** 611-622.

[49] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267-288.

[50] Wang, H., Li, R. and Tsai, C. L. (2007). On the consistency of SCAD tuning parameter selector. *Biometrika* **94** 553–568.

[51] Welling M., Agakov F. and Williams C.K.I.. Extreme components analysis. *In Neural Information Processing Systems*, **16**, Vancouver, Canada, 2003.

[52] Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*. **92** 937-950.

[53] Yin, X. and Cook, R. D. (2002). Dimension reduction for the conditional kth moment in regression. *J. Roy. Statist. Soc. Ser. B* **64** 159–75.

[54] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68** 49–67.

[55] Zhou, J. and He, X. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. *Ann. Statist.* **36** 1649–1668.

[56] Zhu, L.-X. and Ng, K. W. (1995). Asymptotics of sliced inverse regression. *Statist. Sinica* **5** 727–736.

[57] Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429.

[58] Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286.