2021

DIMENSION REDUCTION TECHNIQUES IN REGRESSION

Pei Wang
University of Kentucky, wangpeinihao@gmail.com
Digital Object Identifier: https://doi.org/10.13023/etd.2021.255

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Pei Wang, Student

Dr. Xiangrong Yin, Major Professor

Dr. Katherine Thompson, Director of Graduate Studies

DIMENSION REDUCTION TECHNIQUES IN REGRESSION

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Arts and Sciences
at the University of Kentucky

By
Pei Wang
Lexington, Kentucky

Co-Directors: Dr. Xiangrong Yin, Professor of Statistics
and Dr. Richard Kryscio, Professor of Statistics
Lexington, Kentucky
2021

ABSTRACT OF DISSERTATION

DIMENSION REDUCTION TECHNIQUES IN REGRESSION

Because of the advances of modern technology, the size of the collected data nowadays is larger and the structure is more complex. To deal with such kinds of data, sufficient dimension reduction (SDR) and reduced rank (RR) regression are two powerful tools. This dissertation focuses on these two tools and it is composed of three projects. In the first project, we introduce a new SDR method through a novel approach of feature filter to recover the central mean subspace exhaustively along with a method to determine the dimension, two variable selection methods, and extensions to multivariate response and large p small n scenarios. In the second project, we propose a novel SDR method by minimizing the distance between the population basis and the sample directions and provide a cross-validation method to determine dimension. In large p small n case, by adding a group lasso type penalty term to the objective function, simultaneous dimension reduction and variable selection are achieved. In the third project, we propose a new model by applying the RR idea to multinomial logistic regression (MLR) and combining RR-MLR with the first-order Markov Chain. Then, the model is applied to a dataset from a longitudinal study of aging and dementia.

KEYWORDS: Dimension Reduction, Reduced Rank, Variable Selection

Pei Wang

July 13, 2021

DIMENSION REDUCTION TECHNIQUES IN REGRESSION

By

Pei Wang

Dr. Xiangrong Yin

Co-Director of Dissertation

Dr. Richard Kryscio

Co-Director of Dissertation

Dr. Katherine Thompson

Director of Graduate Studies

July 13, 2021

Date

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my advisors Drs. Xiangrong Yin and Richard Kryscio for their strong support, nice encouragement and tremendous patience during the completion of this dissertation and my time at the University of Kentucky. Their enthusiasm, optimism, and confidence have set great examples for me to be a successful scientific researcher.

I would also like to thank Dr. Solomon Harrar, Dr. Arnold Stromberg, Dr. Katherine Thompson, and Dr. Chi Wang for serving on my dissertation committee. Their thoughtful comments and feedbacks make this dissertation better.

I would like to extend my gratitude to all faculty members in the Dr. Bing Zhang Department of Statistics at the University of Kentucky for their consistent support and encouragements through courses, seminars and discussions. I am also grateful to all my fellow graduate students for the time we spent together on discussing the course materials, preparing the exams and getting together to recharge ourselves.

Last but not least, I want to thank my family for their unconditional love and endless support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

1.1 Introduction

Due to the advances in computational procedures and the decreasing cost of collection, data is becoming more complex with high volume and dimension. Very large datasets are now routinely collected in biomedicine, gerontology, genetics, systems biology, finance, public health, and the social sciences. Most traditional statistical methods cannot be applied directly to analyze these data due to the complex structure or the high dimensionality. In the regression setting high dimensionality is encountered when the number of predictors is large compared to the number of observations and in some cases exceeds the number of observations. One particular issue is multicollinearity and to address this issue, several novel methods have been proposed including ridge regression [30], partial least squares regression [65] and least absolute shrinkage and selection operator (Lasso) [57]. To process large and complex data sets effectively, dimension reduction also draws significant attention from statisticians. In this dissertation, we investigate two dimension reduction techniques, sufficient dimension reduction (SDR) and reduced rank regression (RRR).

1.2 Sufficient Dimension Reduction

SDR refers to the application of dimension reduction while retaining all the relevant information in the regression model. In a classical regression setting with a response $Y \in \mathbb{R}$ and a predictor vector $\mathbf{X} \in \mathbb{R}^p$, we are looking for a $p \times d$ dimension reduction matrix $\mathbf{B}$ such that $Y \perp \mathbf{X}|\mathbf{B}^T\mathbf{X}$, where $d < p$ and $\perp$ means independence. That is, Y is independent of $\mathbf{X}$ given $\mathbf{B}^T\mathbf{X}$. The subspace spanned by the columns of $\mathbf{B}$ is called dimension reduction subspace[11][38](DRS). Neither $\mathbf{B}$ nor DRS is unique because the columns of multiple $\mathbf{B}$'s are possible to span the same subspace and any matrix comprised of $\mathbf{B}$ and any additional columns will make the independent condition holds. Thus, the primary interest is to find the central subspace (CS), $\mathcal{S}_{Y|\mathbf{X}}$,

which is defined as the intersection of all DRS with itself is a DRS [12]. The latter always exists and is unique under mild conditions[12][75].

Following CS, many novel specific subspaces have been proposed including central mean subspace[15], central moment subspace[72] and central variance subspace[79]. More generally, see Luo et al. [43] for a central T-subspace. One common example is that we sometimes are interested in the conditional mean function $E(Y|\mathbf{X})$, rather than the full conditional distribution of $Y|\ X$. Along this line, dimension reduction on the conditional mean $E(Y|\mathbf{X})$ is the focus. We call the associated subspace as central mean subspace (CMS). In referring the CMS, we are working on the mean function $E(Y|\mathbf{X})$ and seeking a matrix $\mathbf{B}$ such that $Y \perp E(Y|\mathbf{X})|\mathbf{B}^T\mathbf{X}$. Then the subspace spanned by the columns of such a $\mathbf{B}$ is called a dimension reduction mean subspace [15]. Similarly, the central mean subspace (CMS) is defined as the intersection of all dimension reduction mean subspaces if itself is a dimension reduction mean subspace and it is denoted as $\mathcal{S}_{E(Y|\mathbf{X})}$.

SDR serves as a pre-processor or an intermediate step in fitting models or analyzing data that greatly reduces the data from high dimensional to a relatively low one where the classical parametric and nonparametric modeling techniques can afterward be readily applied. Cook [12] pointed out that the dimension of the reduced predictor is often very small after pre-precessing the data with the SDR methods, usually 2-3 dimensions are enough.

Many SDR methods have been proposed since two pioneering methods, slice inverse regression (SIR) [38] and sliced average variance estimation (SAVE) [18]. These methods are roughly classified into three categories: Inverse methods such as parametric inverse regression (PIR)[7], directional regression (DR) [35] etc.; Forward methods such as minimum average variance estimation (MAVE) [67] and sliced regression (SR) [60]; and joint methods such as principal hessian direction (PHD) [39] and Kullback-Leibler distance [73]. Some of the proposed methods in the literature can accommodate multivariate response. See, Aragon [5], Cook and Setodji[17], Yin and Bura[71], and Li et al.[36]. Different to the eigen-decomposition technique in most of the SDR methods, Cook and Ni [16][44] estimated the target subspace by minimizing

a discrepancy function based on the inverse regression (IRE; RIRE). However, most of the methods in literature often involve kernel smoothing, information index optimization or tend to have strong assumptions on the distribution of predictors or on the link functions. The most common assumptions used in SDR are the linearity condition: $E(\mathbf{X}|\boldsymbol{\beta}^T\mathbf{X})$ is a linear function of $\boldsymbol{\beta}^T\mathbf{X}$ and the constant variance condition: $\text{var}(\mathbf{X}|\boldsymbol{\beta}^T\mathbf{X})$ is a nonrandom matrix, where $\boldsymbol{\beta}$ is the basis for the target subspace. These two conditions are quite strong and usually we have to assume the normality on $\mathbf{X}$ to satisfy the conditions. The linear and constant variance assumptions can be relaxed by using nonparametric methods (MAVE and Kullback Leibler in Table 1.1) but the computations involved become cumbersome. We summarize the above methods in Table 1.1.

Table 1.1: Summary of Existing Methods

Method	**Assumptions**	**Estimation Method**
SIR	$\text{span}(\Sigma^{-1}(E(\mathbf{X}\|Y) - E(\mathbf{X}))) \subseteq \mathcal{S}_{Y\|\mathbf{X}}$	
	Linearity	Eigen-decomposition
IRE/RIRE	$\text{span}(\Sigma^{-1}(E(\mathbf{X}\|Y) - E(\mathbf{X}))) \subseteq \mathcal{S}_{Y\|\mathbf{X}}$	
	Linearity	Quadratic function minimization
SAVE	$\text{span}(\Sigma - \text{var}(\mathbf{X}\|Y)) \subseteq \Sigma\mathcal{S}_{Y\|\mathbf{X}}$	
	Linearity & constant variance	Eigen-decomposition
PIR	$\text{span}(\Sigma^{-1}\text{cov}(\mathbf{X}, F(Y))) \subseteq \mathcal{S}_{Y\|\mathbf{X}}$	
	Linearity	Eigen-decomposition
DR	$\Sigma^{-1/2}\text{span}(2I_p - E(\mathbf{Z} - \tilde{\mathbf{Z}})(\mathbf{Z} - \tilde{\mathbf{Z}})^T\|Y, \tilde{Y}) \subseteq \mathcal{S}_{Y\|\mathbf{X}}$	
	Linearity & constant variance	Eigen-decomposition
PHD	$\text{span}(\Sigma^{-1}E((\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))^T(Y - E(Y))\Sigma^{-1}) \subseteq \mathcal{S}_{E(Y\|\mathbf{X})}$	
	Linearity & constant variance	Eigen-decomposition
MAVE	$L(a_1, \cdots, a_n, b_1, \cdots, b_n, B) = \sum_{j=1}^n \sum_{i=1}^n (Y_i - \{a_j + b_j^T\mathbf{B}^T(\mathbf{X}_i - \mathbf{X}_j)\})^2 w_{ij}$	
	/	Need nonparametric smoothing
Kullback Leibler	$\mathcal{I}(\boldsymbol{\beta}) = E(\log \frac{p(\boldsymbol{\beta}^T\mathbf{X}, Y)}{p(Y)p(\boldsymbol{\beta}^T\mathbf{X})})$	
	(Conditional) Normality	Nonparametric density estimation and successive optimization

Note, $\Sigma = \text{cov}(\mathbf{X})$; $F(Y)$ is formed by a set of functions of Y; operation $\tilde{}$ is used to denote the iid copy and $\mathbf{Z}$ is the standardized version of $\mathbf{X}$; $p(\cdot)$ is used to denote the density function.

All these difficulties lead us to develop new SDR methods with weak assumptions and cheap computational cost. Without using the linearity and constant variance conditions, we propose new SDR methods in this dissertation based on a assumption that $(\boldsymbol{\beta}, \boldsymbol{\beta}_0)$ forms an orthogonal matrix and $(Y, \boldsymbol{\beta}^T\mathbf{X})$ is independent of $\boldsymbol{\beta}_0^T\mathbf{X}$. Under

normality, this condition is satisfied and with $\boldsymbol{\beta}^T\mathbf{X} \perp \boldsymbol{\beta}_0^T\mathbf{X}$, $Y \perp \mathbf{X}|\boldsymbol{\beta}^T\mathbf{X}$ is equivalent to $(Y, \boldsymbol{\beta}^T\mathbf{X}) \perp \boldsymbol{\beta}_0^T\mathbf{X}$. However, the normality is not necessary. For instance, $\boldsymbol{\beta}^T\mathbf{X} \perp \boldsymbol{\beta}_0^T\mathbf{X}$ when $X_1 \perp (X_2, \cdots, X_p)$ and $\boldsymbol{\beta} = (1, 0, \cdots, 0)^T$ [78]. In addition, when p is reasonably large, the independence condition will hold approximately. Based on these points, the condition used in this dissertation is considered as weak. See Cook [13] and Sheng and Yin [54] for detailed discussion about this condition. Under this condition, the new SDR method ends up to calculate sample average and eigen-decomposition procedure which is computationally cheaper than the nonparametric kernel smoothing technique. See Section 2 for details.

SDR projects the original predictors to a lower dimension subspace and the reduced predictors still contain all original predictors. This issue hampers the interpretation of the results. To overcome this problem, sufficient variable selection (SVS), closely related to SDR, is proposed. SVS refers to selecting a subset of predictors that contains all the regression information. Similar to SDR, the goal is to find a $p \times k$ matrix $\mathbf{A}$ such that $Y \perp \mathbf{X}|\mathbf{A}^T\mathbf{X}$, where $k < p$ and the columns of $\mathbf{A}$ consist of unit vectors, $\mathbf{e}_i$, whose ith element is 1 [74]. Li [40] propose to combine the SDR method with a penalty to produce a sparse estimation to overcome this issue. See also, Chen et al. [9]. Different from the penalization methods, screening approach aims to select the important variables only. See Fan and Lv [23], Huang et al. [33], Fan et al. [24] and Yang et al. [68]. In this dissertation, we combine the SDR procedure with a penalty function to get the sparse matrix $\mathbf{A}$.

1.3 Reduced Rank Regression

RRR is another example of dimension reduction [4] and it is widely used in multivariate regression. Consider the regression of $k \times 1$ response $\mathbf{Y}$ on $(q + p) \times 1$ covariates $(\mathbf{W}^T, \mathbf{Z}^T)^T$, where $\mathbf{W}, \mathbf{Z}$ are $q \times 1$, $p \times 1$ vector respectively. We can write the model as $\mathbf{Y}_i = \mathbf{W}_i^T\mathbf{C}_1 + \mathbf{Z}_i^T\mathbf{C}_2 + \boldsymbol{\epsilon}_i$. In RRR, we introduce a rank constraint on the coefficient matrix $\mathbf{C}_2$. Instead of directly estimating the original $p \times k$ coefficients matrix, we estimate two lower-rank coefficient matrices, $p \times T$ matrix $\mathbf{A}$ and $T \times k$ matrix $\mathbf{G}$, with $T \leq \min\{p, k\}$; and then take their product, $\mathbf{C}_2 = \mathbf{AG}$. Aside from

the multivariate regression analysis in the ordinary least square setting, the reduced rank idea is commonly used in other regression methods. For example, Anderson first applied the reduced rank idea to multinomial logistic regression (MLR) and named it as stereotype model [3]. This stereotype model is limited to one dimension, and Goodman [28] as well as Yee and Hastie [70] extended it to multidimensional setting. Fiocco et al. applied the reduced rank idea to multi-state models in the proportional hazards model [25] and applied the proposed model to analyze some survival data.

To describe the transitions among multiple states in longitudinal studies, multi state models are useful. When successive observations are equally spaced in time, multi state Markov chains are frequently operationalized via MLR models to govern the transitions within each row of the process one-step transition matrix ("P matrix")[1][10][58]. This modeling facilitates the study of risk factors associated with transitions among states. However, because multinomial models require the estimation of many unknown parameters, model fitting can become cumbersome whenever the number of states or the number of potential transitions among states or the number of risk factors under consideration increases. This is often compounded when certain transitions are rare in comparison to others. Thus it is necessary to reduce the number of parameters that must be estimated and it motivates us to combine the idea of reduced rank MLR to a Markov chain.

1.4 Overview of the Dissertation

In this dissertation, we first propose two novel SDR methods using characteristic function along with their sparse solutions and then a new reduced rank model by applying the reduced rank MLR to the one step transition matrix in a markov chain.

The dissertation is organized as follows.

In Chapter 2, we introduce a new computationally efficient SDR method based on weak conditions and it is through a novel approach of feature filter. We first characterize a series of directions belong to a target subspace; then combine them into a candidate matrix; and extract the directions by applying eigen-decomposition to the candidate matrix. The new method has higher computational efficiency compared

to most of the forward methods which typically use nonparametric estimation. In addition, the proposed method works well for both univariate and multivariate responses. We further provide a permutation test to estimate the structural dimension. To select the informative predictors, we develop a sparse SDR estimator by building on the work of Li [40] and Wu and Yin [66], and Chen et al. [9]. To deal with the large p small n data, we propose two methods by adopting the two-stage selection procedure of Yang et al.[68] and using the sequential approach of Yin and Hilafu[74]. In this project, we not only do extensive numerical analysis but also prove several theoretical properties.

In Chapter 3, we extend the idea in Chapter 2 and develop another new SDR method. We define a vector seed via characteristic function that generates a series of vectors that belong to a certain subspace as in Chapter 2. It is natural to estimate the population subspace spanned by the vector seeds with a d dimensional subspace that is closest to the generated vectors, where d is the structure dimension. Motivated by Cook and Ni [16], we take the quadratic discrepancy function to denote closeness between the vectors and the population basis, and derive the estimator by minimizing this quadratic function. This method has more flexibility since we can take advantage of the inner product matrix and gain more information. In this project, we take two special inner product matrices, the identity matrix and Kronecker product between the identity matrix and the covariance matrix of $\mathbf{X}$. The two estimators from these two special quadratic discrepancy functions yield higher accuracy compared to other classical methods. In addition, we adopt the cross-validation method to test the dimensionality. The first project can be treated as a special case since eigen-decomposition problem can always be reformulated to the ordinary least square minimization problem, which also minimizes a quadratic function. In a high dimension dataset, it may be the case where only some of the predictors are significant for regression. Towards this end, motivated by Qian et al. [46], we develop a new method that works for SDR and SVS simultaneously. The key idea for this method is to add a coordinate-independent penalty to the quadratic function. To minimize the panelized quadratic function, we adopt the Coordinate descent algorithm and

Stiefel manifold optimization as in Qian et al.[46]. This strategy produces two sparse estimators with high accuracy in both direction estimation and variable selection. In addition, the theoretical properties about subspace estimation and variable selection consistency are proved.

In Chapter 4, we propose a new regression model by applying dimension reduction technique (reduced rank; RR) to MLR and combining RR-MLR with the first order Markov Chain. This novel model could highly reduce the number of parameters to be estimated and helps with inference for rare transitions in the chain. The model is applied to a dataset from a longitudinal study of aging and dementia by analyzing Apolipoprotein-E (APOE) gene $\epsilon 4$ allele(s) (i.e., carrying at least one $\epsilon 4$, APOE4) as a risk factor for transitions among cognitive states after adjustment for the presence of eight other covariates.

Chapter 2 Feature Filter for Estimating Central Mean Subspace and Its Sparse Solution

2.1 Introduction

SDR has been widely used in analyzing high-dimensional data. To overcome the drawbacks among the existing methods, it is necessary to develop new methods. In this chapter, we proposed a new SDR method by the technique of feature filter with the target on CMS. This method is based on a weak assumption and the computation cost is cheap. Chapter 2 is organized as follows. Section 2.2 introduces the methodology of our method along with its motivation, theoretical results, estimation algorithm and testing procedure. Section 2.3 illustrates two methods to get the sparse solution (SVS). Section 2.4 shows two approaches to deal with the large p small n problem. Section 3.4 reports numerical results. Section 2.6 concludes this chapter with a brief discussion. Proofs, long derivations and additional numerical results are provided in the appendix.

2.2 The Proposed Method

Let $Y \in \mathbb{R}$ and $\mathbf{X} \in \mathbb{R}^p$, with $\mathbf{\Sigma}_x = \text{cov}(\mathbf{X})$. Suppose that $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_d)$ is a $p \times d$ basis matrix of $\mathcal{S}_{E(Y|\mathbf{X})}$ with $d < p$. Then, $m(\mathbf{X}) = E(Y|\mathbf{X}) = E(Y|\boldsymbol{\beta}^T\mathbf{X}) = m(\boldsymbol{\beta}^T\mathbf{X})$. To facilitate our development, let f be a generic density and $\mathbf{i}^2 = -1$. Define $\psi(\omega) = \int e^{\mathbf{i}\omega^T\mathbf{X}} d[m(\mathbf{X})f(\boldsymbol{\beta}^T\mathbf{X})]$ with $\omega \in \mathbb{R}^p$. Then, we have the following result, whose proof is delayed in the supplement (A1.2).

Lemma 2.2.1 The CMS is exhaustively recovered by the collection of all of $\psi(\omega)$ with $\omega \in \mathbb{R}^p$. That is, $\mathcal{S}_{E(Y|\mathbf{X})} = \text{span}\{\psi(\omega) : \omega \in \mathbb{R}^p\}$.

Assume that density functions $f(\mathbf{X})$, $f(\boldsymbol{\beta}^T\mathbf{X})$ exist and $f(\boldsymbol{\beta}^T\mathbf{X}) \to 0$ as $||\mathbf{X}|| \to \infty$. Then, by simple algebra, we have $\psi(\omega) = -E(\mathbf{i}\omega e^{\mathbf{i}\omega^T\mathbf{X}} Y \frac{f(\boldsymbol{\beta}^T\mathbf{X})}{f(\mathbf{X})})$. Write $\psi(\omega) = \mathbf{a}_\omega + \mathbf{i}\mathbf{b}_\omega$, then $\mathcal{S}_{E(Y|\mathbf{X})} = \text{span}\{\mathbf{a}_\omega, \mathbf{b}_\omega : \omega \in \mathbb{R}^p\}$ [82]. However, using the form of

$\psi(\omega) = -E(\mathbf{i}\omega e^{\mathbf{i}\omega^T\mathbf{X}}Y\frac{f(\boldsymbol{\beta}^T\mathbf{X})}{f(\mathbf{X})})$ involves either estimating the respective densities on $\mathbf{X}$, or assuming a specific distribution of $\mathbf{X}$. To avoid such difficulties, we will use a simplified version: $E(\omega e^{\mathbf{i}\omega^T\mathbf{X}}Y)$, which is achieved under a simple condition as shown in Lemma 2.2.2. On the other hand, this simplified version is also motivated by the Martingale difference divergence (MDD) [51], which is defined as the nonnegative number that satisfies

$$\mathrm{MDD}(Y|\mathbf{X})^2 = \frac{1}{c_p}\int_{\mathbb{R}^p}\frac{|E(e^{\mathbf{i}\omega^T\mathbf{X}}Y) - E(e^{\mathbf{i}\omega^T\mathbf{X}})E(Y)|^2}{|\omega|_p^{1+p}}d\omega.$$

where, $c_p = \frac{\pi^{(1+p)/2}}{\Gamma((1+p)/2)}$ and $\mathrm{MDD}(Y|\mathbf{X})^2 = 0$ iff $E(Y|\mathbf{X}) = E(Y)$. This motivates us to search directions that maximize $\mathrm{MDD}(Y|\boldsymbol{\eta}^T\mathbf{X})^2$ over $\boldsymbol{\eta}$ for CMS. Lemma 2.2.2 shows a condition under which the two motivations are indeed supporting the same idea, whose proof is given in the supplement (A1.2).

Lemma 2.2.2 Assume $(\boldsymbol{\beta}, \boldsymbol{\beta}_0)$ forms an orthogonal matrix and $(Y, \boldsymbol{\beta}^T\mathbf{X})$ is independent of $\boldsymbol{\beta}_0^T\mathbf{X}$, then we have the following two results.

1. $\psi(\omega)$ reduces to $E(\omega e^{\mathbf{i}\omega^T\mathbf{X}}Y)k_0 \propto E(\omega e^{\mathbf{i}\omega^T\mathbf{X}}Y)$, where k_0 is a constant of $\boldsymbol{\beta}$ and the exact form is given in the supplement (A1.2).

2. Let $\boldsymbol{\eta}$ be any $p \times d_2$ matrix and $\boldsymbol{\eta}^T\boldsymbol{\eta} = I_{d_2}$. d_2 can be less, larger or equal to d. Assume $\mathcal{S}(\boldsymbol{\eta}) \nsubseteq \mathcal{S}(\boldsymbol{\beta})$, then $\mathrm{MDD}(Y|\boldsymbol{\eta}^T\mathbf{X})^2 < \mathrm{MDD}(Y|\boldsymbol{\beta}^T\mathbf{X})^2$.

The independence condition that $(Y, \boldsymbol{\beta}^T\mathbf{X})$ is independent of $\boldsymbol{\beta}_0^T\mathbf{X}$ is weaker than the common assumptions in the current literatures [13]. Part 1 of Lemma 2.2.2 indicates that each direction $E(\omega e^{\mathbf{i}\omega^T\mathbf{X}}Y)$ can be used to estimate $\mathcal{S}_{E(Y|\mathbf{X})}$; and Part 2 of Lemma 2.2.2 means that the maximum of $\mathrm{MDD}(Y|\boldsymbol{\eta}^T\mathbf{X})^2$ is achieved when $\mathbf{X}$ is projected onto the subspace that is spanned by the columns of $\boldsymbol{\beta}$. To see why they both support the same idea, we further suppose $E(Y) = 0, \mathrm{var}(Y) = 1$, then we can see that for the purpose of estimating the CMS, part 1 is going to use each individual direction ω, while part 2 means that we will use the direction that maximizes the "average" of these directions. Thus, in this paper, we propose to combine these two ideas to use the directions of ω whose "value" in the sense of correlation of

$E(e^{\mathbf{i}\omega^T\mathbf{X}}Y)$, is big. Then we take the "average" of them. Compare this idea to the approach of Zhu and Zeng [82], we avoid assumption on the multivariate normal distribution or nonparametric estimation on the density of $\mathbf{X}$. Compare this idea to the maximization on MDD, we avoid an optimization problem with a nonlinear constraint. Because these informative directions are selected by the large value of $E(e^{\mathbf{i}\omega^T\mathbf{X}}Y)$, this part is named as feature filter.

To implement our method, next we use two ways to formulate a dimension reduction candidate matrix: One is based on a discrete kernel and the other is based on a continuous kernel.

Euler Approach

Using the Euler formula, $e^{\mathbf{i}\omega^T\mathbf{X}} = \cos(\omega^T\mathbf{X}) + \mathbf{i}\sin(\omega^T\mathbf{X})$, we have

$$E(\omega e^{\mathbf{i}\omega^T\mathbf{X}}Y) = \omega E[\cos(\omega^T\mathbf{X})Y] + \mathbf{i}\omega E[\sin(\omega^T\mathbf{X})Y] = \mathbf{b}_\omega + \mathbf{i}\mathbf{a}_\omega.$$

We can then form a dimension reduction candidate matrix, $\mathbf{M}_e = \mathbf{C}\mathbf{C}^T$, where $\mathbf{C} = (\mathbf{a}_1, \mathbf{b}_1, \cdots, \mathbf{a}_m, \mathbf{b}_m)$. Thus we call such a method Euler Approach. Theoretically we should take all $\omega \in \mathbb{R}^p$. However, this is impossible practically. Fortunately, a finite large number of ωs will be enough to estimate the CMS [76]. Assume that m ωs are given, and $\{(Y_j,\ \mathbf{X}_j^T),\ j = 1, \cdots, n\}$ is a random sample of $(Y, \mathbf{X})$. An estimation of $\mathbf{M}_e$ is a $p \times p$ matrix,

$$\hat{\mathbf{M}}_e = \hat{\mathbf{C}}\hat{\mathbf{C}}^T, \tag{2.1}$$

where $\hat{\mathbf{C}} = (\hat{\mathbf{a}}_1, \hat{\mathbf{b}}_1, \cdots, \hat{\mathbf{a}}_m, \hat{\mathbf{b}}_m)$ is a $p \times 2m$ matrix, $\hat{\mathbf{a}}_k = \omega_k \frac{1}{n}\sum_{j=1}^n \sin(\omega_k^T\mathbf{X}_j)Y_j$, $\hat{\mathbf{b}}_k = \omega_k \frac{1}{n}\sum_{j=1}^n \cos(\omega_k^T\mathbf{X}_j)Y_j$ and $k = 1, \cdots, m$. In theory, every p-dimensional vector ω can serve to recover the CMS. However, in practice, an accurate CMS estimation is related to how ω is generated. Empirically, we follow what is proposed by Zhu et al. [80] to generate ωs: $\omega \overset{iid}{\sim} N(0, \sigma^2\mathbf{I}_p)$ with $\sigma^2 = s\pi^2/E(\mathbf{X}^T\mathbf{X})$ and $s = 0.02$. Furthermore, we normalize each ω to have scale free on ω by taking $\omega = \frac{\omega}{\sqrt{\omega^T\omega}}$. Finally, we only retain the ωs, whose absolute correlations between Y and $\cos(\omega^T\mathbf{X})$, Y and $\sin(\omega^T\mathbf{X})$ are among the peak values. Assume that d is known, estimation for d will be provided later. Now we are ready to describe our detailed algorithm.

Algorithm of Euler Approach

Step 1: Let $Y_i^* = Y_i - \frac{\sum_{i=1}^n Y_i}{n}$. Generate m ωs and normalize each one.

Step 2: Let $\rho_k = \max\{|\rho(Y^*, \cos(\omega_k^T \mathbf{X}))|, |\rho(Y^*, \sin(\omega_k^T \mathbf{X}))|\}$, $k = 1, \ldots, m$, where ρ is the correlation coefficient. Order these m correlations descending and choose τ percent of these m ωs with the largest ρ_k.

Step 3: Using the chosen ωs to construct the matrix $\hat{\mathbf{M}}_e$ in equation (2.1), then perform eigen-decomposition.

Step 4: The matrix formed by the first d eigenvectors, $\hat{\boldsymbol{\beta}} = [\hat{\mathbf{v}}_1, \cdots, \hat{\mathbf{v}}_d]$, corresponding to the largest d eigenvalues is our estimate.

In our limited study, using $\tau = 0.2$ and $m = 20000$ achieved consistent and stable results. Supplement (A2.4) provides additional simulations on various m and τ.

Theorem 2.2.1 Assume $\text{var}(Y \sin(\omega^T \mathbf{X}))$, $\text{var}(Y \cos(\omega^T \mathbf{X}))$ exist and $(Y_j, \mathbf{X}_j^T)$ is a random sample, then $\sqrt{n}(\text{vec}(\hat{\mathbf{M}}_e) - \text{vec}(\mathbf{M}_e)) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_1)$. The exact form of $\boldsymbol{\Sigma}_1$ and the proof of this theorem are given in the supplement (A1.2).

Corollary 2.2.1 Given $\text{vec}(\hat{\mathbf{M}}_e) \to \text{vec}(\mathbf{M}_e)$ at $\sqrt{n}$ rate, we have $\hat{\mathbf{M}}_e \to \mathbf{M}_e$ at the same rate. Then, the eigenvalues and eigenvectors of $\hat{\mathbf{M}}_e$ converge to those of $\mathbf{M}_e$ at the same rate [81].

Theorem 2.2.2 Let $2m > d$ and $p > d$. Then, the asymptotic distribution of $\hat{\Lambda}_d$ is the same as the distribution of

$$T = \sum_{j=1}^{(p-d)(2m-d)} \delta_j T_j,$$

where $\hat{\Lambda}_d = n \sum_{j=d+1}^{p} \hat{\lambda}_j$ is a test statistic for testing the structure dimension [38], T_j's are independently distributed with $T_j \sim \chi_1^2$ and $\delta_1, \delta_2, \cdots, \delta_{(p-d)(2m-d)}$ are the ordered singular values from $\boldsymbol{\Omega}_T$ whose exact form is given in the supplement (A1.2).

The justification of this theorem is included in the supplement (A1.2).

Theorem 2.2.2 enables us to determine dimension via sequential hypothesis tests. We describe the brief procedure below and refer the details to Li[38] and Bura and Cook[8]. In theorem 2, we replace all the unknown quantities by the corresponding consistent estimates. Once $\hat{\mathbf{\Omega}}_T$ is given, the asymptotic distribution of $\hat{\Lambda}_d$ is estimated by

$$\hat{T} = \sum_{j=1}^{(p-d)(2m-d)} \hat{\delta}_j T_j,$$

where $\hat{\delta}_1, \hat{\delta}_2, \cdots, \hat{\delta}_{(p-d)(2m-d)}$ are the ordered singular values from $\hat{\mathbf{\Omega}}_T$.

The sequential test procedure determines the dimension with the hypotheses form $d = k$ versus $d > k$ [8][38]. We start with $k = 0$, and compare $\hat{\Lambda}_k$ to a certain quantile of distribution of $\hat{T}$. If the test statistic is small, which corresponds to a large p-value, then there is no statistical evidence to reject the null hypothesis, we conclude that $d = k$. If the test statistic is large and it means the p-value is small, we conclude that $d > k$. Every large test statistic will result in 1 increment in k. This procedure stops until it encounters a small test statistic.

Such a sequential procedure should work in theory, but it is not effective in practice. We will use an alternative test later.

Note that our approach can be extended to multivariate response straightforwardly based on Cook and Setodji [17]. Let $\mathbf{Y} \in \mathbb{R}^q$, then $\mathcal{S}_{E(\mathbf{Y}|\mathbf{X})} = \sum_{i=1}^{q} \mathcal{S}_{E(Y_i|\mathbf{X})}$ [17], where Y_i is an element of $\mathbf{Y}$. Define $m(\mathbf{X}) = E(\mathbf{Y}^T|\mathbf{X})$, then all the theoretical results hold. And in Step 2 of our algorithm, by letting $\rho_k = \max\{|\rho(Y_1^*, \cos(\omega_k^T\mathbf{X}))|, |\rho(Y_1^*, \sin(\omega_k^T\mathbf{X}))|, \cdots, |\rho(Y_q^*, \cos(\omega_k^T\mathbf{X}))|, |\rho(Y_q^*, \sin(\omega_k^T\mathbf{X}))|\}$, the modified algorithm will work for multivariate response. This is different from selecting ω by maximizing the multivariate version of $\text{MDD}(\mathbf{Y}|\mathbf{X})^2$ [45].

Kernel Approach

Using a kernel, $K(\omega)$, we can form a candidate matrix as

$$\mathbf{M}_k = Re(\int \psi(\omega)\bar{\psi}(\omega)^T K(\omega)d\omega) = \int [\mathbf{a}_\omega \mathbf{a}_\omega^T + \mathbf{b}_\omega \mathbf{b}_\omega^T] K(\omega)d\omega.$$

If we choose the Gaussian function $K(\omega) = (2\pi\sigma_\omega^2)^{-p/2}\exp(-\frac{\| \omega \|^2}{2\sigma_\omega^2})$[82], and let $(Y_1, \mathbf{X}_1)$, $(Y_2, \mathbf{X}_2)$ be iid copies of $(Y, \mathbf{X})$, then $\mathbf{M}_k$ is proportional to $E_{(\mathbf{X}_1,Y_1),(\mathbf{X}_2,Y_2)}\mathbf{J}$, where

$$\mathbf{J} = Y_1 Y_2 e^{-\frac{\sigma_\omega^2 \| \mathbf{X}_1 - \mathbf{X}_2 \|^2}{2}} \sigma_\omega^2[\mathbf{I}_p - \sigma_\omega^2(\mathbf{X}_1 - \mathbf{X}_2)(\mathbf{X}_1 - \mathbf{X}_2)^T]. \tag{2.2}$$

Derivation of equation (2.2) is given in the supplement (A1.1). We call this method Kernel Approach. Let $\{(Y_j,\ \mathbf{X}_j^T),\ j = 1, \cdots, n\}$ be a sample of $(Y, \mathbf{X})$, then an estimate of $\mathbf{M}_k$ is

$$\hat{\mathbf{M}}_k = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} Y_i Y_j e^{-\frac{\sigma_\omega^2 \| \mathbf{X}_i - \mathbf{X}_j \|^2}{2}} \sigma_\omega^2[\mathbf{I}_p - \sigma_\omega^2(\mathbf{X}_i - \mathbf{X}_j)(\mathbf{X}_i - \mathbf{X}_j)^T]. \tag{2.3}$$

We have the following algorithm for Kernel Approach.

Algorithm of Kernel Approach

Step 1: Let $Y_i^* = Y_i - \frac{\sum_{i=1}^n Y_i}{n}$.

Step 2: Use equation (2.3) to construct an estimated candidate matrix $\hat{\mathbf{M}}_k$.

Step 3: Perform eigen-decomposition on $\hat{\mathbf{M}}_k$ and let $\hat{\boldsymbol{\beta}} = [\hat{\mathbf{v}}_1, \cdots, \hat{\mathbf{v}}_d]$, where $\hat{\mathbf{v}}_i$s are the first d eigenvectors associated with the largest d eigenvalues.

Step 4: Output $\hat{\boldsymbol{\beta}}$.

Practically, we use $\sigma_\omega^2 = 0.1$ as suggested by Zhu and Zeng [82], which works quite well. We have the following theorem for the Kernel Approach.

Theorem 2.2.3 Assume $\mathrm{cov}(\mathrm{vec}(\mathrm{J}((\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2))))$ exists. Then,

$$\hat{\mathbf{M}}_k = \mathbf{M}_k + \frac{1}{n}\sum_{i=1}^{n}(\mathrm{J}'(\mathbf{X}_i, Y_i) - 2\mathbf{M}_k) + o_p(n^{-1/2}),\ \text{as } n \to \infty,$$

where $\mathrm{J}'(\mathbf{X}, Y) = E_{(\mathbf{X}_2,Y_2)}[\mathrm{J}((\mathbf{X}, Y), (\mathbf{X}_2, Y_2)) + \mathrm{J}^T((\mathbf{X}, Y), (\mathbf{X}_2, Y_2))]$. Let $\boldsymbol{\Sigma}_2$ be the $p^2 \times p^2$ covariance matrix of $\mathrm{vec}(\mathrm{J}'(\mathbf{X}, Y))$, then $\sqrt{n}(\mathrm{vec}(\hat{\mathbf{M}}_k) - \mathrm{vec}(\mathbf{M}_k)) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_2)$.

Note that Theorem 2.2.3 is a simpler case of Theorem 1 in Zhu and Zeng [82] and our proof of theorem 2.2.3 is exactly the same as the proof of theorem 1 in Zhu and Zeng [82]. Thus, we omit it.

Corollary 2.2.2 Given $\text{vec}(\hat{\mathbf{M}}_k) \to \text{vec}(\mathbf{M}_k)$ at $\sqrt{n}$ rate, we have $\hat{\mathbf{M}}_k \to \mathbf{M}_k$ at the same rate. Then, the eigenvalues and eigenvectors of $\hat{\mathbf{M}}_k$ also converge to those of $\mathbf{M}_k$ at the same rate [81].

Kernel Approach and its algorithm work for multivariate response. See comments in supplement (A1.1).

Permutation Test

Previously, we assume that d is known, however, practically, we need to estimate it. Previous experience and our simulations indicate that the sequential hypotheses test is not effective. Therefore we adopt a permutation test. Permutation test is based on the eigenvalues and eigenvectors of the candidate matrix and thus it works for both Euler Approach and Kernel Approach. The idea behind the permutation test is that for a given real dimension d, at the population level, the first d eigenvalues of the candidate matrix are nonzero and the remaining eigenvalues are exact zero; At the sample level, the first d eigenvalues are much larger than the remaining which are close to zero. We refer more details to Cook and Yin[19], Yin and Cook [72] and Wang et al.[62].

Consider the hypothesis test $d = k$ vs. $d > k$, with $k \in \{0, 1, \cdots, p-2\}$. The range of d is from 0 to $p-1$ under the assumption that the dimension can be reduced at least by 1. Let $\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \cdots, \hat{\mathbf{v}}_p$ be the orthonormal eigenvectors corresponding to the ordered eigenvalues of candidate matrix $\hat{\mathbf{M}}$. Denote $\mathbf{A}_k = (\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \cdots, \hat{\mathbf{v}}_k)$ and $\mathbf{B}_k = (\hat{\mathbf{v}}_{k+1}, \hat{\mathbf{v}}_{k+2}, \cdots, \hat{\mathbf{v}}_p)$. Then, the null hypothesis is equivalent to $Y \perp E(Y|\mathbf{X})|\mathbf{A}_k^T\mathbf{X}$. Based on this idea, Wang et al. [62] proposed a test statistic

$$\hat{\Lambda}_k = \hat{\lambda}_{(k+1)} - \frac{1}{p-(k+1)} \sum_{i=k+2}^{p} \hat{\lambda}_i.$$

The algorithm to estimate d is given below.

Step 1: Obtain the matrix $\hat{\mathbf{M}}$ from data set by the algorithms in Section 2.2.

Step 2: Compute $\hat{\Lambda}_k$.

Step 3: Obtain $(\mathbf{X}\mathbf{A}_k, \mathbf{X}\mathbf{B}_k, Y)$, apply L independent permutations to the rows of $\mathbf{X}\mathbf{B}_k$. For each permutation, compute the test statistic $\hat{\Lambda}_{kl} = \hat{\lambda}_{(k+1)l} - \frac{1}{p-(k+1)} \sum_{i=k+2}^{p} \hat{\lambda}_{il}$, where $l = 1, 2, \cdots, L$ and $\hat{\lambda}_{il}$ is the ith eigenvalue of $\hat{\mathbf{M}}_l$, the candidate matrix from the lth permutation.

Step 4: Construct the p-value $p_k = \frac{\sum_{l=1}^{L} I(\hat{\Lambda}_{kl} > \hat{\Lambda}_k)}{L}$, where I is the indicator function. Given a preset significance level α, reject the null if $p_k < \alpha$.

Step 5: Repeat steps 2-4 for $k = 0, \cdots, p-2$ until the null hypothesis cannot be rejected and choose $\hat{d} = k$ as the estimated dimension.

The setting of $\alpha = 0.05$ and $L = 100$ works well in our limited study [20]. We use $m = 1000$ and $\tau = 0.2$ for Euler Approach in our simulations. A smaller m drastically reduced computational cost while keep its accuracy, see the supplementary file (A2.4) for additional simulations. Note that other estimation methods for d may be used here, such as elbow plot of eigenvalues, variation plot of Ye and Weiss [69] and ladle plot of Luo and Li [42].

2.3 Sufficient Variable Selection

It may still be difficult to interpret the reduced variables by SDR, as such variables contain all the predictors while perhaps, only a few x_i's are informative. Thus, the goal in this section is to select these informative x_i's by combining SDR with penalization. We will investigate two approaches: directly incorporate the method of Wu and Yin [66] which is modified from Li [40]; and embed the idea of Chen et al. [9]. In both of these two SVS approaches next, we fix $\tau = .2$ and $m = 1000$ to reduce the computational cost for Euler Approach, while keep its accuracy. See supplementary file (A2.4) for additional simulations.

Adaptive Lasso Type of SVS

Li [40] develop a general sparse estimator for eigen-decompostion approach, which is further modified by Wu and Yin [66] who used adaptive lasso [83]. Wu and Yin [66] further used $\mathbf{M}_\delta = \mathbf{M} + \delta\mathbf{\Sigma}$ instead of the candidate matrix $\mathbf{M}$, where δ is a small positive constant to improve Li's method [40], by reducing computation cost while keeping its accuracy. The choice of δ has little effect [66]. We propose our algorithm and present the details below.

Wu and Yin (2015) worked on the optimization problem

$$\underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{min}\{\sum_{i=1}^{p} \parallel \mathbf{\Sigma}^{-1}\mathbf{m}_i - \boldsymbol{\alpha}\boldsymbol{\beta}^T\mathbf{m}_i \parallel_{\mathbf{\Sigma}}^2 + \sum_{j=1}^{d}\lambda_j\sum_{h=1}^{p}\frac{|\beta_{jh}|}{w_h}\},$$

subject to $\boldsymbol{\alpha}^T\mathbf{\Sigma}\boldsymbol{\alpha} = \mathbf{I}_d$, where $\mathbf{\Sigma} = \text{cov}(\mathbf{X})$, $\mathbf{m}_i, i = 1, 2, \cdots, p$, are the columns of $\mathbf{M}^{1/2}$ and $\mathbf{w} = (w_1, \cdots, w_p)^T$ is a known weight vector. In our algorithm, this weight vector is chosen as the absolute OLS estimates.

The algorithm of adaptive lasso type SVS is summarized as follows.

Step 1: Calculate the candidate matrix $\hat{\mathbf{M}}$ and $\hat{\mathbf{M}}_\delta$ for a given $\delta = 0.001$.

Step 2: Get the initial α from the candidate matrix without the penalty term.

Step 3: Fixed $\boldsymbol{\alpha}$, updated $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_d)$ by replacing $\hat{\beta}_j$ with

$$\beta_j^* = \underset{\beta_j}{min}\{\parallel \hat{\mathbf{M}}_\delta^{1/2}\alpha_j - \hat{\mathbf{M}}_\delta^{1/2}\beta_j \parallel^2 + \lambda_j\sum_{h=1}^{p}\frac{|\beta_{jh}|}{w_h}\},$$

where w_i is the absolute OLS estimate of β_j for $j = 1, \cdots, p$.

Step 4: Fixed $\hat{\boldsymbol{\beta}}$, by $\mathbf{\Sigma}^{-1/2}\hat{\mathbf{M}}_\delta\hat{\boldsymbol{\beta}} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, we have $\boldsymbol{\alpha} = \mathbf{\Sigma}^{-1/2}\mathbf{U}\mathbf{V}^T$.

Step 5: Repeat steps 3 and 4 until the maximum absolute value of difference between two consecutive $\hat{\boldsymbol{\beta}}$'s $< 10^{-3}$.

We use the following BIC criteria [40][59] to select the tuning parameter λ_j.

$$\min_{\lambda}\{\sum_{i=1}^{p}||\mathbf{\Sigma}^{-1}\mathbf{m}_i - \hat{\boldsymbol{\beta}}_\lambda\hat{\boldsymbol{\beta}}_\lambda^T\mathbf{m}_i||_{\mathbf{\Sigma}}^2 + \log(n) * p_\lambda/n\},$$

where $\hat{\boldsymbol{\beta}}_\lambda$ is an estimated direction for a given λ and p_λ is the effective number of parameters which is estimated by the number of nonzero components in $\hat{\boldsymbol{\beta}}_\lambda$.

Group Lasso Type of SVS

Previous section for the algorithm of SVS is an element-wise penalization. Chen et al. [9] pointed out that non-active variables should have zeros on the entire row of $\boldsymbol{\beta}$, and proposed a coordinate-independent penalty function based on the formulation of Li [40]. The main idea is to shrink $\boldsymbol{\beta}$ by rows instead of columns. Follow Chen et al. [9], we propose a group lasso type algorithm and present the details as below.

Chen et al. [9] worked on the optimization problem $min_{\boldsymbol{\beta}}\{-\text{tr}(\boldsymbol{\beta}^T\mathbf{M}\boldsymbol{\beta}) + \rho(\boldsymbol{\beta})\}$, subject to $\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta} = \mathbf{I}_d$, where $\rho(\boldsymbol{\beta}) = \Sigma_{i=1}^{p}\theta_i \parallel \beta_i^* \parallel_2$ is a coordinate-independent penalty function and the $d \times 1$ row vector β_i^* is the ith row of $\boldsymbol{\beta}$. When $d = 1$, this method is the same as Li's [40] proposal. Let $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\Gamma}}$ and use a local quadratic function[22] to approximate the penalty term, it ends up to an optimization function

$$\min_{\boldsymbol{\Gamma}}\{-tr(\boldsymbol{\Gamma}^T\mathbf{M}_1\boldsymbol{\Gamma}) + \frac{1}{2}\text{tr}(\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}^{-1/2}\mathbf{H}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Gamma})\}, \tag{2.4}$$

where $\mathbf{M}_1 = \boldsymbol{\Sigma}^{-1/2}\mathbf{M}\boldsymbol{\Sigma}^{-1/2}$ and $\mathbf{H} = \text{diag}(\frac{\theta_1}{\parallel \hat{\beta}_1^* \parallel_2}, \frac{\theta_2}{\parallel \hat{\beta}_2^* \parallel_2}, \cdots, \frac{\theta_p}{\parallel \hat{\beta}_p^* \parallel_2})$. Thus, $\boldsymbol{\Gamma}$ can be easily solved by eigen-decomposition on $\mathbf{M}_1 - \frac{1}{2}\boldsymbol{\Sigma}^{-1/2}\mathbf{H}\boldsymbol{\Sigma}^{-1/2}$. See Chen et al.[9] for more details.

The algorithm can be summarized as follows.

Step 1: Get the candidate matrix $\hat{\mathbf{M}}$ and find initial $\hat{\boldsymbol{\Gamma}}$ from $min_{\boldsymbol{\Gamma}}\{-tr(\boldsymbol{\Gamma}^T\hat{\mathbf{M}}_1\boldsymbol{\Gamma})\}$ and let $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\Gamma}}$.

Step 2: Update $\hat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{H}\hat{\boldsymbol{\Sigma}}^{-1/2}$ based on a given $\hat{\boldsymbol{\beta}}$.

Step 3: Update $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\beta}}$ by eigen-decomposition of matrix $\hat{\mathbf{M}}_1 - \frac{1}{2}\hat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{H}\hat{\boldsymbol{\Sigma}}^{-1/2}$ and $\hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\Gamma}}$ respectively.

Step 4: Repeat steps 2 to 3 until the angle between two consecutive $\hat{\boldsymbol{\beta}}$'s $< 10^{-6}$.

During the procedure, variable X_i will be removed if $\parallel \hat{\beta^*}_i \parallel_2 < 10^{-6}$. The hard threshold used here is different to the soft threshold in Lasso where x_i is removed for an exact zero. We use the BIC criteria of Chen et al. [9] to select the tuning parameter θ. The criteria has a form $-\text{tr}(\hat{\boldsymbol{\beta}}_\theta^T\mathbf{M}\hat{\boldsymbol{\beta}}_\theta) + \text{df}_\theta \log(n)/n$, where $\hat{\boldsymbol{\beta}}_\theta$ is the

estimated basis given θ, df_θ is estimated by $(p_\theta - d)d$ and p_θ is the number of nonzero rows of $\hat{\boldsymbol{\beta}}_\theta$.

2.4 Large p small n

In modern data analysis, one critical issue is to deal with the scenario where $p > n$. In this section, we embed the two-stage selection procedure [68], where a sparse model is assumed, and the sequential approach [74] into our proposed method. In both of these two approaches, we fix $\tau = .2$ and $m = 1000$ for Euler Approach to reduce the computational cost, while keep its accuracy. See supplementary file (A2.4) for additional simulations.

Variable Screening

Yang et al. [68] proposed a two-stage selection procedure with distance correlation [56] to screen the variables. We use their method first to select informative variables so that the number of active predictors is less than the number of sample size, then apply our methods to further reduce the data. Note that the two-stage selection procedure not only considers the relationship between the response and the predictors, but also considers the relationship among predictors. Thus, it achieves better selection results. The details of this selection procedure are referred to Yang et al. [68] and our algorithm is shown below.

Let $\hat{c}(\mathbf{U}, \mathbf{V})$ denote the sample distance correlation between random vectors $\mathbf{U}$ and $\mathbf{V}$. See Székely et al. [56] for details of distance correlation.

The algorithm for this two-stage variable screening can be summarized as follows.

Step 1: Calculate $\hat{c}_j = \hat{c}(Y, X_j)$, $j = 1, \cdots, p$, and select d_1 X_j's corresponding to the largest $\hat{c}_j$'s.

Step 2: Obtain the conditional set:

1. Slice Y into 2 non-overlapping slices;

2. Calculate $\hat{c}_j^* = \sum_{s=1}^{2} \hat{c}_{j,s}$, where $\hat{c}_{j,s} = \hat{c}((\mathbf{X}_{-j}, X_j)|Y = s)$, where $\mathbf{X}_{-j}$ means deleting the j th column from $\mathbf{X}$.

3. Take d_2 X_j's corresponding to the largest $\hat{c}_j^*$'s, but not selected in step 1.

Step 3: Union the X_j's from steps 1 and 2.

Where $p' = n/\log(n)$, $d_1 = 0.95p'$ and $d_2 = p' - d_1$ as suggested by Yang et al. [68] and they work quite well in our simulations.

Sequential Procedure

Yin and Hilafu [74] propose a sequential sufficient dimension reduction (SSDR), to deal with the ultrahigh dimensional dataset. The SSDR is based on the next proposition, which is adopted from the proposition 1 of Yin and Hilafu [74].

Proposition 2.4.1 Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be two random vectors, then either (a) or (b) will imply (c) below:

(a) $\mathbf{X}_1 \perp (\mathbf{X}_2, Y)|\mathbf{B}^T\mathbf{X}_1$;

(b) $\mathbf{X}_1 \perp \mathbf{X}_2|(\mathbf{B}^T\mathbf{X}_1, Y)$ and $\mathbf{X}_1 \perp Y|\mathbf{B}^T\mathbf{X}_1$;

(c) $\mathbf{X}_1 \perp Y|(\mathbf{B}^T\mathbf{X}_1, \mathbf{X}_2)$.

Based on (c), one can see that $p(Y|\mathbf{X}_1, \mathbf{X}_2) = p(Y|\mathbf{B}^T\mathbf{X}_1, \mathbf{X}_2)$. If the dimension of $\mathbf{B}^T\mathbf{X}_1$ is less than $\mathbf{X}_1$, then the goal of dimension reduction is achieved. Under the case of $p >> n$, we can partition $\mathbf{X}$ into $\mathbf{X}_1$ and $\mathbf{X}_2$ such that the dimension of $\mathbf{X}_1$ is less than n. After reducing $\mathbf{X}_1$ to $\mathbf{B}^T\mathbf{X}_1$, we consider $\mathbf{B}^T\mathbf{X}_1, \mathbf{X}_2$ as new predictors and partition the new predictors to have new $\mathbf{X}_1$ and $\mathbf{X}_2$. Repeat this until there is no further reduction can be achieved. The target here is to find a matrix $\mathbf{B}$ such that (c) is satisfied. In order to make (c) true, we can find $\mathbf{B}$ such that either (a) or (b) is true. Two paths are proposed corresponding to (a) and (b). It is path I if (a) is used, which may be better for quantitative response, and path II if (b) is used, which may be better for qualitative response. We use path I as the response is quantitative.

We describe the SSDR algorithm below.

Step 1: Rearrange the order of predictors by distance correlation of Székely et al. [56].

Step 2: Partition $\mathbf{X} \in \mathbb{R}^p$ to $\mathbf{X}_1 \in \mathbb{R}^{p_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{p_2}$ such that $p_1 < n$. Combine $\mathbf{X}_2$ and Y as a new response and let $\mathbf{X}_1$ be the predictors.

Step 3: Use the algorithm described in Section 2.2 to get the reduced variable $\mathbf{B}^T\mathbf{X}_1$.

Step 4: Let the new predictors be $\mathbf{X} = (\mathbf{B}^T\mathbf{X}_1, \mathbf{X}_2)$ and go back to step 1.

Step 5: Repeat steps 2-4 until no further reduction can be achieved.

We use $p_1 = 20$ in our simulations [74]. To obtain a sparse estimation, we incorporate a penalty term to obtain a sparse solution, $\mathbf{B}^s$. At each partition, we obtain a sparse basis. Thus, in the procedure, we follow above SSDR algorithm except in Step 3, where we replace $\mathbf{B}^T\mathbf{X}_1$ with $\mathbf{B}^{s^T}\mathbf{X}_1$. We call this approach as sequential sufficient variable selection (SSVS). See Yin and Hilafu [74] for details.

2.5 Numerical Study

This section will evaluate the efficacy of our methods via simulations and application to a real data example. Our code is written in **R** and is available at the github with link *https://github.com/wangpeinihao/code1*. The **R** code for adaptive lasso type sparse estimation follows from Wu and Yin [66]; the **R** code for CISE type sparse estimation is based on Chen et al.'s **Matlab** code [9]; the **R** code for sequential SVS is modified from Yin and Hilafu's code [74]. We run 100 replicates to report our results.

The estimation accuracy between the population $p \times d$ matrix $\boldsymbol{\beta}$ and the estimated matrix $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_d)$ is measured by some commonly used criteria. Assume that $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ have orthonormal columns. We use trace correlation, $r = \sqrt{\sum_{i=1}^{d} \rho_i^2/d}$ [69], to measure the similarity, where ρ_i^2's are the eigenvalues of $\hat{\boldsymbol{\beta}}^T\boldsymbol{\beta}\boldsymbol{\beta}^T\hat{\boldsymbol{\beta}}$. The larger the value is, the better the estimate is. Also, distances such as $\Delta_m(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}})$ defined as the spectral norm of matrix $\mathbf{A}$ and Frobenius norm $\Delta_f(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) = \sqrt{trace(\mathbf{A}\mathbf{A}^T)}$ [37]

are used, where $\mathbf{A} = \boldsymbol{\beta}\boldsymbol{\beta}^T - \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T$. Another distance is $m_i^2(\hat{\boldsymbol{\beta}}_i, \boldsymbol{\beta}) = |(I - \boldsymbol{\beta}\boldsymbol{\beta}^T)\hat{\boldsymbol{\beta}}_i|^2$ [67]. The smaller the distance is, the better the estimate is. Additionally, $|r_1|$, defined as the absolute correlation between the true sufficient predictors and its estimate [74] is used.

In sufficient variable selection, we report the true positive rate (TPR) and false positive rate (FPR), where TPR stands for the ratio of the number of correctly selected active predictors to the number of true active predictors and FPR stands for the ratio of the number of falsely selected active predictors to the number of true inactive predictors. Better estimate has bigger TPR and smaller FPR.

Simulations

We use the following models for our simulations and comparisons.

M1: $Y = (\mathbf{X}^T\beta)^3 + 0.4\epsilon$.

M1a: $Y = (\mathbf{X}^T\beta)^2 + 0.4\epsilon$.

M2: $Y = \cos(2\mathbf{X}^T\beta_1) - \cos(\mathbf{X}^T\beta_2) + 0.5\epsilon$.

M3: $Y_1 = 1 + (\mathbf{X}^T\beta_1)^2 + \epsilon_1$, $Y_2 = \mathbf{X}^T\beta_2 + \epsilon_2$, $Y_3 = \epsilon_3$, $Y_4 = \epsilon_4$.

M4: $X_Y = \sqrt{7/8}\beta Y + 0.5\epsilon$.

M5: $Y = \sqrt{7/8}\mathbf{X}^T\beta + 1.5\epsilon$.

The cubic Model M1 is modified from the quadratic model of Cook and Weisberg [18]; the quadratic model M1a is from Section 2 of Cook and Weisberg [18] and it has the same setting as M1; M2 is example 8.1 from Li [39]; M3 is a multivariate model 4.5 of Li et al. [36]; M4 is the model in Section 5.1 of Cook [14], and M5 is example 1 in the supplementary file of Yin and Hilafu[74]. All model settings are summarized in Table 2.1 below.

Example 2.5.1

This example uses M1, M1a and M2 to show the estimation accuracy of our methods in SDR for a univariate response with comparisons to SIR [38], SAVE [18], PHD [39] and DR [35]. The results are reported in Tables 2.2–2.4. The entries are the means of the criterion values and their respective standard errors (in parentheses). Since

Table 2.1: Model Settings

No.	Variables	$\boldsymbol{\beta}$	p	q	n	ϵ	d
M1	$\mathbf{X} \sim N(\mathbf{0}, \mathbf{I})$	$\mathbf{e}_1 + \mathbf{e}_2$	10	1	200	$N(0,1)$	1
M2	$\mathbf{X} \sim N(\mathbf{0}, \mathbf{I})$	$(\mathbf{e}_1\ \mathbf{e}_2)$	10	1	400	$N(0,1)$	2
M3	$\mathbf{X} \sim N(\mathbf{0}, \mathbf{I})$	$(\mathbf{e}_1\ 2\mathbf{e}_2 + \mathbf{e}_3)$	6	4	100/200	$N(\mathbf{0}, \Sigma_3)$	2
M4	$Y \sim N(0, 0.5)$	$\mathbf{e}_{100} + \mathbf{e}_{200} + \cdots + \mathbf{e}_{800}$	1000	1	200	$N(\mathbf{0}, \mathbf{I})$	1
M5	$\mathbf{X} \sim N(\mathbf{0}, \mathbf{I})$	$\mathbf{e}_{100} + \mathbf{e}_{200} + \cdots + \mathbf{e}_{800}$	1000	1	200	$N(0,1)$	1

Note: $\mathbf{e}_i$ is a $p \times 1$ vector with the ith element 1 and the other elements 0; q is the dimension of the response; Except in M4, $\mathbf{X}$ is independent of ϵ, while in M3, $\Sigma_3 = \begin{bmatrix} 1 & -0.5 & 0 & 0 \\ -0.5 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$.

MAVE is based on local polynomial approach which is different to other methods, we present the MAVE results in the supplementary file (A2.3).

Table 2.2 indicates that SIR has the largest correlation and smallest distance measures. It is expected since M1 has strong linear trend and SIR always performs the best for such a model. Kernel Approach is the second best, close to SIR, with the next two bests are Euler Approach and DR. Table 2.3 shows the results for M1a and it says our Euler and Kernel Approaches have the best results. SIR fails to work because of the symmetric pattern in this model. Table 2.4 reports the results for M2. In this model, Euler Approach is the best. Since there is a symmetric pattern in this model, SIR fails to recover the true direction as we expected. From these three models, we conclude that our methods are consistent and stable with the best performance, closely followed by DR and SAVE.

Table 2.2: SDR results of Model M1

method	r	Δ_m	Δ_f	m_1
SIR	**0.997** (0.002)	**0.075**(0.019)	**0.106** (0.027)	**0.075** (0.019)
SAVE	0.989 (0.012)	0.143 (0.055)	0.202 (0.078)	0.143 (0.055)
DR	0.993 (0.002)	0.113 (0.026)	0.160 (0.037)	0.113 (0.026)
PHD	0.652 (0.179)	0.722 (0.152)	1.021 (0.214)	0.722 (0.152)
Euler (Ours)	0.993 (0.007)	0.115 (0.045)	0.163 (0.065)	0.115 (0.045)
Kernel (Ours)	0.996 (0.002)	0.087 (0.022)	0.123 (0.031)	0.087 (0.022)

Example 2.5.2

This example uses modified M1a to show the efficacy of our methods in a case where the error term is not normally distributed. Here we use the same M1a model

Table 2.3: SDR results of Model M1a

method	r	Δ_m	Δ_f	m_1
SIR	0.293 (0.237)	0.917 (0.130)	1.298 (0.184)	0.917 (0.130)
SAVE	0.975 (0.011)	0.217 (0.052)	0.306 (0.073)	0.217 (0.052)
DR	0.976 (0.011)	0.211 (0.052)	0.298 (0.073)	0.211 (0.052)
PHD	0.966 (0.032)	0.248 (0.071)	0.350 (0.100)	0.248 (0.071)
Euler (Ours)	**0.978** (0.010)	**0.202** (0.049)	**0.286** (0.069)	**0.202** (0.049)
Kernel (Ours)	0.978 (0.010)	0.205 (0.049)	0.290 (0.070)	0.205 (0.049)

Table 2.4: SDR results of Model M2

method	r	Δ_m	Δ_f	m_1	m_2
SIR	0.436 (0.131)	0.973 (0.041)	1.776 (0.133)	0.887 (0.112)	0.880 (0.116)
SAVE	0.947 (0.029)	0.376 (0.090)	0.624 (0.146)	0.302 (0.093)	0.313 (0.090)
DR	0.949 (0.027)	0.367 (0.082)	0.615 (0.137)	0.292 (0.087)	0.314 (0.084)
PHD	0.969 (0.016)	0.293 (0.070)	0.483 (0.112)	0.236 (0.071)	0.240 (0.068)
Euler (Ours)	**0.971** (0.013)	**0.278** (0.064)	**0.464** (0.102)	**0.220** (0.063)	**0.235** (0.071)
Kernel (Ours)	0.933 (0.057)	0.435 (0.182)	0.669 (0.243)	0.335 (0.171)	0.307 (0.133)

except the error term ϵ has heavy tail, either t_5 or χ^2_9 distribution. The results are in Table 2.5. In both situations, our Euler Approach has the best performance as in the normal error term case and the Kernel Approach has a very competitive results. It means that our proposed method is robust against heavy tail cases.

Table 2.5: SDR results of M1a with a heavy tail error

error term	method	r	Δ_m	Δ_f	m_1
$\epsilon \sim t_5$	SIR	0.217 (0.187)	0.955 (0.074)	1.351 (0.105)	0.955 (0.074)
	SAVE	0.974 (0.014)	0.217 (0.056)	0.306 (0.080)	0.217 (0.056)
	DR	0.969 (0.017)	0.238 (0.063)	0.337 (0.089)	0.238(0.063)
	PHD	0.964 (0.019)	0.257 (0.067)	0.363 (0.095)	0.257 (0.067)
	Euler (Ours)	**0.977** (0.011)	**0.209** (0.049)	**0.296**(0.069)	**0.209**(0.049)
	Kernel (Ours)	0.966 (0.024)	0.246 (0.074)	0.348 (0.105)	0.246 (0.074)
$\epsilon \sim \chi^2_9$	SIR	0.272 (0.185)	0.941 (0.082)	1.331 (0.117)	0.941 (0.082)
	SAVE	0.959 (0.021)	0.273 (0.070)	0.387 (0.100)	0.273 (0.070)
	DR	0.949 (0.030)	0.301 (0.083)	0.426 (0.118)	0.301 (0.083)
	PHD	0.959 (0.020)	0.275 (0.067)	0.389 (0.095)	0.275 (0.067)
	Euler (Ours)	**0.971** (0.014)	**0.233** (0.056)	**0.329** (0.079)	**0.233**(0.056)
	Kernel (Ours)	0.946 (0.070)	0.296 (0.109)	0.418 (0.153)	0.296 (0.109)

Example 2.5.3

This example uses M3 to show the efficacy of our methods for multivariate response. Li et al. [36] proposed projective resampling (PR) idea and developed PR-SIR and PR-SAVE which outperformed K-means estimators [50] and central moment spaces

estimator [71] using Forbenius norm as the criterion. Under the same criterion, results in Table 2.6 indicate that Euler Approach is the best, Kernel Approach is the second best, both their performances are better than PR-SIR and PR-SAVE.

Table 2.6: SDR results (Δ_f) of Model M3

methods	PR-SIR	PR-SAVE	Euler (Ours)	Kernel (Ours)
n=100	1.205 (0.255)	0.612 (0.237)	0.353 (0.143)	0.527 (0.222)
n=200	1.203 (0.262)	0.313 (0.089)	0.241 (0.078)	0.309 (0.109)

Example 2.5.4

This example uses M1, M1a, M2 and M3 to show the performance of permutation test in estimating dimension. The sample sizes are 400, 800 and 1200 for the four models. The proportions of correct dimension determination for each model are reported in Table 2.7, which clearly indicate that the results are better with bigger sample size, and that it detects the dimension quite well, especially for Euler Approach and when $n \geq 800$.

Table 2.7: Dimension Test with correctly identified percentage

model	method	n=400	n=800	n=1200
M1	Euler	0.93	0.94	0.98
	Kernel	0.90	0.97	0.98
M1a	Euler	0.95	0.93	0.93
	Kernel	0.98	0.97	0.95
M2	Euler	0.80	0.93	0.91
	Kernel	0.20	0.89	0.97
M3	Euler	0.96	0.93	0.93
	Kernel	0.90	0.96	0.93

Example 2.5.5

This example uses M1 and M2 to show the results of variable selection by adaptive lasso method (ALasso, Section 3.1) and group Lasso as coordinator-independence sparse estimation (CISE, Section 3.2). For ALasso sparse estimation, we compare our results to SIR, SAVE, and PHD embedded with ALasso. Comparison to sparse MAVE [61], is also available in the supplementary file (A2.3). For CISE type, we

compare our results to SIR, SAVE and PHD embedded with CISE. Note that Chen et al. [9] only reported CISE with SIR, but we extend CISE to SAVE and PHD.

Table 2.8 reports the results using TPR and FPR. SIR, Euler Approach and Kernel Approach perform very well in M1, but SIR fails in M2. While PHD, SAVE, Euler Approach and Kernel Approach all perform well in M2, PHD fails in M1, and SAVE with ALasso fails in M1. Thus overall, Euler Approach and Kernel Approach are the most consistent methods with large TPR and small FPR.

Table 2.8: Variable Selection Results

Penalty	Model	d	Criteria	SIR	PHD	SAVE	Euler (Ours)	Kernel (Ours)
ALasso	M1	1	TPR	1.000	0.995	0.995	1.000	0.915
			FPR	0.001	0.844	0.568	0.001	0.000
	M1a	1	TPR	0.630	1.000	1.000	0.995	1.000
			FPR	0.035	0.859	0.618	0.063	0.026
	M2	2	TPR	0.390	1.000	1.000	1.000	0.990
			FPR	0.155	0.004	0.184	0.019	0.038
CISE	M1	1	TPR	1.000	0.885	0.980	1.000	0.975
			FPR	0.000	0.663	0.008	0.004	0.000
	M1a	1	TPR	0.170	1.000	0.995	1.000	0.895
			FPR	0.163	0.353	0.024	0.024	0.006
	M2	2	TPR	0.265	1.000	1.000	0.998	0.975
			FPR	0.211	0.013	0.019	0.021	0.013

Example 2.5.6

This example uses M4 to illustrate the estimation accuracy of our methods in ultra-high dimension settings. As we discussed before, we incorporate all the comparison methods with the variable screening procedure. The results are reported in Table 2.9, indicating that PHD and SAVE fail to capture the direction. Euler Approach is the best, DR and Kernel Approach are the closest second and third best. The results for MAVE with variable screening procedure is given in the supplementary file (A2.3).

Example 2.5.7

This example uses M5 to show the usefulness of our method with the SSVS approach in ultra-dimensional setting. We report Δ_f, absolute correlation $|r_1|$, FPR and TPR in Table 2.10. In terms of TPR, both of our methods are better than SIR sequential method, Euler Approach is the best.

Table 2.9: Estimation accuracy for model M4

	$\|r_1\|$	r	Δ_m	Δ_f	m_1
SIR	0.973 (0.006)	0.793 (0.042)	0.605 (0.054)	0.856 (0.077)	0.605 (0.054)
SAVE	0.036 (0.030)	0.039 (0.041)	0.999 (0.003)	1.412 (0.004)	0.999 (0.003)
DR	0.991 (0.003)	0.926 (0.018)	0.375 (0.042)	0.530 (0.060)	0.375 (0.042)
PHD	0.167 (0.108)	0.068 (0.051)	0.996 (0.006)	1.409 (0.008)	0.996 (0.006)
Euler (Ours)	**0.992** (0.002)	**0.936** (0.011)	**0.352** (0.029)	**0.498** (0.041)	**0.352** (0.029)
Kernel (Ours)	0.990 (0.003)	0.917 (0.014)	0.397 (0.033)	0.561 (0.046)	0.397 (0.033)

Table 2.10: Sequential SVS approaches

	Δ_f	$\|r_1\|$	TPR	FPR
SSVSSIR	0.602 (0.222)	0.905 (0.063)	0.888	0.005
SSVSEuler (Ours)	0.713 (0.168)	0.917 (0.034)	0.994	0.095
SSVSKernel (Ours)	0.646 (0.147)	0.886 (0.068)	0.970	0.063

Note: The results for SSVSSIR are copied from Yin and Hilafu [74].

Example 2.5.8

We use this example to compare our method to Zhu and Zeng's [82] Fourier transformation method. To make a fair comparison, the same evaluation criterion, $1 - \sqrt{\text{tr}(\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T)/d}$, from Zhu and Zeng's [82] paper is used, where $\mathbf{A}$, $\mathbf{B}$ are two $p \times d$ matrices. Boxplots are drawn based on the distances from 500 replicates. We run their method on **R** code of Zhu and Zeng[82]. The models used include the example 1 of Zhu and Zeng[82] and M2 in this chapter.
M6 (Example 1, Zhu and Zeng, 2006): $Y = (\mathbf{X}^T\beta_1)^2/(3 + (\mathbf{X}^T\beta_2 + 2)^2) + 0.2\epsilon$, $\beta_1 = (1, 1, 1, 1, 0, 0, 0, 0, 0, 0)^T$, $\beta_2 = (0, 0, 0, 0, 0, 0, 1, 1, 1, 1)^T$ and $\mathbf{X} = (X_1, \cdots, X_{10})^T$, ϵ are iid $N(0, 1)$. Sample size, $n = 500$.

The boxplot in Figure 2.1 reports the results from Model M6. We can see that Kernel Approach performs a slightly better than the FCN method and Euler Approach performs similar to FMN, where FCN stands for Fourier transform method of Zhu and Zeng[82] with target of CS and FMN stands for Fourier transform method of Zhu and Zeng[82] with target of CMS.

Figure 2.2 is the boxplot for Model M2 and it shows that Euler Approach is the best compared to other approaches. Kernel Approach has a similar performance compared to FMN which is definitely better than FCN.

These two simulations demonstrated that our approaches not only result in a

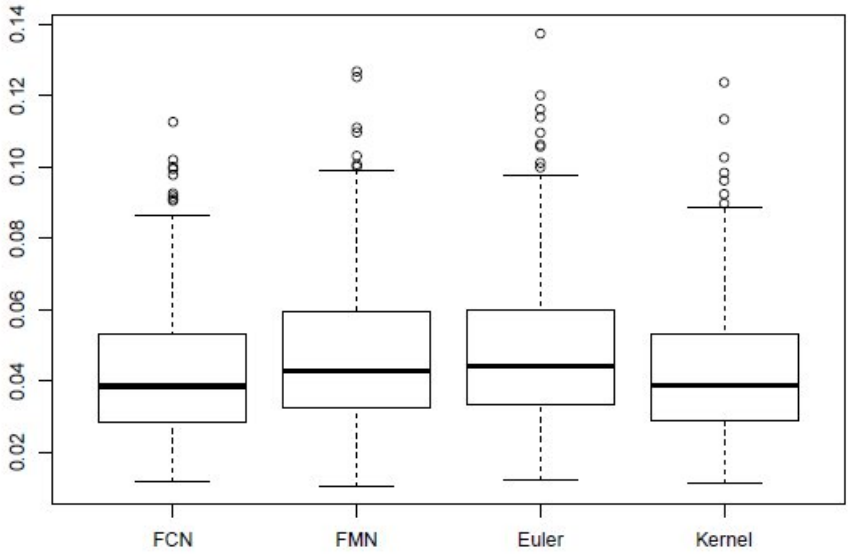


Figure 2.1: Side-by-Side boxplot of model M6 for comparing the performance for methods FCN, FMN, Euler and Kernel.

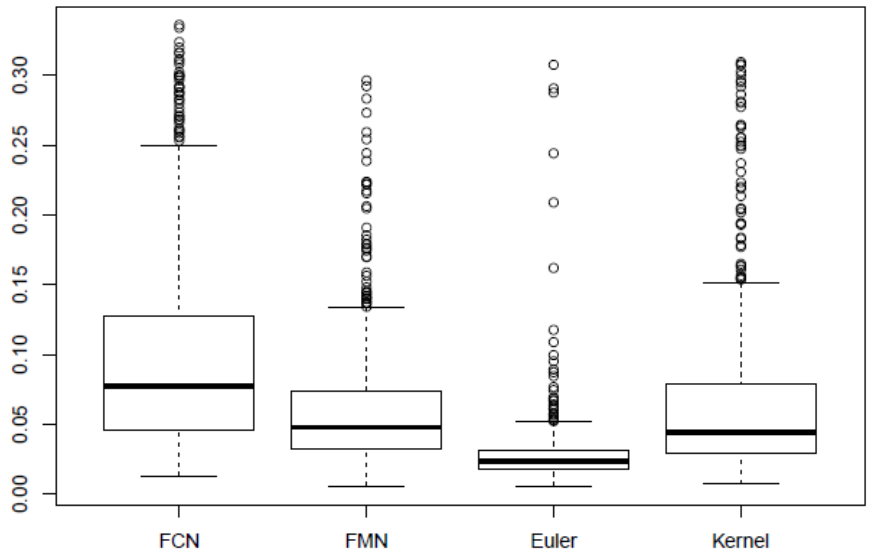


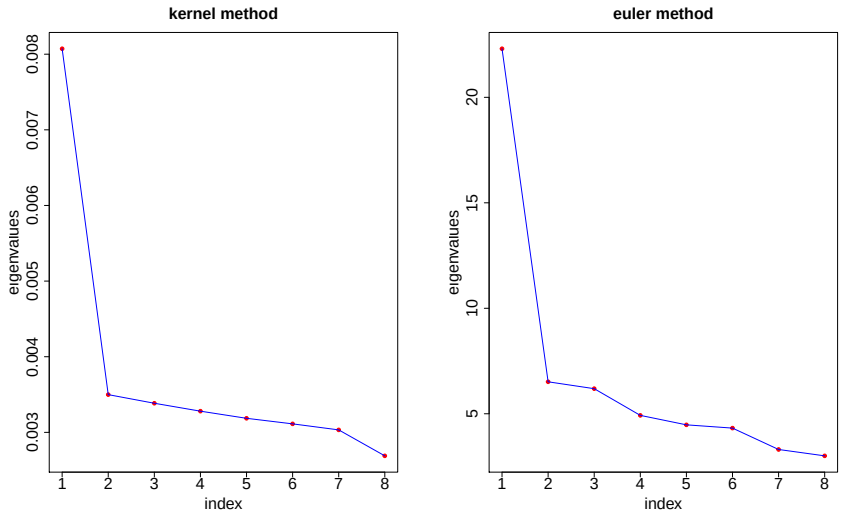
Figure 2.2: Side-by-Side boxplot of model M2 for comparing the performance for methods FCN, FMN, Euler and Kernel.

simpler form but also better results compared to the Fourier transform approaches of Zhu and Zeng[82].

To summarize, across these models, our proposed methods are the best or perform consistently in the top group.

Real Data Example

In this section, we apply our proposed approaches to Prostate Cancer Data, which is from **R** package *lasso2*: *//cran.r-project.org/web/packages/lasso2/lasso2.pdf*. The goal is to study the relation between the level of prostate specific antigen and a series of clinical measures in men who were about to have a radical prostatectomy. There are $n = 97$ cases and 9 variables. The 9 variables are lcavol (X_1; log(cancer volume)), lweight (X_2; log(prostate weight)), age(X_3), lbph (X_4; log(benign prostatic hyperplasia amount)), svi (X_5; seminal vesicle invasion), lcp (X_6; log(capsular penetration)), gleason (X_7; gleason score), pgg45(X_8; percentage gleason score 4 or 5) and lpsa (Y; log(prostate specific antigen)).

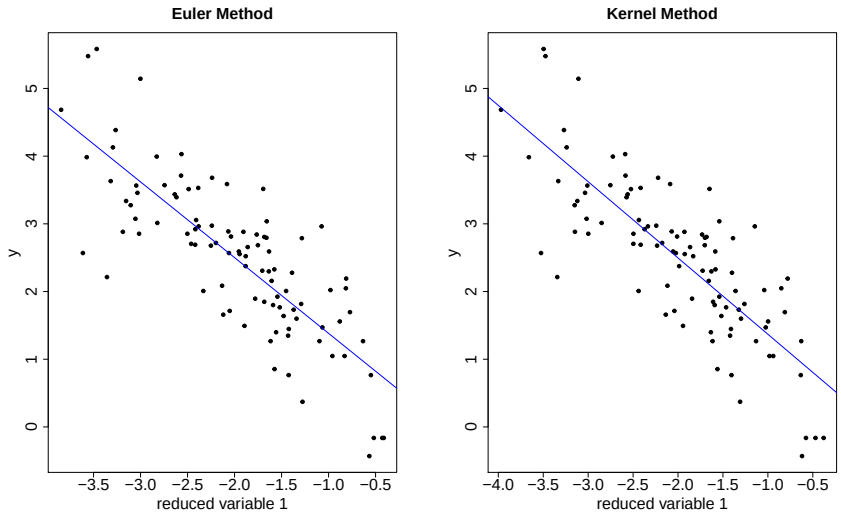


(a) (b)

Figure 2.3: Scree plots of two methods

Permutation test may not be effective because of small sample size ($n = 97$). Thus, we use the elbow plots to determine the dimension of the central mean subspace. The plots in figure 2.3, suggest $d = 1$ for both of our methods. The scatter plots between y

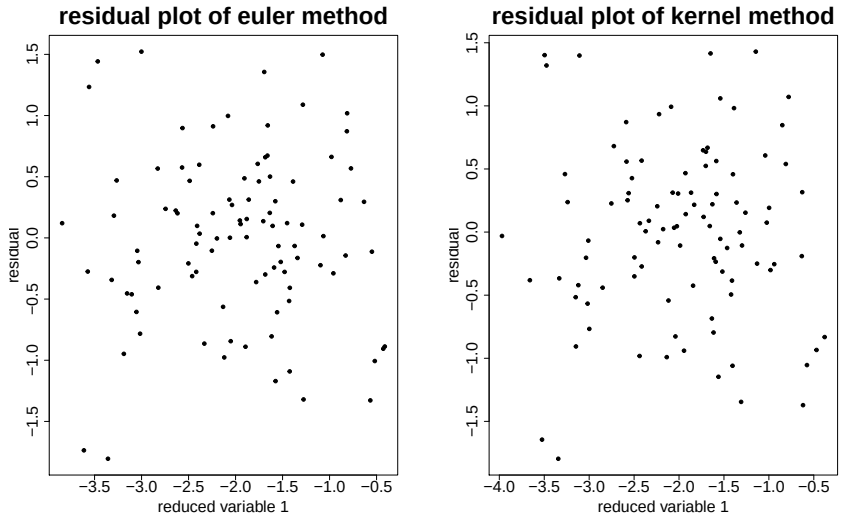
and the first reduced variables for both methods in Figure 2.4, clearly show the linear trend: plot (a) and plot (b) with added respective OLS fit. However, the patterns between the response and the respective second reduced variables are hard to tell (not shown here). Thus, we pursue further analysis using $d = 1$. The residual plots for the OLS fit respectively using the first reduced variable are reported in Figure 2.5. For both residual plots from the two approaches, we see the residual points are randomly distributed along the panels without any pattern. Therefore, we conclude that the mean model fits well based on the first respective reduced variable, i.e., it is reasonable to infer $d = 1$.



(a) (b)

Figure 2.4: Scatter plot between Y and the first respective reduced variables

To obtain the sparse estimates, adaptive lasso type method described in section 2.3 is used since $d = 1$. For both nonsparse and sparse estimates, we compare our results to SIR. The results are reported in Table 2.11, which show very similar results. The nonsparse estimates indicate that the direction is mainly determined by X_1, and X_5. Indeed, all three sparse estimates pick up X_1 and X_5, additionally, SIR picks up X_2 and Euler Approach picks up X_4. Due to the fact that our methods focus on the mean function, we further analyze with the first sparse reduced variables by Euler Approach and Kernel Approach.



(a) (b)

Figure 2.5: Residual plots for Euler and Kernel Approaches

Table 2.11: Estimated Directions for Prostate Dateset

	Predictors	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Nonparse	SIR	-0.4805	-0.3445	0.0132	-0.0539	-0.7895	0.1508	-0.0340	-0.0053
	Kernel	-0.4344	-0.4076	0.0115	-0.0798	-0.7958	0.0528	-0.0503	-0.0045
	Euler	-0.4251	-0.4646	0.0154	-0.0566	-0.7712	0.0326	-0.0638	-0.0036
Sparse	SIR	-0.986	-0.038	0	0	-0.160	0	0	0
	Kernel	-0.639	0	0	0	-0.769	0	0	0
	Euler	-0.612	0	0	-0.020	-0.790	0	0	0

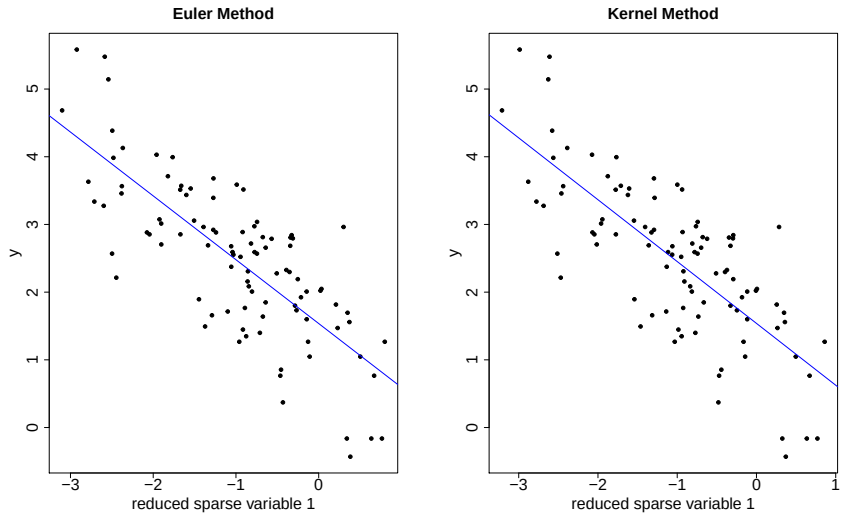
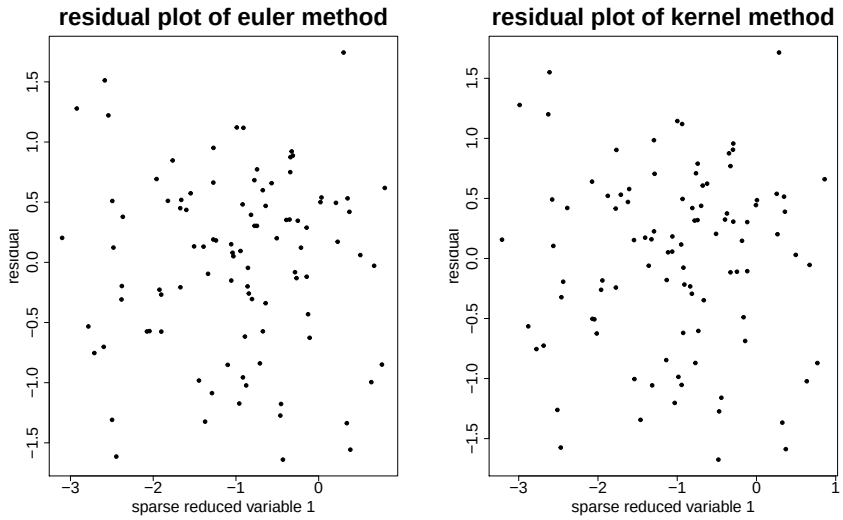


Figure 2.6: Scatter plots between Y and reduced sparse variables for sparse estimates

To verify our conclusion, we plot the first respective reduced sparse variables vs the response, with added OLS fit. Figure 2.6 shows a clear linear trend for both methods, which is similar to the nonsparse case. Thus, the sparse estimates are quite reasonable. The residual plots from the OLS by regressing the response on the first reduced sparse variables is reported in Figure 2.7, respectively. In both residual plots, there are no clear patterns and all the points are scattered around 0, indicating that the respective first reduced sparse variable is adequate for capturing the regression information.

We further analyze the importance of the extra variable X_4 that is selected by Euler Approach. The two reduced first sparse variables by Euler Approach and Kernel Approach have correlation coefficient 0.999. Thus with or without X_4, it is not affecting the reduced predictor. See figure 2.8.



(a) (b)

Figure 2.7: Residual plots of two methods

2.6 Discussion

In this project, we propose a new SDR method to recover the CMS by using characteristic function, together with the novel filtering idea. Our method avoids to use slicing technique and nonparametric estimation which overcomes the sensitivity

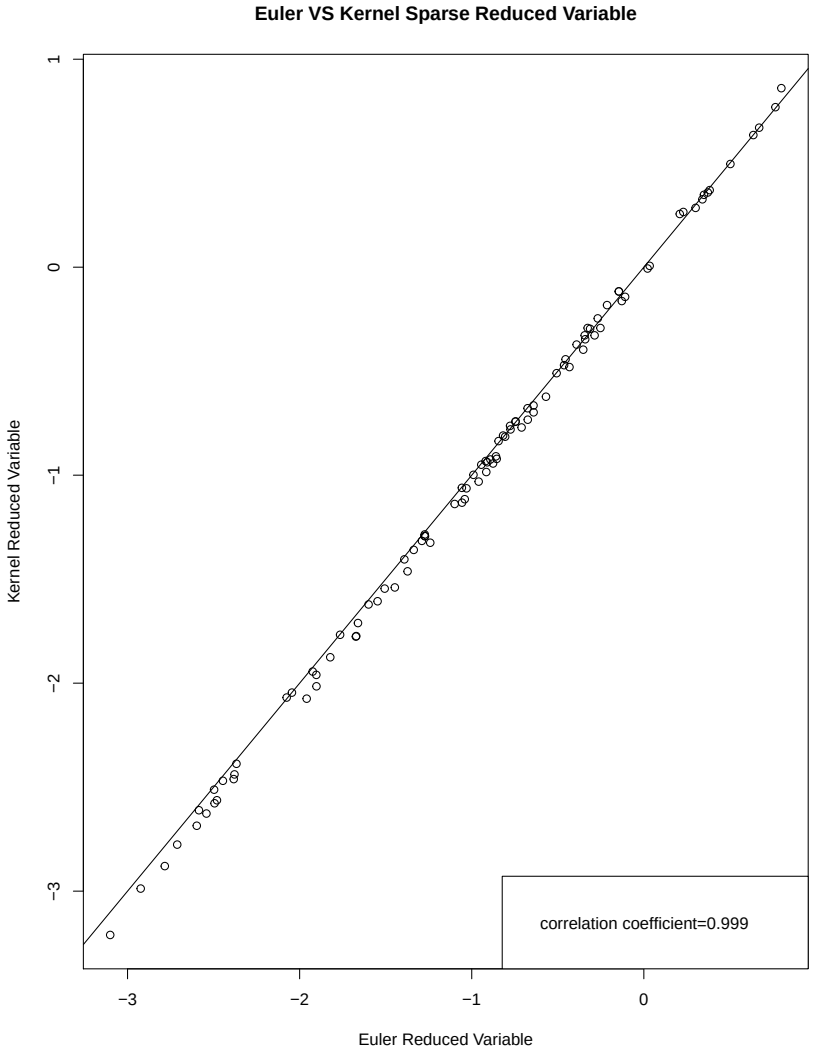


Figure 2.8: The plot of the first reduced sparse variables by Euler Approach and Kernel Approach.

of slicing number choice and computational complexity, respectively. Our method works well both in univariate and multivariate cases. We propose two choices of weighting schemes. Other ideas of different weights may be pursued in the future. For instance, after selecting important ωs, we can directly use ω as directions instead of $\omega E(Ye^{\mathbf{i}\omega^T\mathbf{X}})$, where $E(Ye^{\mathbf{i}\omega^T\mathbf{X}})$ serves as a weight.

The condition $(Y, \beta^T\mathbf{X})$ is independent of $\beta_0^T X$ is frequently used in sufficient dimension reduction area. It is equivalent to $Y \perp X|\beta^T\mathbf{X}$ under the assumption that $P_\beta\mathbf{X} \perp P_{\beta_0}\mathbf{X}$. Sheng and Yin [53] discussed that $P_\beta\mathbf{X} \perp P_{\beta_0}\mathbf{X}$ holds when $\mathbf{X}$ is multivariate normal, but the normality is not necessary. For arbitrary predictors, low dimensional projections of the predictor are approximately multivariate normal when p is large. Thus, this condition is not as strong as it appears to be. In practice, most of the predictors will satisfy the condition. See Section 3.5 of Sheng and Yin (2013) for more details. We include a non-normal predictor example in the supplementary file, see M9 for more details and leave the case where the independence condition is severely violated for future investigation.

Chapter 3 Minimum Discrepancy Approach for Sufficient Dimension Reduction Using Characteristic Function

3.1 Introduction

In Chapter 2, all vectors that belong to the target CMS are aggregated to form a candidate matrix and then by eigen-decomposition, we extract the eigen-vectors to estimate the CMS. Different to the traditional eigen-decomposition approach on a candidate matrix, Ni and Cook [16] proposed to minimize a quadratic function to estimate the target subspace. Assume the structure dimension d is known, the quadratic objective function is defined as

$$F_d(\mathbf{B}, \mathbf{C}) = (\text{vec}(\hat{\boldsymbol{\xi}}\mathbf{R}_n) - \text{vec}(\mathbf{BC}))^T \mathbf{V}_n (\text{vec}(\hat{\boldsymbol{\xi}}\mathbf{R}_n) - \text{vec}(\mathbf{BC})), \tag{3.1}$$

where $\hat{\boldsymbol{\xi}}$ is a $p \times m$ matrix, $\mathbf{V}_n \in \mathbb{R}^{pm \times pm}$ is a positive definite matrix, $\mathbf{B} \in \mathbb{R}^{p \times d}$ is a basis of span$\{\boldsymbol{\xi}\mathbf{R}_n\}$, $\boldsymbol{\xi}$ is the population version of $\hat{\boldsymbol{\xi}}$, $\mathbf{C} \in \mathbb{R}^{d \times m}$ is a matrix that is used for fitting, $\mathbf{R}_n \in \mathbb{R}^{m \times m}$ is matrix to organize the columns of $\hat{\boldsymbol{\xi}}$ and vec represents an operator that stacks the columns of a matrix to a single long vector. Then, by minimizing this objective function through iteration on $\mathbf{B}$ and $\mathbf{C}$, we can use the final $\mathbf{B}$ as an estimated basis for $\mathcal{S}_{y|\mathbf{x}}$. Since every nonsingular matrix $\mathbf{R}_n$ can be used to estimate the basis and it will not influence the estimation result, we can let $\mathbf{R}_n$ be the identity matrix for simplicity [16]. With different choices for the inner product matrix $\mathbf{V}_n$, we have more flexibility and gain more information compared to the eigen-decomposition type SDR methods. Motivated by this method, we can apply the formulation here to our proposed vector seed in Section 2.2.

The rest of this chapter is structured as follows. Section 3.2 introduces the new method for SDR along with its formulation, algorithm and the approach to determine the structure dimension. Section 3.3 presents the SVS formulation, algorithm and theoretical properties for high-dimensional data. We report the numerical results of the proposed methods in Section 3.4. This chapter ends with a discussion in Section 3.5.

3.2 The Proposed Method for SDR

For generality, let $\mathbf{x} = (x_1, \cdots, x_p)^T$ be a p dimensional vector, $y \in \mathbb{R}^1$ be a response and $\mathbf{\Sigma} > 0$ be the covariance matrix of $\mathbf{x}$. If $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_d)$ forms a basis for $\mathcal{S}_{\mathrm{E}(y|\mathbf{x})}$, then the mean function $m(\mathbf{x}) = m(\boldsymbol{\beta}^T\mathbf{x})$. Based on the fact that the gradient of $m(\mathbf{x})f(\boldsymbol{\beta}^T\mathbf{x})$ is a linear combination of $\boldsymbol{\beta}$ for any fixed $\mathbf{x}$ and thus it is in the $\mathcal{S}_{E(y|\mathbf{x})}$, we define a vector seed by taking a Fourier transformation of the gradient. That is, $\psi_0(\omega) = \int e^{\mathbf{i}\omega^T\mathbf{x}} d(m(\mathbf{x})f(\boldsymbol{\beta}^T\mathbf{x}))$, where $\omega \in \mathbb{R}^p$ is a constant vector. Wang et al.[63] proved that the $\mathcal{S}_{E(y|\mathbf{x})}$ is fully recovered by the collection of all the $\psi_0(\omega)$. We further assume that the densities $f(\mathbf{x})$, $f(\boldsymbol{\beta}^T\mathbf{x})$ exist, $f(\boldsymbol{\beta}^T\mathbf{x}) \to 0$ as $||\mathbf{x}|| \to \infty$, and $(\boldsymbol{\beta}, \boldsymbol{\beta}_0)$ form an orthogonal matrix with $(y, \boldsymbol{\beta}^T\mathbf{x})$ being independent to $\boldsymbol{\beta}_0^T\mathbf{x}$. We have $\psi_0(\omega) = k\mathrm{E}(\omega y e^{\mathbf{i}\omega^T\mathbf{x}}) \propto \mathrm{E}(\omega y e^{\mathbf{i}\omega^T\mathbf{x}})$, where k is a constant of $\boldsymbol{\beta}$. See more details in Wang et al.[63]. If $\psi(\omega) = \mathrm{E}(\omega y e^{\mathbf{i}\omega^T\mathbf{x}}) = E(\omega y \cos(\omega^T\mathbf{x})) + \mathbf{i}E(\omega y \sin(\omega^T\mathbf{x}))$, then the CMS is spanned by the collection of $\psi(\omega)$.

Motivated by Cook and Ni[16], we develop a quadratic discrepancy function in terms of our proposal. Before stating the objective function, we define the following: $\psi(\omega_i) = \boldsymbol{\beta}\zeta_i$, and $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_m) \in \mathbb{R}^{p\times 2m}$, where $\boldsymbol{\xi}_i = (\mathrm{Re}(\psi(\omega_i)), \mathrm{Im}(\psi(\omega_i)))$, $i = 1, \cdots, m$ and m is the total number of ω used. Then, we have $\boldsymbol{\xi} = \boldsymbol{\beta}\nu$, with $\nu = (\mathrm{Re}(\zeta(\omega_1)), \mathrm{Im}(\zeta(\omega_1)), \cdots, \mathrm{Re}(\zeta(\omega_m)), \mathrm{Im}(\zeta(\omega_m)))$ where operators $\mathrm{Re}(A)$ and $\mathrm{Im}(A)$ are the real and imaginary part of A respectively. Given a random sample $(\tilde{y}_j, \tilde{\mathbf{x}}_j)$, the sample version of $\boldsymbol{\xi}_i$ is defined as

$$\hat{\boldsymbol{\xi}}_i = (\frac{1}{n}\sum_{j=1}^{n}\omega_i\tilde{y}_j\cos(\omega_i^T\tilde{\mathbf{x}}_j), \frac{1}{n}\sum_{j=1}^{n}\omega_i\tilde{y}_j\sin(\omega_i^T\tilde{\mathbf{x}}_j)), i = 1, \cdots, m.$$

Theorem 3.2.1 Assume $\mathrm{var}(y\sin(\omega^T\mathbf{x}))$, $\mathrm{var}(y\cos(\omega^T\mathbf{x}))$ exist for every given ω, then we have $\sqrt{n}(\mathrm{vec}(\hat{\boldsymbol{\xi}}) - \mathrm{vec}(\boldsymbol{\beta}\nu)) \xrightarrow{d} N(\mathbf{0}, \mathbf{\Gamma})$, where $\mathbf{\Gamma}$ is a $2pm \times 2pm$ variance-covariance matrix with its exact form given in Supplement B1.1.

Then we define the following quadratic objective function,

$$F_d^{cf}(\mathbf{B}, \mathbf{C}) = (\mathrm{vec}(\hat{\boldsymbol{\xi}}) - \mathrm{vec}(\mathbf{BC}))^T\mathbf{V}_n(\mathrm{vec}(\hat{\boldsymbol{\xi}}) - \mathrm{vec}(\mathbf{BC})). \tag{3.2}$$

If m ωs are used to estimate the basis, then $\hat{\boldsymbol{\xi}}$ is a $p \times 2m$ matrix, $\mathbf{B}$ is a $p \times d$ matrix, $\mathbf{C}$ is a $d \times 2m$ matrix and $\mathbf{V}_n$ is a $2mp \times 2mp$ inner product matrix.

For the unknown matrix $\mathbf{V}_n$ in equation 3.2, Cook and Ni[16] point out that it is determined by the SDR method and different choices of $\mathbf{V}_n$ result in different SDR methods. In this project, we choose $\mathbf{V}_n = \mathbf{I}_{2m} \otimes \hat{\mathbf{\Sigma}}$ and $\mathbf{V}_n = \mathbf{I}_{2mp}$ to formulate two SDR methods.

Covariance Characteristic Function Estimator

In this section, we let $\mathbf{V}_n = \mathbf{I}_{2m} \otimes \hat{\mathbf{\Sigma}}$, and the objective function is

$$F_d^{ccf}(\mathbf{B}, \mathbf{C}) = (\text{vec}(\hat{\boldsymbol{\xi}}) - \text{vec}(\mathbf{BC}))^T (\mathbf{I}_{2m} \otimes \hat{\mathbf{\Sigma}})(\text{vec}(\hat{\boldsymbol{\xi}}) - \text{vec}(\mathbf{BC})). \tag{3.3}$$

If $(\hat{\boldsymbol{\beta}}, \hat{\nu}) = \text{argmin}_{\mathbf{B},\mathbf{C}} F_d^{ccf}(\mathbf{B}, \mathbf{C})$, then $\hat{\boldsymbol{\beta}}$ is the estimated basis of $\mathcal{S}_{E(y|\mathbf{x})}$. We call this estimator the covariance characteristic function estimator (ccfe) of $\boldsymbol{\beta}$.

Let $(\dfrac{\partial \text{vec}(\mathbf{BC})}{\partial \text{vec}(\mathbf{B})}, \dfrac{\partial \text{vec}(\mathbf{BC})}{\partial \text{vec}(\mathbf{C})})$ be the Jacobian matrix, and the value evaluated at $(\mathbf{B} = \boldsymbol{\beta}, \mathbf{C} = \nu)$ is $\Delta = (\nu^T \otimes \mathbf{I}_p, \mathbf{I}_{2m} \otimes \boldsymbol{\beta})$, a $2mp \times (2m + p)d$ matrix. Then the asymptotic properties of this estimation are established in theorem 3.2.2.

Theorem 3.2.2 Assume $\text{var}(y \sin(\omega^T \mathbf{x}))$ and $\text{var}(y \cos(\omega^T \mathbf{x}))$ exist for every given ω and a random sample $(\tilde{\mathbf{x}}, \tilde{y})$ is given. Let $\mathcal{S}_{\boldsymbol{\xi}} = \sum_{i=1}^m \boldsymbol{\xi}_i$ and $d = \dim(\mathcal{S}_{\boldsymbol{\xi}})$. Then, we have:

- $\sqrt{n}(\text{vec}(\hat{\boldsymbol{\beta}}\hat{\nu}) - \text{vec}(\boldsymbol{\beta}\nu)) \rightarrow N(\mathbf{0}, \Delta(\Delta^T \mathbf{V} \Delta)^- \Delta^T \mathbf{V} \Gamma \mathbf{V} \Delta (\Delta^T \mathbf{V} \Delta)^- \Delta^T)$, where $\mathbf{V} = \mathbf{I}_{2m} \otimes \mathbf{\Sigma}$;
- $n\hat{F}_d^{ccf}$ is asymptotically distributed as $\sum_{i=1}^{2mp} \lambda_i \chi_i^2(1)$, where $\hat{F}_d^{ccf}$ is the minimum value of F_d^{ccf}, $\chi_i^2(1)$s are independent chi-square random variable with degrees of freedom 1 and the weights λ_is are the eigenvalues of $Q_\Phi \mathbf{V}^{1/2} \Gamma \mathbf{V}^{1/2} Q_\Phi$, where $\Phi = \mathbf{V}^{1/2} \Delta$ and $Q_\Phi = \mathbf{I} - \Phi(\Phi^T \Phi)^- \Phi^T$;
- $\text{span}(\hat{\boldsymbol{\beta}})$ is a consistent estimator of $\mathcal{S}_{\boldsymbol{\xi}}$.

Identity Characteristic Function Estimator

In this section, if $\mathbf{V}_n = \mathbf{I}_{2mp}$ then the objective function in (3.1) becomes

$$F_d^{icf}(\mathbf{B}, \mathbf{C}) = (\text{vec}(\hat{\boldsymbol{\xi}}) - \text{vec}(\mathbf{BC}))^T \mathbf{I}_{2mp} (\text{vec}(\hat{\boldsymbol{\xi}}) - \text{vec}(\mathbf{BC})). \tag{3.4}$$

If $(\hat{\boldsymbol{\beta}}, \hat{\nu}) = \arg\min_{\mathbf{B},\mathbf{C}} F_d^{icf}(\mathbf{B}, \mathbf{C})$, then $\hat{\boldsymbol{\beta}}$ is the estimated basis of $\mathcal{S}_{E(y|\mathbf{x})}$. We call this estimator the identity characteristic function estimator (icfe) of $\boldsymbol{\beta}$.

Similarly, let $(\frac{\partial \text{vec}(\mathbf{BC})}{\partial \text{vec}(\mathbf{B})}, \frac{\partial \text{vec}(\mathbf{BC})}{\partial \text{vec}(\mathbf{C})})$ be the Jacobian matrix, and the value evaluated at $(\mathbf{B} = \boldsymbol{\beta}, \mathbf{C} = \nu)$ is $\Delta = (\nu^T \otimes \mathbf{I}_p, \mathbf{I}_{2m} \otimes \boldsymbol{\beta})$, a $2mp \times (2m + p)d$ matrix. The asymptotic properties of this estimation are established in theorem 3.2.3. Note in this formulation, $\mathbf{V}_n = \mathbf{V}$ is an identity matrix and no estimation is needed for this $\mathbf{V}_n$.

Theorem 3.2.3 Assume $\text{var}(y \sin(\omega^T \mathbf{x}))$, $\text{var}(y \cos(\omega^T \mathbf{x}))$ exist for every given ω and a random sample $(\tilde{\mathbf{x}}, \tilde{y})$ is given. Let $\mathcal{S}_{\boldsymbol{\xi}} = \sum_{i=1}^m \boldsymbol{\xi}_i$, $d = \dim(\mathcal{S}_{\boldsymbol{\xi}})$. Then we have:

- $\sqrt{n}(\text{vec}(\hat{\boldsymbol{\beta}}\hat{\nu}) - \text{vec}(\boldsymbol{\beta}\nu)) \to N(\mathbf{0}, \Delta(\Delta^T\Delta)^-\Delta^T\Gamma\Delta(\Delta^T\Delta)^-\Delta^T)$;
- $n\hat{F}_d^{icf}$ is asymptotically distributed as $\sum_{i=1}^{2mp} \lambda_i \chi_i^2(1)$, where $\hat{F}_d^{icf}$ is the minimum value of F_d^{icf}, $\chi_i^2(1)$s are independent chi-square random variable with degrees of freedom 1 and the weights λ_is are the eigenvalues of $Q_\Delta \Gamma Q_\Delta$;
- $\text{span}(\hat{\boldsymbol{\beta}})$ is a consistent estimator of $\mathcal{S}_{\boldsymbol{\xi}}$.

From theorems 3.2.2 and 3.2.3, we know that minimizing the objective function $F_d^{cf}(\mathbf{B}, \mathbf{C})$ gives a consistent estimate for $\text{vec}(\boldsymbol{\beta}\nu)$. And from the second conclusion, we can use the sequential hypothesis test to determine the structure dimension with all the population quantities replaced by the sample version[38][16].

Comment 1: Except $\mathbf{I}_{2m} \otimes \hat{\boldsymbol{\Sigma}}$ and $\mathbf{I}_{2mp}$, we can let $\mathbf{V}_n = \hat{\boldsymbol{\Gamma}}^{-1}$. In theory, $\boldsymbol{\Gamma} = \mathbf{W}\boldsymbol{\Sigma}_E\mathbf{W}^T$, where $\boldsymbol{\Sigma}_E = \text{cov}\begin{bmatrix} y\sin(\omega_1^T\mathbf{x}) \\ y\cos(\omega_1^T\mathbf{x}) \\ \vdots \\ y\sin(\omega_m^T\mathbf{x}) \\ y\cos(\omega_m^T\mathbf{x}) \end{bmatrix}$ and $\mathbf{W} = \begin{bmatrix} \omega_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \omega_1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \omega_2 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ 0 & 0 & 0 & \cdots & \omega_m & 0 \\ 0 & 0 & 0 & \cdots & 0 & \omega_m \end{bmatrix}$ is a $2mp \times 2m$ matrix. The maximum rank of $\boldsymbol{\Gamma}$ is $2m$ and thus $\boldsymbol{\Gamma}$ is not full rank. However, the formulation needs a positive definite $\mathbf{V}_n$. In sample version, the ωs are selected by large correlations between y and $\sin(\omega^T\mathbf{x})$ and $\cos(\omega^T\mathbf{x})$. If we use a large

number of ωs, it is possible that we are repeating the information from ω. Choosing a large number of ωs is not any better than choosing a small number. Therefore, we are mainly focused on the two cases discussed in this project.

Comment 2: As Wang et al.[63] pointed out, all the p dimensional vectors ω can be used to recover the CMS in theory. However, the selection of ω is crucial in practice. Thus we generate a large number of ωs from $N(0, \sigma^2\mathbf{I}_p)$ independently with $\sigma^2 = 0.02\pi^2/E(\mathbf{x}^T\mathbf{x})$ as Zhu et al. suggested (2010). Among all the ωs, we retain those ωs having large absolute correlations between y and $\cos(\omega^T\mathbf{x})$, y and $\sin(\omega^T\mathbf{x})$.

Assume d is given, the objective functions $F_d^{ccf}(\mathbf{B}, \mathbf{C})$ and $F_d^{icf}(\mathbf{B}, \mathbf{C})$ can be minimized by an alternating least squares method[16]. We summarize the algorithm as follows.

Step 1: Prepare $\hat{\boldsymbol{\xi}}$.

Step 2: Choose a constant matrix as the initial $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_d)$, where $\mathbf{b}_i$ is a unit vector with the ith element being 1 and the remaining being 0. Let $ite = 0$.

Step 3: Fix $\mathbf{B}$, update $\text{vec}(\mathbf{C})$ with a least square coefficient by regressing $\mathbf{V}_n^{1/2}\text{vec}(\hat{\boldsymbol{\xi}})$ on $\mathbf{V}_n^{1/2}(\mathbf{I}_{2m}\otimes\mathbf{B})$, that is $\text{vec}(\mathbf{C}) = ((\mathbf{I}_{2m}\otimes\mathbf{B}^T)\mathbf{V}_n(\mathbf{I}_{2m}\otimes\mathbf{B}))^{-1}(\mathbf{I}_{2m}\otimes\mathbf{B}^T)\mathbf{V}_n\text{vec}(\hat{\boldsymbol{\xi}})$.

Step 4: Fix $\mathbf{C}$, for each $k = 1, \cdots, d$, let $\alpha_k = \text{vec}(\hat{\boldsymbol{\xi}} - \mathbf{B}_{(-k)}\mathbf{C}_{(-k)})$, where $\mathbf{B}_{(-k)}$ and $\mathbf{C}_{(-k)}$ denote that the kth column of $\mathbf{B}$ and kth row of $\mathbf{C}$ are removed respectively. Then update $\hat{\mathbf{b}}_k = Q_{\mathbf{B}_{(-k)}}[Q_{\mathbf{B}_{(-k)}}(\mathbf{c}_k^T\otimes\mathbf{I}_p)\mathbf{V}_n(\mathbf{c}_k\otimes\mathbf{I}_p)Q_{\mathbf{B}_{(-k)}}]^{-}Q_{\mathbf{B}_{(-k)}}(\mathbf{c}_k^T\otimes\mathbf{I}_p)\mathbf{V}_n\alpha_k$, where $\mathbf{c}_k$ is the kth row of $\mathbf{C}$ and $Q_{\mathbf{B}_{(-k)}}$ is projecting onto the orthogonal complement of $\text{span}(\mathbf{B}_{(-k)})$ with the usual inner product. Normalize $\hat{\mathbf{b}}_k$ to have unit length and update $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_{k-1}, \hat{\mathbf{b}}_k, \mathbf{b}_{k+1}, \cdots, \mathbf{b}_d)$.

Step 5: Calculate $Fv = F_d(\mathbf{B}, \mathbf{C})$ and let $ite = ite + 1$.

Step 6: Repeat steps 3 to 5 until no decrease on Fv and output $\hat{\boldsymbol{\beta}} = \mathbf{B}$.

In our simulation, as Wang et al.[63] suggested, we generated 20,000 ωs. Since the number of ω has heavy impact on the computational cost, we tend to select a small

number of ω and 5% of the generated ωs that have the high correlation property are selected to estimate the CMS.

Dimension Determination

In the above formulations, we have assumed d, the dimension of the CMS, is known. However, in practice we have to estimate this d. In theory, we can use the sequential hypothesis test [38][7] to determine d by theorem 3.2.2 and 3.2.3. However, the results are not effective. Thus we adopt Xia et al.'s cross-validation approach to determine d[67].

Suppose we have the estimated directions $\hat{\boldsymbol{\beta}}_1, \cdots, \hat{\boldsymbol{\beta}}_{d_0}$ for the assumed dimension d_0, where $d_0 \in \{1, \cdots, p\}$. Let

$$\hat{y}_{j,d_0} = \sum_{i=1, i\neq j}^{n} K_{h_{d_0}}^{i,j} y_i / \sum_{i=1, i\neq j}^{n} K_{h_{d_0}}^{i,j},$$

where $K_{h_{d_0}}^{i,j} = K_{h_{d_0}}[\hat{\boldsymbol{\beta}}_1^T(\mathbf{x}_i - \mathbf{x}_j), \cdots, \hat{\boldsymbol{\beta}}_{d_0}^T(\mathbf{x}_i - \mathbf{x}_j)]$, h_{d_0} is the bandwidth corresponding to d_0 and $K(\cdot)$ represents a kernel function. Then, define the corresponding cross-validation value as $\mathrm{CV}(d_0) = \dfrac{\sum_{j=1}^{n}(y_j - \hat{y}_{j,d_0})^2}{n}$. For the trivial case, define $\mathrm{CV}(0) = \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$. Then, we choose the estimated dimension $\hat{d} = \mathrm{argmin}_{0\leq d\leq p}\mathrm{CV}(d)$. For the bandwidth h, as Xia et al. suggested, it is taken to be proportional to $n^{-1/(d+4)}$[67].

3.3 The Proposed Method for SVS

In a high dimensional dataset, it is common that only a small portion of the variables are important for regressing y on $\mathbf{x}$. In this section, we propose a new method to do variable selection by imposing a coordinate-independence penalty on the objective function. The coordinate-independence penalty is originally proposed by Chen et al.[9] and further studied by Qian et al.[46]. The exact form of this penalty term is $\rho_\theta(\mathbf{B}) = \sum_{k=1}^{p} \theta_k \parallel \mathbf{B}_k \parallel_2$, where $\theta = (\theta_1, \cdots, \theta_p)$ are the penalty weights and B_k is the kth row of matrix $\mathbf{B} = (B_1, \cdots, B_p)^T$. If we assume the ith variable, $\mathbf{x}_i$, is redundant, we have $\mathbf{e}_i^T\boldsymbol{\beta} = \mathbf{0}$, where $\mathbf{e}_i$ is a unit vector with the ith element being 1 and the

remaining elements being 0. Then, we can denote $\mathcal{A} = \{1 \leq i \leq p : \mathbf{e}_i^T \boldsymbol{\beta}\boldsymbol{\beta}^T \mathbf{e}_i > 0\}$ as the index set for important variables. We further let $s = |\mathcal{A}|$ be the number of important variables.

Corresponding to the two choices of $\mathbf{V}_n$ in Section 3.2, we develop two methods to perform variable selection and central mean subspace estimation for the large p small n scenario.

Covariance Weights: $\mathbf{V}_n = \mathbf{I}_{2m} \otimes \hat{\boldsymbol{\Sigma}}$

Objective function $F_d^{ccf}(\mathbf{B}, \mathbf{C})$ can be reformulated as

$$F_d^{ccf}(\mathbf{B}, \mathbf{C}) = \text{tr}((\hat{\boldsymbol{\xi}} - \mathbf{BC})^T \hat{\boldsymbol{\Sigma}} (\hat{\boldsymbol{\xi}} - \mathbf{BC})), \tag{3.5}$$

where tr is the operator for trace. By adopting the coordinate-independence penalty term, we get a new objective function,

$$L(\mathbf{B}, \mathbf{C}) = \frac{1}{2}\text{tr}((\hat{\boldsymbol{\xi}} - \mathbf{BC})^T \hat{\boldsymbol{\Sigma}} (\hat{\boldsymbol{\xi}} - \mathbf{BC}) + \lambda \rho_\theta(\mathbf{B}), \tag{3.6}$$

subject to $\mathbf{C}\mathbf{C}^T = I_d$. Note, according to proposition 1 of Qian et al. [46], most SDR methods put constraints on $\mathbf{B}$, but having constraints on $\mathbf{C}$ works the same. Thus, to facilitate the algorithm, we choose $\mathbf{C}\mathbf{C}^T = I_d$ instead of $\mathbf{B}^T\mathbf{B} = I_d$.

Once the minimizer $(\hat{\mathbf{B}}, \hat{\mathbf{C}}) = \text{argmin}_{\mathbf{B},\mathbf{C}} L(\mathbf{B}, \mathbf{C})$ is given, we simultaneously estimate the CMS by $\text{span}(\hat{\mathbf{B}})$ and the index set for important variables by $\{1 \leq i \leq p : \mathbf{e}_i^T \hat{\mathbf{B}}\hat{\mathbf{B}}^T \mathbf{e}_i > 0\}$. For the penalty weights, we adopt the method from adaptive lasso[83] and let $\theta_i = (\mathbf{e}_i^T \hat{\mathbf{B}}_\lambda \hat{\mathbf{B}}_\lambda^T \mathbf{e}_i)^{-\rho/2}$, where $\hat{\mathbf{B}}_\lambda$ is the minimizer of $L(\mathbf{B}, \mathbf{C})$ with a given λ, $\rho = 0.5$ and $\theta = \mathbf{1}_p$. With the following conditions assumed, the theoretical properties of these estimators are summarized in Theorem 3.3.1.

1. C1: $y \in \mathbb{R}$ is a sub-Gaussian random variable.

2. C2: The covariance matrix $\boldsymbol{\Sigma} = \sigma_{ij}$ has an element-wise upper bound and its minimum eigenvalue is bounded away from 0. That is, there are constants $\sigma_l, \sigma_u > 0$ such that $\sigma_{ij} < \sigma_u$ for every $1 \leq i, j \leq p$, and $\lambda_{min}(\boldsymbol{\Sigma}) > \sigma_l$, where $\lambda_{min}(\boldsymbol{\Sigma})$ is denoted as the minimum eigenvalue of matrix $\boldsymbol{\Sigma}$.

3. C3: Assume $m^2 s \log p_n = O(n^{1-2\eta})$ and $ds^2 \log p_n = O(n^{1-2\eta})$ for some constant $\eta \in (0, 1/2)$ and $p_n = \max\{p, n\}$.

4. C4: Assume $\min_{i \in \mathcal{A}} \mathbf{e}_i^T \boldsymbol{\beta}\boldsymbol{\beta}^T \mathbf{e}_i > C_\phi n^{-\phi}$ for some $\phi \in [0, 2\eta)$ and constant C_ϕ.

5. C5: Assume the nonzero singular values of $\hat{\boldsymbol{\xi}}$ are bounded away from 0.

Theorem 3.3.1 Let $\| \cdot \|_F$ denote the Frobenius norm, under conditions $C1 - C5$, the minimizer $(\hat{\mathbf{B}}, \hat{\mathbf{C}})$ and the estimated index set $\hat{\mathcal{A}}$ satisfy

- CMS estimation consistency: $\| P_{\mathcal{S}_{\hat{\mathbf{B}}}} - P_{\mathcal{S}_{E(y|\mathbf{x})}} \|_F = O_p(\sqrt{ms \log p_n / n})$,
- SVS estimation consistency: $P(\hat{\mathcal{A}} = \mathcal{A}) \to 1$ as $n \to \infty$.

Identity Weights: $\mathbf{V}_n = \mathbf{I}_{2mp}$

For identity weights, we add a penalty term to the objective function $F_d^{icf}(\mathbf{B}, \mathbf{C})$, and we have

$$L_I(\mathbf{B}, \mathbf{C}) = \frac{1}{2} \text{tr}((\hat{\boldsymbol{\xi}} - \mathbf{B}\mathbf{C})^T (\hat{\boldsymbol{\xi}} - \mathbf{B}\mathbf{C}) + \lambda \rho_\theta(\mathbf{B}), \tag{3.7}$$

subject to $\mathbf{C}\mathbf{C}^T = I_d$. Similarly, if the minimizer $(\hat{\mathbf{B}}, \hat{\mathbf{C}}) = \text{argmin}_{\mathbf{B},\mathbf{C}} L_I(\mathbf{B}, \mathbf{C})$ is given, we simultaneously estimate the CMS by $\text{span}(\hat{\mathbf{B}})$ and the index set for important variables by $\{1 \leq i \leq p : \mathbf{e}_i^T \hat{\mathbf{B}}\hat{\mathbf{B}}^T \mathbf{e}_i > 0\}$. For the penalty weights, we use the same adaptive lasso technique as suggested in the previous section. For the estimators, we have the following theoretical properties.

Theorem 3.3.2 Under conditions $C1 - C5$, the minimizer $(\hat{\mathbf{B}}, \hat{\mathbf{C}})$ and estimated index set $\hat{\mathcal{A}}$ satisfy

- CMS estimation consistency: $\| P_{\mathcal{S}_{\hat{\mathbf{B}}}} - P_{\mathcal{S}_{E(y|\mathbf{x})}} \|_F = O_p(\sqrt{sm \log p_n / n})$,
- SVS estimation consistency: $P(\hat{\mathcal{A}} = \mathcal{A}) \to 1$ as $n \to \infty$.

Algorithm

To optimize the penalized objective functions, we adopt the iterative parallelizable coordinate decent (IPCD) algorithm from Qian et al.[46] which updates $\mathbf{B}$ and $\mathbf{C}$ iteratively until the algorithm converges.

Algorithm for Covariance Weights

To update $\mathbf{B}$, we first let

$$U(\mathbf{B},\mathbf{C}) = \frac{\partial F_d^{ccf}(\mathbf{B},\mathbf{C})}{2\partial \mathrm{vec}(\mathbf{B}^T)} = -[\mathbf{I}_p \otimes (\mathbf{C}\Upsilon^T)]\mathrm{vec}(\mathbf{I}_p) + [\hat{\mathbf{\Sigma}} \otimes \mathbf{I}_d]\mathrm{vec}(\mathbf{B}^T),$$

where $\Upsilon = \hat{\mathbf{\Sigma}}\hat{\boldsymbol{\xi}}$. We further define $U_t = U(\mathbf{B}_{(t)},\mathbf{C}_{(t)})$, $\tilde{h} = \lambda_{\max}(\hat{\mathbf{\Sigma}})$ and $\tilde{\mathbf{\Sigma}} = \tilde{h}\mathbf{I}_p \otimes \mathbf{I}_d$, where $(\mathbf{B}_{(t)},\mathbf{C}_{(t)})$ is the estimator of $(\mathbf{B},\mathbf{C})$ after t-th iteration. To get $\mathbf{B}_{(t+1)}$, we make use of a quadratic approximation of $L(\mathbf{B},\mathbf{C})$. Given the estimator from the tth estimator $(\mathbf{B}_{(t)},\mathbf{C}_{(t)})$,the quadratic approximation $L_q^{(t)}(\mathbf{B})$ has the following expression

$$U_t^T(\mathrm{vec}(\mathbf{B}^T) - \mathrm{vec}(\mathbf{B}_{(t)}^T)) + \frac{1}{2}(\mathrm{vec}(\mathbf{B}^T) - \mathrm{vec}(\mathbf{B}_{(t)}^T))^T\tilde{\mathbf{\Sigma}}(\mathrm{vec}(\mathbf{B}^T) - \mathrm{vec}(\mathbf{B}_{(t)}^T)) + \lambda\rho_\theta(\mathbf{B}).$$

We update $\mathbf{B}$ by minimizing $L_q^{(t)}(\mathbf{B})$. That's $\mathbf{B}_{(t+1)} = \mathrm{argmin}_{\mathbf{B}} L_q^{(t)}(\mathbf{B})$. Qian et al.[46] pointed out that the optimization guarantees that $L(\mathbf{B}_{(t+1)},\mathbf{C}_{(t+1)}) \leq L(\mathbf{B}_{(t)},\mathbf{C}_{(t)})$. Another descent property is that the minimizer of $L_q^{(t)}(\mathbf{B})$ holds a closed form and it is computationally efficient with the facilitation of parallel computing. Further if the Karush-Kuhn-Tucker condition holds, then the jth row of $\mathbf{B}$, $\mathbf{B}_{(t+1)^j}$, can be explicitly expressed as

$$\frac{1}{\tilde{h}}(1 - \frac{\lambda\theta_j}{\parallel \mathbf{C}_{(t)}\Upsilon^T\mathbf{e}_j - \sum_{i=1}^p h_{ji}\mathbf{B}_{(t)^i} + \tilde{h}\mathbf{B}_{(t)^j} \parallel_2})_+(\mathbf{C}_{(t)}\Upsilon^T\mathbf{e}_j - \sum_{i=1}^p h_{ji}\mathbf{B}_{(t)^i} + \tilde{h}\mathbf{B}_{(t)^j}),$$

where $h_{ji} = (\hat{\mathbf{\Sigma}})_{ji}$ and $a_+ = \max\{0,a\}$. Now, it is obvious that the p rows can be simultaneously estimated in a parallel manner.

To update $\mathbf{C}$, we let $\mathbf{B} = \mathbf{B}_{(t+1)}$ be fixed and then solve a Stiefel manifold optimization problem with the facilitation of reduced rank procrustes rotation [83]. That is

$$\mathbf{C}_{(t+1)} = \mathrm{argmin}_{\mathbf{C}} - \mathrm{tr}(\Upsilon^T\mathbf{B}_{(t+1)}\mathbf{C}) = \mathrm{argmax}_{\mathbf{C}} tr(\Lambda_1\mathbf{D}\Lambda_2\mathbf{C}) = \Lambda_2\Lambda_1,$$

where $\Lambda_1 \mathbf{D} \Lambda_2$ is the consequence of applying SVD on $\Upsilon^T \mathbf{B}_{(t+1)}$.

Now, we summarize the algorithm as follows:

Step 1: Get the initial $\mathbf{B}$ and $\mathbf{C}$.

Step 2: Update each row of $\mathbf{B}$ by $\mathbf{B}_{(t+1)^j}$ and let $\mathbf{B}_{(t+1)} = (\mathbf{B}_{(t+1)^1}, \cdots, \mathbf{B}_{(t+1)^p})^T$.

Step 3: Update $\mathbf{C}$ by $\mathbf{C}_{(t+1)} = \text{argmin}_{\mathbf{C}} - \text{tr}(\Upsilon^T \mathbf{B}_{(t+1)} \mathbf{C})$.

Repeat steps 2 and 3 until the $\mathbf{B}$ and $\mathbf{C}$ converge.

Algorithm for Identity Weights

To update $\mathbf{B}$, we first let

$$U_I(\mathbf{B}, \mathbf{C}) = \frac{\partial F^{icf}(\mathbf{B}, \mathbf{C})}{2 \partial \text{vec}(\mathbf{B}^T)} = -[\mathbf{I}_p \otimes (\mathbf{C} \boldsymbol{\xi}^T)] \text{vec}(\mathbf{I}_p) + \text{vec}(\mathbf{B}^T).$$

We further define $U_{It} = U_I(\mathbf{B}_{(t)}, \mathbf{C}_{(t)})$, where $(\mathbf{B}_{(t)}, \mathbf{C}_{(t)})$ is the estimator of $(\mathbf{B}, \mathbf{C})$ after tth iteration. To get $\mathbf{B}_{(t+1)}$, we make use of an quadratic approximation of $L_I(\mathbf{B}, \mathbf{C})$. Given the estimator from the tth estimator $(\mathbf{B}_{(t)}, \mathbf{C}_{(t)})$, we have the quadratic approximation, $L_{Iq}^{(t)}(\mathbf{B})$, expressed as

$$U_{It}^T (\text{vec}(\mathbf{B}^T) - \text{vec}(\mathbf{B}_{(t)}^T)) + \frac{1}{2} (\text{vec}(\mathbf{B}^T) - \text{vec}(\mathbf{B}_{(t)}^T))^T (\text{vec}(\mathbf{B}^T) - \text{vec}(\mathbf{B}_{(t)}^T)) + \lambda \rho_\theta(\mathbf{B}).$$

By minimizing $L_{Iq}^{(t)}(\mathbf{B})$, we update $\mathbf{B}$. That is $\mathbf{B}_{(t+1)} = \text{argmin}_{\mathbf{B}} L_{Iq}^{(t)}(\mathbf{B})$. Similarly, according to Qian et al. [46], the optimization guarantees that $L_I(\mathbf{B}_{(t+1)}, \mathbf{C}_{(t+1)}) \leq L_I(\mathbf{B}_{(t)}, \mathbf{C}_{(t)})$. For jth row of $\mathbf{B}$, $\mathbf{B}_{(t+1)^j}$, can be explicitly expressed as

$$\mathbf{B}_{(t+1)^j} = (1 - \frac{\lambda \theta_j}{\parallel \mathbf{C}_{(t)} \boldsymbol{\xi}^T \mathbf{e}_j \parallel_2})_+ (\mathbf{C}_{(t)} \boldsymbol{\xi}^T \mathbf{e}_j).$$

Again, the p rows can be simultaneously estimated by a parallel manner.

As in the covariance weights method, $\mathbf{C}_{(t+1)} = \text{argmin}_{\mathbf{C}} - \text{tr}(\boldsymbol{\xi}^T \mathbf{B}_{(t+1)} \mathbf{C}) = \text{argmax}_{\mathbf{C}} tr(\Lambda_1 \mathbf{D} \Lambda_2 \mathbf{C}) = \Lambda_2 \Lambda_1$, where $\Lambda_1 \mathbf{D} \Lambda_2$ is the consequence of applying SVD on $\boldsymbol{\xi}^T \mathbf{B}_{(t+1)}$.

The detailed algorithm to solve $L_I(\mathbf{B}, \mathbf{C})$ is summarized here.

Step 1: Get the initial $\mathbf{B}$ and $\mathbf{C}$.

Step 2: Update each row of $\mathbf{B}$ by $\mathbf{B}_{(t+1)^j}$ and let $\mathbf{B}_{(t+1)} = (\mathbf{B}_{(t+1)^1}, \cdots, \mathbf{B}_{(t+1)^p})^T$.

Step 3: Update $\mathbf{C}$ by $\mathbf{C}_{(t+1)} = \text{argmin}_{\mathbf{C}} - \text{tr}(\boldsymbol{\xi}^T \mathbf{B}_{(t+1)} \mathbf{C})$.

Step 4: Repeat step 2 and step 3 until the $\mathbf{B}$ and $\mathbf{C}$ converge.

In our simulation, when the maximum element-wise difference between two consecutive $\mathbf{B}$ and $\mathbf{C}$ is smaller than 1e-12, we say the algorithm converges. For the tuning parameter, we use the cross-validation technique.

3.4 Numerical Study

In this section, we will show the efficacy of our method by simulation studies. These numerical studies are mainly performed in **Matlab**. The SDR and dimension test code is available on github with the link *https://github.com/wangpeinihao/code2*. The SVS code is modified from Qian et al.'s [46] **Matlab** code; the dimension determination code is modified from Xia et al.'s **R** code.

The SDR accuracy of the estimation is measured by the vector correlation coefficient [32], trace correlation[31] and the angles[16], which are denoted by q^2, r^2 and θ. They are expressed as $\prod_{i=1}^{d} \rho_i$, $\sum_{i=1}^{d} \rho_i / d$ and $180 \cos^{-1}(\frac{|\boldsymbol{\beta}_i^T \hat{\boldsymbol{\beta}}_i|}{||\boldsymbol{\beta}_i|| ||\hat{\boldsymbol{\beta}}_i||})/\pi$ respectively, where ρ_i is the ith largest eigenvalue of $\boldsymbol{\beta}^T \hat{\boldsymbol{\beta}} \hat{\boldsymbol{\beta}}^T \boldsymbol{\beta}$. Large values of q^2, r^2 (close to 1) and small values of θ (close to 0) are preferred. For the angle, we present the benchmark angle θ_r for comparison. We compare our methods to IRE, RIRE and PHD.

For SVS, to evaluate the performance, we use average true positive (ATP): the number of correctly selected as active predictors; average false positive (AFP): the number of falsely selected as actively predictors; Frobenius norm: $||\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}^T \hat{\boldsymbol{\beta}})^{-1} \hat{\boldsymbol{\beta}}^T - \boldsymbol{\beta}(\boldsymbol{\beta}^T \boldsymbol{\beta})^{-1} \boldsymbol{\beta}^T||_F$ and trace correlation. The better estimation will have ATP close to the real number of active predictors, AFP close to 0, norm close to 0 and r^2 close to 1.

Simulation studies

Example 3.4.1

In this example, we use the following three models to show the efficacy of our SDR methods (ICFE and CCFE). The models are

Model 1 $Y = \cos(2X_1) - \cos(X_2) + 0.5\epsilon$;

Model 2 $Y = X_1 + X_2 + X_3 + X_4 + 0.2\epsilon$;

Model 3 $X_y = \sqrt{7/8}\beta y + 0.5\epsilon$.

Model 1 is example 8.1 from Li[39]. $X \sim N(\mathbf{0}, \mathbf{I})$, ϵ is from standard normal distribution, $n = 400$ and $p = 10$. Model 2 is a linear model with $p = 10$, $d = 1$, $n = 200$, $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I})$ and $\epsilon \sim N(0, 1)$. Model 3 is a modified inverse model from section 5.1 of Cook [14] with $p = 10$, $d = 1$, $n = 200$, $\beta = c(1, 0, 1, 0, 1, 0, 1, 0, 1, 0)$, $y \sim N(0, 0.5)$ and $\epsilon \sim N(\mathbf{0}, \mathbf{I})$.

The SDR results of these models are listed in Table 3.1 along with the comparison to IRE, RIRE and PHD. For model 1, our ICFE performs the best and then followed by CCFE and PHD with the large correlation and small angle. It is within our expectation that the IRE and RIRE fail to find the true direction because they cannot detect the symmetric pattern. For linear model 2, all IRE, RIRE, ICFE and CCFE work well. For model 3, CCFE performs the best and ICFE is the second best. PHD fails to work for both models 2 and 3 because as a second order method it easily loses the linear pattern direction. Across the three models with different patterns, our proposed new methods work very well.

Example 3.4.2

This example uses models 1, 2 and 3 to show the performance of cross-validation in determining the structure dimension d. The results are reported in Table 3.2. From the table, we see that the cross-validation method works well to determine the dimension especially for model 2. In some cases, this method tends to overestimate

Table 3.1: SDR results comparison

Models	Methods	r^2	q^2	θ_1	θ_2	θ_r
Model 1 $d=2$	ICFE	0.955(0.018)	0.912(0.035)	15.030	18.066	79.687
	CCFE	0.950(0.022)	0.902(0.042)	15.972	19.034	79.687
	IRE	0.216(0.116)	0.032(0.047)	77.433	77.142	79.687
	RIRE	0.215(0.116)	0.031(0.046)	77.823	76.830	79.687
	PHD	0.936(0.027)	0.872(0.051)	19.507	20.233	79.687
Model 2 $d=1$	ICFE	0.982(0.008)	0.982(0.008)	10.551	/	83.847
	CCFE	0.976(0.010)	0.976(0.010)	12.251	/	83.847
	IRE	0.992(0.004)	0.992(0.004)	6.876	/	83.847
	RIRE	0.987(0.007)	0.987(0.007)	8.784	/	83.847
	PHD	0.241(0.201)	0.241(0.201)	75.630	/	83.847
Model 3 $d=1$	ICFE	0.967(0.024)	0.967(0.024)	14.161	/	84.975
	CCFE	0.971(0.021)	0.971(0.021)	13.080	/	84.975
	IRE	0.904(0.042)	0.904(0.042)	24.675	/	84.975
	RIRE	0.885(0.049)	0.885(0.049)	27.152	/	84.975
	PHD	0.304(0.163)	0.304(0.163)	71.905	/	84.975

the dimension. When the dimension cannot be determined accurately, we prefer the overestimation because it guarantees no loss on the important regression information. In addition, the method works better when the sample size is larger.

Example 3.4.3

Model 4 $\mathbf{X} = \sqrt{7/8}\beta y + 0.5\epsilon$,

where $\beta = \mathbf{e}_{100} + \mathbf{e}_{200} + \cdots + \mathbf{e}_{800}$, $p = 1000, n = 200$, $\epsilon \sim N(\mathbf{0}, \mathbf{I})$ and $y \sim N(0, 0.5)$. This model is from section 5.1 of Cook[14].

Model 5 $Y = \sin(X^T\beta)^2 + X^T\beta + 0.2\epsilon$,

where β has coefficient 1 at 5 random positions, $p = 500, n = 200$, $\epsilon \sim N(0, 1)$ and $\mathbf{X} \sim N(0, \mathbf{\Sigma})$ and $\mathbf{\Sigma}$ is a block diagonal matrix with 100 blocks and in each block, it has 1 in the diagonal, 0.5 in the off-diagonal. This model is equation 3.1 in Weng and Yin[64].

In this example, we use models 4 and 5 to evaluate the performance of the SVS proposed in section 3.3. The results are reported in Table 3.3. We compare our results

Table 3.2: Dimension Determination in CV

Models	Methods	n	$d=0$	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$
Model 1	ICFE	200	0	1	89	9	1	0
$d=2$		400	0	0	99	1	0	0
		800	0	0	100	0	0	0
	CCFE	200	0	0	91	9	0	0
		400	0	0	96	4	0	0
		800	0	0	100	0	0	0
Model 2	ICFE	200	0	100	0	0	0	0
$d=1$		400	0	100	0	0	0	0
		800	0	100	0	0	0	0
	CCFE	200	0	100	0	0	0	0
		400	0	100	0	0	0	0
		800	0	100	0	0	0	0
Model 3	ICFE	200	0	64	21	11	1	3
$d=1$		400	0	69	20	6	0	5
		800	0	75	14	5	5	1
	CCFE	200	0	94	0	5	0	1
		400	0	96	0	2	0	2
		800	0	99	0	1	0	0

Table 3.3: SVS results comparison

Models	Methods	ATP	AFP	r^2	norm
	ICFE	8.00	1.97	0.991	0.183
Model 4	CCFE	7.95	7.34	0.934	0.489
$d=1$	SSIR	8.00	54.96	0.756	0.910
	ICFE	5.00	7.33	0.895	0.586
Model 5	CCFE	5.00	34.43	0.793	0.856
$d=1$	SSIR	4.63	11.72	0.825	0.497

to sparse SIR (SSIR) by Qian et al.[46]. For model 4, the ICFE works best with the largest ATP and smallest AFP for variable selection and largest correlation and smallest distance for direction estimation; CCFE has a similar performance to SSIR in variable selection but with a much better performance in direction estimation. For model 5, ICFE is the best with large values in ATP/r^2 and small values in AFP/norm. The SSIR results here is taken from Weng and Yin since we are using their model [64]. We conclude that our ICFE outperforms SSIR based on the example presented here.

3.5 Discussion

In the project, we let the matrix $\mathbb{R}_n$ in equation 3.1 be $\mathbf{I}_{2m}$ because we believe all the directions $\psi(\omega_i)$ are of equal importance. We can also choose it as a diagonal matrix with different weights for $\psi(\omega)$ on its diagonal. Although we start from different methods to organize $\psi(\omega)$, we stop with the same quadratic discrepancy function since the weight matrix $\mathbb{R}_n$ gets canceled if we use the optimal choose of $\mathbf{V}_n$ in equation 3.1. Thus theoretically, all other properties will follow the same. In practice, the $\psi(\omega)$s with small weights may not contribute to estimate the SDR and the targeted subspace is mainly controlled by those with large weights.

Chapter 4 Reduced Rank Multinomial Logistic Regression in Markov Chains with Application to Cognitive Data

4.1 Introduction

Dimension reduction is an importance topic in statistics, especially when the given data is in high volume and complex. Reduced-rank regression [4] is one of the examples. In reduced-rank regression, instead of directly estimating the $p \times J$ coefficients matrix $\boldsymbol{\beta}$, it estimates two lower-rank coefficient matrices, $p \times T$ matrix $\mathbf{A}$ and $T \times J$ matrix $\mathbf{G}$, with $T \leq \min\{p, J\}$; and then takes their product, $\boldsymbol{\beta} = \mathbf{AG}$. We refer more details to Izenman [34], Schmidli [49], Yee and Wild [70] and Fiocco et al. [26].

In this project, we propose a new model, called reduced rank (partial reduced rank) multinomial logistic regression for Markov chains, abbreviated as RR-MLRfMC (PRR-MLRfMC). In this novel model, we combine the RR-MLR idea with a first order Markov chain. We then fit this novel model to a dataset from a longitudinal study of aging and dementia. In addition, we show how the RR idea can be applied to only a subset of the risk factors.

The rest of the chapter is outlined as follows: we introduce the RR-MLRfMC, PRR-MLRfMC, along with its algorithm, in Section 4.2; the detailed development and application to data using RR-MLRfMC are given in Section 4.3; and a discussion concludes the project in Section 4.4.

4.2 The Proposed Method

RR-MLR

The MLR model is often used to analyze nominal responses with more than two categories. Assume Y takes discrete values from $1, 2, \cdots, J$, and $P(Y = j)$ is the probability that the response falls into the jth category. Based on the fact that $\sum_{j=1}^{J} P(Y = j) = 1$, we only work on $J - 1$ categories. In MLR, we assess whether

the response probabilities depend on covariates $\mathbf{z} = (\mathbf{z}_1, \cdots, \mathbf{z}_p)$ and assume the log-odds of the response probability is linearly related to the covariates. That is,

$$\log(\frac{P(Y=j)}{P(Y=1)}) = \mathbf{z}^T \boldsymbol{\beta}_j, \tag{4.1}$$

where $j = 2, \cdots, J$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_J)$ is a $p \times (J-1)$ matrix. Note the reference category $Y = 1$ could be replaced by other categories.

A crucial problem of MLR is that the number of parameters to be estimated is large, especially when p is large (e.g., $p(J-1)$ with the intercept excluded). In RR-MLR, $\boldsymbol{\beta} = \mathbf{AG}$, where $\mathbf{A}$ is a $p \times T$, and $\mathbf{G}$ is a $T \times (J-1)$ matrix respectively, $T = 1, \cdots, \min\{p, J-1\}$. Then, the number of parameters need to be estimated is $T(p+J-1-T)$ and the difference $(J-1-T)(p-T)$ is large when T is small. When dimension is T, by optimizing the log-likelihood function with respect to $\mathbf{A}$ and $\mathbf{G}$, the candidate coefficient matrix is given as $\hat{\boldsymbol{\beta}}_T = \hat{\mathbf{A}}\hat{\mathbf{G}}$. The optimal T can be selected by several information criteria, e.g., Akaike Information Criterion (AIC)[2].

RR-MLR for Markov Chains

In this section we show how to apply RR-MLR to a Markov Chain. Assume each person $i = 1, \cdots, n$ provides n_i observations $\{X_{i1}, \cdots, X_{in_i}\}$, where X_{ij} is the state person i occupies at time j. Let $X_{i,j-1}$ and $X_{i,j}$ denote the prior and current states for each observation respectively in the dataset. We use multinomial logits to compute the probability of current states given the prior states. That is, for non-reference categories,

$$P(X_{ij} = l | X_{i,j-1} = u) = \frac{e^{\mathbf{z}_i^T \lambda_{ul}}}{1 + \sum_{l^*=1, l^* \neq u}^{L} e^{\mathbf{z}_i^T \lambda_{ul^*}}}; \tag{4.2}$$

and for reference category (prior state),

$$P(X_{ij} = u | X_{i,j-1} = u) = \frac{1}{1 + \sum_{l^*=1, l^* \neq u}^{L} e^{\mathbf{z}_i^T \lambda_{ul^*}}}, \tag{4.3}$$

where $l = 1, \cdots, u-1, u+1, \cdots, L$ and $u = 1, \cdots, U$. Combining the $\boldsymbol{\lambda}_{ul}$s from all the transitions defines the coefficient matrix and $\boldsymbol{\beta} = (\boldsymbol{\lambda}_{1,2}, \cdots, \boldsymbol{\lambda}_{1,L}, \cdots, \boldsymbol{\lambda}_{U,1}, \cdots, \boldsymbol{\lambda}_{U,U-1}, \boldsymbol{\lambda}_{U,U+1}, \boldsymbol{\lambda}_{U,L}) = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_{U*(L-1)})$, where $\boldsymbol{\lambda}_{u,l} = \boldsymbol{\beta}_{(u-1)L+l}$ for $u > l$ and $\boldsymbol{\lambda}_{u,l} =$

$\boldsymbol{\beta}_{(u-1)L-u+l}$ for $u < l$. With reduced-rank, $\boldsymbol{\beta} = \mathbf{AG}$ and $\boldsymbol{\beta}_b = \mathbf{AG}_b$, where $\mathbf{A}$ and $\mathbf{G}$ are $p \times T$ full column and $T \times p$ full row rank matrices respectively, $T = 1, \cdots, \min\{p, U*(L-1)\}$. That is each coefficient $\boldsymbol{\beta}_b$ is a linear combination of a common matrix, $\mathbf{A}$, for each transition.

Then, the log likelihood for these data is $\ln(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{n_j} \ln P(X_{ij}|X_{i,j-1})$; and for each observation, the log likelihoods for without and with reduced-rank are

$$L_{ij} = -\log(1 + \sum_{l=1, l\neq X_{i,j-1}}^{L} e^{\mathbf{z}_i^T \boldsymbol{\lambda}_{X_{i,j-1},l}}) + (1 - \delta_{X_{i,j-1},X_{ij}})\mathbf{z}_i^T \boldsymbol{\lambda}_{X_{i,j-1},X_{ij}}, \tag{4.4}$$

and

$$L_{ij} = -\log(1 + \sum_{l=1, l\neq u}^{L} e^{\mathbf{z}_i^T \mathbf{AG}_b}) + (1 - \delta_{u,X_{ij}})\mathbf{z}_i^T \mathbf{AG}_b, \tag{4.5}$$

where $b = (u-1)L + l$ for $u > l$ and $b = (u-1)L - u + l$ for $u < l$.

Since there is no explicit solution for $\mathbf{A}$ and $\mathbf{G}$, we use Newton-Raphson Algorithm to get the numerical solution by iteratively updating $\mathbf{A}$ and $\mathbf{G}$. Note, the choice of $\mathbf{A}$ and $\mathbf{G}$ are not unique because for any nonsingular matrix $\mathbf{M}$, $\boldsymbol{\beta} = \mathbf{AG} = \mathbf{AMM}^{-1}\mathbf{G} = \mathbf{A}'\mathbf{G}'$, where $\mathbf{A}' = \mathbf{AM}$ and $\mathbf{G}' = \mathbf{M}^{-1}\mathbf{G}$. In this project, to get the unique $\mathbf{A}$ and $\mathbf{G}$, we use singular value decomposition as Fiocco et al. suggested[26]. That is $\mathbf{AG} = (UD^{1/2})(D^{1/2}V^T)$, where $\boldsymbol{\beta} = UDV^T$. See Yee and Hastie [70] for other remedies.

Algorithm 4.2.1 *We summarize the algorithm as follows:*

- *Step 1: Set an initial* $\mathbf{G}$*, initial* $\Delta = 10$*,* $L_1 = 100$*;*
- *Step 2: If* $\Delta > 10^{-5}$*, continue; otherwise, stop;*
- *Step 3: (Zig Step) Fix* $\mathbf{G}$*, update* $\mathbf{A}$ *by Newton-Raphson method until it converges, which is* $vec(\mathbf{A})^q = vec(\mathbf{A})^{q-1} - D_{a2}^{-1}D_{a1}/\mathbf{A} = \mathbf{A}^{q-1}$*, where* D_{a1} *and* D_{a2} *are the first and second derivatives of log-likelihood function with respect to* $vec(\mathbf{A})$*;*
- *Step 4: (Zag Step) Fix* $\mathbf{A}$*, update* $\mathbf{G}$ *by* U *individual multinomial logistic regressions; here, the covariates are* $\mathbf{z}_i^T\mathbf{A}$ *instead of* $\mathbf{z}_i$ *and the initial* $\mathbf{A}$ *is a constant matrix with elements 0.01;*

- *Step 5: Find* $\boldsymbol{\beta} = \mathbf{AG}$*, calculate the log likelihood* L_2*, get* $\Delta = L_1 - L_2$ *and let* $L_1 = L_2$*;*

- *Step 6: Repeat step 2 to step 5.*

Comment: The algorithm here is called the Zigzag method; the convergence of this method is not guaranteed.

Partial RR-MLR for Markov Chains

It is possible to have some scenarios where the RR is applied only to a subset of the covariates. In this case, we assume the covariates comprise a $(q+p)\times 1$ vector, $(\mathbf{w}_1, \cdots, \mathbf{w}_q, \mathbf{z}_1, \cdots, \mathbf{z}_p)^T$ and RR applies to the covariates in $\mathbf{z}$. Thus, we label this method as Partial RR-MLR for Markov Chain (PRR-MLRfMC). Given the dataset, the log-likelihood for each observation, L_{ij}, in Equations 4.4 and 4.5 are modified as

$$-\log(1+\sum_{l=1,l\neq X_{i,j-1}}^{L} e^{\mathbf{w}_i^T\boldsymbol{\rho}_{X_{i,j-1},l}+\mathbf{z}_i^T\boldsymbol{\lambda}_{X_{i,j-1},l}})+(1-\delta_{X_{i,j-1},X_{ij}})(\mathbf{w}_i^T\boldsymbol{\rho}_{X_{i,j-1},X_{ij}}+\mathbf{z}_i^T\boldsymbol{\lambda}_{X_{i,j-1},X_{ij}}), \tag{4.6}$$

and

$$-\log(1+\sum_{l=1,l\neq u}^{L} e^{\mathbf{w}_i^T\boldsymbol{\rho}_{u,l}+\mathbf{z}_i^T\mathbf{AG}_b})+(1-\delta_{u,X_{ij}})(\mathbf{w}_i^T\boldsymbol{\rho}_{u,l}+\mathbf{z}_i^T\mathbf{AG}_b), \tag{4.7}$$

where $b=(u-1)L+l$ for $u>l$ and $b=(u-1)L-u+l$ for $u<l$.

Let $\boldsymbol{\theta}=(\boldsymbol{\rho}_{1,2},\cdots\boldsymbol{\rho}_{1,L},\cdots,\boldsymbol{\rho}_{U,1},\cdots,\boldsymbol{\rho}_{U,U-1},\boldsymbol{\rho}_{U,U+1},\cdots,\boldsymbol{\rho}_{U,L})=(\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_{U*(L-1)})$. Now the goal is to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. We use the same zigzag method as indicated in Algorithm 4.2.1 and it is modified as follows:

Algorithm 4.2.2 *This is the algorithm for PRR-MLRfMC.*

- *Step 1: Set an initial* $\mathbf{G}$*, initial* $\boldsymbol{\theta}$*, initial* $\Delta = 10$*,* $L_1 = 100$*;*

- *Step 2: If* $\Delta > 10^{-5}$*, continue; otherwise, stop;*

- *Step 3: (Zig Step) Fix* $\mathbf{G}$ *and* $\boldsymbol{\theta}$*, update* $\mathbf{A}$ *by Newton-Raphson method until it converges, which is* $vec(\mathbf{A})^q = vec(\mathbf{A})^{q-1} - D_{a2}^{-1}D_{a1}/\mathbf{A} = \mathbf{A}^{q-1}$*, where* D_{a1} *and*

D_{a2} are the first and second derivatives of log-likelihood function with respect to vec($\mathbf{A}$);

- *Step 4: (Zag Step) Fix $\mathbf{A}$, update $\boldsymbol{\theta}$ and $\mathbf{G}$ by U individual multinomial logistic regressions; here, the covariates are $(\mathbf{w}_i^T, \mathbf{A}^T\mathbf{z}_i)^T$ instead of $\mathbf{z}_i$ and the initial $\mathbf{A}$ is a constant matrix with elements 0.01;*

- *Step 5: Use $\boldsymbol{\theta}$ and $\boldsymbol{\beta} = \mathbf{AG}$ to calculate the log-likelihood L_2, get $\Delta = L_1 - L_2$ and let $L_1 = L_2$;*

- *Step 6: Repeat step 2 to step 5.*

4.3 Application to longitudinal data on cognitive assessments

To illustrate the methodology we analyze Apolipoprotein-E (APOE)[47] gene $\epsilon 4$ allele(s) (i.e., carrying at least one $\epsilon 4$, APOE4) as a risk factor for transitions among congnitive states. A seven-state Markov Chain model was constructed with four transient states: normal cognition, amnestic MCI (A-MCI; a classification of mild impairment due solely to poor performance on a memory exam taken annually but otherwise cognitively normal), mixed MCI (M-MCI; poor performance on a non-memory cognitive test, instead of or in addition to, in the memory domain; otherwise cognitively normal), and MCI (a clinician diagnosis of mild cognitive impairment verified by an informant and poor cognitive test scores). The model also includes three absorbing states: dropout, death without a diagnosis of clinical dementia, or a diagnosis of a clinical dementia. Table 4.1 provides the frequency of the one-tep transitions based on the data discussed by Abner et al. [1] and collected annually on the first 649 participants in the BRAiNS (Biologically Resilient Adults in Neurological Studies) cohort at the University of Kentucky's Alzheimer's Disease Center.[49] Notice that once in the clinical MCI state a participant can only transition forward to an absorbing state or remain in MCI. The purpose of the analysis is to compute log-odds, can be transformed back to transition probabilities, associated with APOE4 carrier status adjusted for the presence of eight known risk factors for a dementia (see

next section). Referring to Table 4.1, if $X_{i,j-1} = 1, 2, 3$, then $X_{i,j} = 1, 2, 3, 4, 5, 6, 7$; and if $X_{i,j-1} = 4$, then $X_{i,j} = 4, 5, 6, 7$.

Table 4.1: One-step Transition Matrix

Prior State($X_{i,j-1}$)	Current State($X_{i,j}$)							Total
	Normal	Amnestic MCI	Mixed MCI	MCI	Dementia	Dropout	Death	
Normal	2634(69.1)	524(13.8)	464(12.2)	40(1.1)	15(0.4)	33(0.9)	101(2.7)	3811(100)
Amnestic MCI	497(57.6)	172(19.9)	129(15.0)	23(2.7)	9(1.0)	13(1.5)	20(2.3)	863(100)
Mixed MCI	404(30.7)	97(7.4)	601(45.7)	66(5.0)	35(2.7)	30(2.3)	80(6.2)	1313(100)
MCI	/	/	/	154(61.4)	50(19.9)	16(6.4)	31(12.4)	251(100)
Dementia	/	/	/	/	/	/	/	
Dropout	/	/	/	/	/	/	/	
Death	/	/	/	/	/	/	/	

Note: The entries are occurrence (percentage).

RR-MLRfMC in the Application

The eight covariates used as adjustments to the effects of APOE4 are baseline age (centered at age 72), family history of dementia, self reported high blood pressure (high BP), self reported head injury (head injury), low education, and cigarette smoking level (none versus < 10, 11-19, and ≥ 20 pack years). Dimension reduction is appropriate to consider here because the covariates are established risk factors for dementia or mortality, and they are not confounders of APOE's association with dementia because they are not causes of APOE. In modeling, we use prior state as references. That is, $X_{ij} = 1, 2, 3, 4$ for $X_{i,j-1} = 1, 2, 3, 4$ as references respectively, then the log likelihood for each observation,L_{ij}, is

$$\begin{cases} -\log(1 + \sum_{l=1, l\neq X_{i,j-1}}^{7} e^{\mathbf{z}_i^T \boldsymbol{\lambda}_{X_{i,j-1},l}}) + (1 - \delta_{X_{i,j-1},X_{ij}})\mathbf{z}_i^T \boldsymbol{\lambda}_{X_{i,j-1},X_{ij}}, & \text{if } X_{i,j-1} = 1, 2, 3 \\ -\log(1 + \sum_{l=5}^{7} e^{\mathbf{z}_i^T \boldsymbol{\lambda}_{4,l}}) + (1 - \delta_{4,X_{ij}})\mathbf{z}_i^T \boldsymbol{\lambda}_{4,X_{ij}}, & \text{if } X_{i,j-1} = 4, \end{cases}$$

where $\boldsymbol{\lambda}_{u,l}$ is a 8 dimensional vector and it determines the effect of the covariates on the transition from state u to l ($l = 1, \cdots, u-1, u+1, \cdots, 7$ if $u = 1, 2, 3$ and $l > 4$ if $u = 4$). Then, our target is $\boldsymbol{\beta} = (\boldsymbol{\lambda}_{12}, \cdots, \boldsymbol{\lambda}_{17}; \boldsymbol{\lambda}_{21}, \boldsymbol{\lambda}_{23}, \cdots, \boldsymbol{\lambda}_{27}; \boldsymbol{\lambda}_{31}, \boldsymbol{\lambda}_{32}, \boldsymbol{\lambda}_{34}, \cdots, \boldsymbol{\lambda}_{37}; \boldsymbol{\lambda}_{45}, \boldsymbol{\lambda}_{46}, \boldsymbol{\lambda}_{47}) = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_{21})$, a 8×21 matrix. Under the decomposition $\boldsymbol{\beta} = \mathbf{AG}$, the log likelihood function, L_{ij}, is modified to

$$\begin{cases} -\log(1 + \sum_{l=1, l\neq u}^{7} e^{\mathbf{z}_i^T \mathbf{AG}_b}) + (1 - \delta_{u,X_{ij}})\mathbf{z}_i^T \mathbf{AG}_b, & \text{if } u = 1, 2, 3 \\ -\log(1 + \sum_{l=5}^{7} e^{\mathbf{z}_i^T \mathbf{AG}_{(14+l)}}) + (1 - \delta_{4,X_{ij}})\mathbf{z}_i^T \mathbf{AG}_{(14+X_{ij})}, & \text{if } u = 4, \end{cases}$$

where $b = 6(u-1) + l$ for $u > l$ and $b = 6(u-1) - u + l$ for $u < l$. To use Newton-Raphson Method, we need the first and second derivatives of $\ln(\mathbf{A}, \mathbf{G})$ w.r.t $\mathbf{A}$. We refer to the corresponding derivatives in Appendix A1.

Given the log-likelihood function and both derivatives, we use Algorithm 4.2.1 to get the estimated $\hat{\mathbf{A}}$, $\hat{\mathbf{G}}$ and $\hat{\boldsymbol{\beta}}$. In terms of determining the initial $\mathbf{G}$, we first fit four individual MLR without intercepts, one MLR for each possible prior state (normal cognition, A-MCI, M-MCI, and MCI), and get $\boldsymbol{\beta}_0$. Then, let $\boldsymbol{\beta}_0 = UDV^T$ by singular value decomposition and choose $\mathbf{G}_0 = D^{1/2}V^T$ as initial value, where D is a diagonal matrix with the non-zero singular values of $\boldsymbol{\beta}_0$ along its diagonal and V^T has rows

that are the right singular vectors. This type of initial $\mathbf{G}$ highly improved the speed of convergence compared to some random generated initial $\mathbf{G}$.

Another point is that the algorithm may fail to converge if we directly use $\mathbf{A}$ and $\mathbf{G}$ in each iteration due to the non-uniqueness of $\mathbf{A}$ and $\mathbf{G}$. In our computation, we use $\mathbf{A} = UD^{1/2}$ and $\mathbf{G} = D^{1/2}V^T$, where $\boldsymbol{\beta} = UDV^T$ during each iteration. This modification reduces the time and improves the likelihood that convergence is attained.

To get the 8×21 matrix, $\boldsymbol{\beta}$, we iteratively fit the RR-MLRfMC with rank $T = 1, 2, \cdots, 8$ to the dataset. The partial results are reported in Table 4.2. From Table 4.2, we see that $T = 2$ results in the smallest AIC. Since the purpose of this project is to investigate the effect of APOE4 as a fixed term, we do not report further results.

Table 4.2: Fit statistics based on 8 adjusting covariates (no intercept)

Rank T	Log-likelihood	Number of Parameters	AIC
1	-8010.84	28	16077.68
2	-7909.92	54	**15927.85**
3	-7890.74	78	15937.48
4	-7873.61	100	15947.22
5	-7861.91	120	15963.83
6	-7853.48	138	15982.97
7	-7849.65	154	16007.30
8	-7847.93	168	16031.87

PRR-MLRfMC in the Application

In this section, in addition to the 8 covariates mentioned in Section 4.3, we consider a study risk factor, APOE4, that is not involved in the dimension reduction.

Given the observations as in Section 4.3, the log-likelihood L_{ij} for each observation is, for $X_{i,j-1} = 1, 2, 3$,

$$-\log(1+\sum_{l=1,l\neq X_{i,j-1}}^{7} e^{\mathbf{w}_i^T\mathbf{D}_{X_{i,j-1},l}+\mathbf{z}_i^T\boldsymbol{\lambda}_{X_{i,j-1},l}})+(1-\delta_{X_{i,j-1},X_{ij}})(\mathbf{w}_i^T\mathbf{D}_{X_{i,j-1},X_{ij}}+\mathbf{z}_i^T\boldsymbol{\lambda}_{X_{i,j-1},X_{ij}})$$

and for $X_{i,j-1} = 4$,

$$-\log(1+\sum_{l=5}^{7} e^{\mathbf{w}_i^T\mathbf{D}_{4,l}+\mathbf{z}_i^T\boldsymbol{\lambda}_{4,l}}) + (1-\delta_{4,X_{ij}})(\mathbf{w}_i^T\mathbf{D}_{4,X_{ij}} + \mathbf{z}_i^T\boldsymbol{\lambda}_{4,X_{ij}}).$$

The goal now is to get $\mathbf{D}$, $q \times 21$ matrix and $\boldsymbol{\beta}$, $p \times 21$ matrix.

With the partial reduced-rank idea involved, the log-likelihood L_{ij} for each observation is modified as, for $X_{i,j-1} = 1, 2, 3$, $-\log(1 + \sum_{l=1, l \neq X_{i,j-1}}^{7} e^{\mathbf{w}_i^T \mathbf{D}_b + \mathbf{z}_i^T \mathbf{A}\mathbf{G}_b}) + (1 - \delta_{X_{i,j-1}, X_{ij}})(\mathbf{w}_i^T \mathbf{D}_b + \mathbf{z}_i^T \mathbf{A}\mathbf{G}_b)$, where $b = 6(X_{i,j-1} - 1) + l$ for $X_{i,j-1} > l$ and $b = 6(X_{i,j-1} - 1) - u + l$ for $X_{i,j-1} < l$; and for $X_{i,j-1} = 4$, $-\log(1 + \sum_{l=5}^{7} e^{\mathbf{w}_i^T \mathbf{D}_{14+l} + \mathbf{z}_i^T \mathbf{A}\mathbf{G}_{14+l}}) + (1 - \delta_{4, X_{ij}})(\mathbf{w}_i^T \mathbf{D}_{(14+l)} + \mathbf{z}_i^T \mathbf{A}\mathbf{G}_{(14+l)})$.

To optimize the log-likelihood function, the Algorithm 2 in Section 4.2 which involves six steps is used. In the zig step, we update $\mathbf{A}$ with fixed $\mathbf{D}$ and $\mathbf{G}$ by Newton-Raphson's method; in the zag step, we update $\mathbf{D}$ and $\mathbf{G}$ by fixing $\mathbf{A}$ which is easily obtained by using standard software to fit four individual multinomial logistic regressions with $(\mathbf{w}^T, \mathbf{A}^T\mathbf{z})^T$ as covariates and stratified by 4 prior states.

In this section we now measure the effect of APOE4 adjusted for the 8 covariates in Section 4.3. Note that the indicator for APOE4 is not involved in the reduced rank data. In addition, we include the intercept as a fixed term in the model. The optimum value of T is determined by Table 4.3.

Table 4.3: Fit statistics using 8 adjusting covariates, fixed intercept, and APOE4

Rank T	Log-likelihood	Number of Parameters	AIC
1	–6800.70	70	13741.39
2	-6768.80	96	**13729.60**
3	-6746.20	120	13732.40
4	-6732.92	142	13749.85
5	-6722.19	162	13768.38
6	-6713.80	180	13787.60
7	-6710.73	196	13813.46
8	-6709.02	210	13838.03

We choose $T = 2$ since rank 2 is the best fit to the data. The estimates of the elements of $\mathbf{A}$ and $\mathbf{G}$ are listed in Tables 4.4–4.7. While there is not much interest in interpreting the eight adjusting covariates in the reduced rank model, there is interest in interpreting the effect of APOE 4 on each one step transition. To this end we suggest using the standard errors and p values for the beta coefficients associated with APOE 4 obtained from the last iteration of the Zag step. In this instance

APOE 4 is found to be a significant predictor of a transition from normal cognition to dementia ($P = 0.029$) and from mixed MCI to Normal ($P = 0.011$).

Table 4.4: Parameter Estimates for Rank 2 Model with Normal as Prior State

Covariates	A-MCI	M-MCI	MCI	Dementia	Dropout	Death	$\mathbf{A}_1$	$\mathbf{A}_2$
intercept	-1.471	-1.582	-4.568	-5.371	-3.740	-5.264		
se	0.077	0.081	0.294	0.460	0.196	0.369		
P	0.000	0.000	0.000	0.000	0.000	0.000		
APOE4	-0.184	-0.225	0.428	1.142	-0.314	0.075		
se	0.109	0.119	0.328	0.523	0.249	0.376		
P	0.092	0.058	0.192	**0.029**	0.208	0.841		
family history	-0.047	-0.100	-0.033	-0.177	-0.179	0.001	0.003	0.226
high BP	-0.022	-0.016	0.271	-0.383	0.431	0.698	-0.473	-0.115
$<$ 10 pack years	-0.138	-0.269	0.146	-0.776	-0.091	0.595	-0.393	0.478
11-19 pack years	-0.026	-0.043	0.093	-0.213	0.102	0.272	-0.183	0.038
$\geq$ 20 pack years	0.021	0.087	0.419	-0.346	0.806	0.986	-0.672	-0.411
low education	0.170	0.348	-0.032	0.804	0.379	-0.373	0.240	-0.705
baseline age	0.034	0.072	0.021	0.130	0.125	-0.006	0.001	-0.161
head injury	-0.002	0.014	0.168	-0.185	0.296	0.413	-0.281	-0.120
	0.097	0.142	-0.538	1.003	-0.721	-1.482		
	-0.209	-0.446	-0.138	-0.798	-0.783	0.025		
	$\mathbf{G}_1$	$\mathbf{G}_2$	$\mathbf{G}_3$	$\mathbf{G}_4$	$\mathbf{G}_5$	$\mathbf{G}_6$		

Table 4.5: Parameter Estimates for Rank 2 Model with A-MCI as Prior State

Covariates	Normal	M-MCI	MCI	Dementia	Dropout	Death	$\mathbf{A}_1$	$\mathbf{A}_2$
intercept	1.172	-0.442	-1.784	-3.365	-2.617	-2.685		
se	0.144	0.195	0.343	0.620	0.434	0.496		
P	0.000	0.023	0.000	0.000	0.000	0.000		
APOE4	0.050	0.232	0.459	0.897	0.001	-1.393		
se	0.205	0.262	0.473	0.695	0.547	1.056		
P	0.807	0.375	0.332	0.197	0.999	0.187		
family history	0.049	-0.028	-0.013	-0.077	-0.040	0.004	0.003	0.226
high BP	-0.152	0.073	-0.431	0.015	0.375	0.274	-0.473	-0.115
$<$ 10 pack years	-0.004	-0.009	-0.400	-0.183	0.215	0.243	-0.393	0.478
11-19 pack years	-0.041	0.018	-0.173	-0.022	0.131	0.108	-0.183	0.038
$\geq$ 20 pack years	-0.268	0.134	-0.595	0.106	0.574	0.383	-0.672	-0.411
low education	-0.086	0.055	0.274	0.253	-0.062	-0.159	0.240	-0.705
baseline age	-0.034	0.019	0.013	0.055	0.026	-0.005	0.001	-0.161
head injury	-0.101	0.050	-0.252	0.027	0.231	0.161	-0.281	-0.120
	0.269	-0.126	0.929	0.051	-0.751	-0.586		
	0.213	-0.122	-0.072	-0.342	-0.168	0.025		
	$\mathbf{G}_7$	$\mathbf{G}_8$	$\mathbf{G}_9$	$\mathbf{G}_{10}$	$\mathbf{G}_{11}$	$\mathbf{G}_{12}$		

Table 4.6: Parameter Estimates for Rank 2 Model with M-MCI as Prior State

Covariates	Normal	A-MCI	MCI	Dementia	Dropout	Death	$\mathbf{A}_1$	$\mathbf{A}_2$
intercept	-0.415	-1.920	-2.327	-3.198	-2.437	-2.966		
se	0.103	0.178	0.214	0.305	0.214	0.294		
P	0.000	0.000	0.000	0.000	0.000	0.000		
APOE4	-0.381	-0.092	0.331	0.447	-0.347	0.300		
se	0.150	0.243	0.272	0.358	0.284	0.391		
P	**0.011**	0.707	0.224	0.212	0.222	0.442		
family history	0.047	0.072	-0.020	0.009	-0.055	0.060	0.003	0.226
high BP	0.186	0.195	-0.038	0.225	0.378	-0.086	-0.473	-0.115
$<$ 10 pack years	0.278	0.350	-0.084	0.215	0.181	0.080	-0.393	0.478
11-19 pack years	0.089	0.102	-0.022	0.091	0.127	-0.011	-0.183	0.038
$\geq$ 20 pack years	0.211	0.197	-0.031	0.308	0.595	-0.188	-0.672	-0.411
low education	-0.258	-0.348	0.089	-0.151	-0.014	-0.158	0.240	-0.705
baseline age	-0.035	-0.053	0.015	-0.008	0.037	-0.042	0.001	-0.161
head injury	0.099	0.099	-0.017	0.131	0.237	-0.065	-0.281	-0.120
	-0.445	-0.492	0.102	-0.488	-0.743	0.117		
	0.214	0.327	-0.091	0.048	-0.234	0.265		
	$\mathbf{G}_{13}$	$\mathbf{G}_{14}$	$\mathbf{G}_{15}$	$\mathbf{G}_{16}$	$\mathbf{G}_{17}$	$\mathbf{G}_{18}$		

Table 4.7: Parameter Estimates for Rank 2 Model with MCI as Prior State

Covariates	Normal	A-MCI	M-MCI	Dementia	Dropout	Death	$\mathbf{A}_1$	$\mathbf{A}_2$
intercept	/	/	/	-0.809	-2.108	-1.890		
se	/	/	/	0.282	0.393	0.442		
P	/	/	/	0.004	0.000	0.000		
APOE4	/	/	/	0.585	-0.007	0.043		
se	/	/	/	0.340	0.430	0.570		
P	/	/	/	0.085	0.987	0.940		
family history	/	/	/	-0.003	-0.056	0.045	0.003	0.226
high BP	/	/	/	-0.598	0.348	-0.384	-0.473	-0.115
$<$ 10 pack years	/	/	/	-0.516	0.152	-0.211	-0.393	0.478
11-19 pack years	/	/	/	-0.234	0.115	-0.133	-0.183	0.038
$\geq$ 20 pack years	/	/	/	-0.841	0.554	-0.592	-0.672	-0.411
low education	/	/	/	0.328	0.007	0.051	0.240	-0.705
baseline age	/	/	/	0.007	0.038	-0.029	0.001	-0.161
head injury	/	/	/	-0.353	0.219	-0.238	-0.281	-0.120
	/	/	/	1.272	-0.678	0.766		
	/	/	/	-0.032	-0.240	0.188		
	/	/	/	$\mathbf{G}_{19}$	$\mathbf{G}_{20}$	$\mathbf{G}_{21}$		

4.4 Discussion

In this project we show how to study the effect of a risk factor for transitions in a complex Markov chain adjusted for the presence of a set of covariates after subjecting the effect of the covariates to dimension reduction. The methodology assumes that transitions within each row of the one step transition matrix are governed by a multinomial logistic regression model. Extensions to other models are possible such as a probit model or a multinomial model with random effects (see e.g. Salazar et al [48], or Song et al. [55]) although the latter would involve additional numerical integration steps. The methodology is useful since it is not clear how to adjust for covariates in a complex chain which are often characterized by select transitions being relatively rare as illustrated in Table 4.1. In the example chosen to illustrate this methodology the number of unknown parameters to be estimated and associated with the eight covariates was reduced by over two thirds (in Tables 4.2 and 4.3 we estimated 54 unknown parameters instead of 168). Our method is limited to homogenous Markov chains and does not apply to chains having time dependent covariates (see e.g. Yu et al.[77]).

Dimension reduction can also be achieved by applying a proportional odds model or a stereotype model[41] to each row of the one step transition matrix in Equation (2). A proportional odds model will apply only to the situation where the states of the chain are ordered while a stereotype model can be applied without this assumption. In a proportional odds model the parameter vectors λ_{ul} in Equation 4.2 depend only on the prior sate u but not on the next state l while in a stereotype model this parameter vector expresses dependence on the covariates through d latent factors where d varies from 1 to $\min(J-1, p+q)$. Both models can be fit to data using standard statistical software but require each covariate in the model to meet these assumptions. If these strategies were applied to the example discussed in Section 4.3 the number of unknown parameters to be estimated would be reduced from 210 to 57 under a proportional odds model and to $72-210$ under various stereotype models. Our solution estimated 96 unknown parameters and has the advantage that APOE

4 carrier status, the risk factor of interest, is not required to meet the proportional odds or the stereotype model assumptions. As seen in Section 4.3 the APOE 4 risk affects only a few select transitions in the chain.

The proposed methodology makes a generalized linear model into a nonlinear model by introducing a product of two unknown matrices (**A** and **G** in Section 4.3). Nonlinearity is handled by maximizing the likelihood function using an iterative zig-zag Algorithm (see e.g. Heckman [29]). The zig sub-step (**G** fixed) is nonlinear and it is possible that the iteration will not converge to a solution. However, we found that if we account for the non-uniqueness of the solutions for **A** and **G** using single value decomposition convergence is obtained. In the zag step (**A** fixed), the problem reduces to fitting a multinomial model using standard software (package **nnet** in **R**) to estimate **G** and the effect of the risk factors of interest.

The choice of initial value plays an important role regarding the convergence speed. In our numerical studies, we take advantage of the four individual multinomial regressions stratified by prior states to get the initial value for **G**. And this method greatly reduces the time to converges compared to random initial values. Another point is that, in each iteration, the higher rank model takes more time than the lower rank model. However, there is no necessary connection between the iteration numbers towards convergence and rank. See Table 4.8 for the computing time needed to construct Table 4.3.

Table 4.8: Computing time needed (in second) to get Table 3

Rank T	number of iteration	time
1	31	133.53
2	18	98.35
3	40	250.09
4	41	296.36
5	27	229.32
6	19	179.03
7	30	421.26
8	2	28.76

In this project we fit multinomial logistic models to the data and it is well known that these models have stability problems in the presence of complete/quasi complete

separation of the covariates or a collinearity among the covariates. These latter are often caused by small sample sizes and/or rare categories in the multinomial. These issues occur more frequently when evaluating the exact likelihood for the Markov chain since the latter fits individual multinomial models to each row of the one step transition matrix. Hence, each row is at risk for encountering these problems. A reduced rank regression addresses these issues by borrowing strength from neighboring rows of the one step transitions matrix in much the same way as done in an empirical Bayes method [26].

By using the AIC we assume we have chosen the correct value of T from among a finite set of possible values for T. This has consequences for the estimate of $\boldsymbol{\beta}$ since that estimator depends directly on the choice for T. The need to account for the error in estimating T when making inference for $\boldsymbol{\beta}$ is debated in the literature (see e.g. a discussion of oracle estimation in Goldberg et al.[27]).The problem is more complicated in this project since our main focus is on inference for the target parameter matrix $\boldsymbol{\theta}$ (Section 4.2 and the application). As far as we know this is not discussed in the literature and remains an open problem for further investigation.

Appendices

A Supplementary Materials for Chapter 2

A1. Formulation and Proofs

A1.1 Formulation of equation (2)

If we choose the Gaussian function $K(\omega) = (2\pi\sigma_\omega^2)^{-p/2} exp(-\frac{\| \omega \|^2}{2\sigma_\omega^2})$, then

$$
\begin{aligned}
\mathbf{M}_k &= Re(\int\int e^{\mathbf{i}\omega^T\mathbf{X}_1} d[m(\mathbf{X}_1)f(\boldsymbol{\beta}^T\mathbf{X}_1)][\int e^{-\mathbf{i}\omega^T\mathbf{X}_2} d[m(\mathbf{X}_2)f(\boldsymbol{\beta}^T\mathbf{X}_2)]]^T(2\pi\sigma_\omega^2)^{-p/2} e^{\frac{\parallel \omega \parallel^2}{-2\sigma_\omega^2}} d\omega) \\
&= Re(\int\int\int (2\pi\sigma_\omega^2)^{-p/2} e^{-\frac{\parallel \omega \parallel^2 -2\sigma_\omega^2 \mathbf{i}\omega^T(\mathbf{X}_1-\mathbf{X}_2)}{2\sigma_\omega^2}} d\omega d[m(\mathbf{X}_1)f(\boldsymbol{\beta}^T\mathbf{X}_1)] d[m(\mathbf{X}_2)f(\boldsymbol{\beta}^T\mathbf{X}_2)]^T) \\
&= \int\int e^{-\frac{\sigma_\omega^2 \parallel \mathbf{X}_1-\mathbf{X}_2 \parallel^2}{2}} d[m(\mathbf{X}_1)f(\boldsymbol{\beta}^T\mathbf{X}_1)] d[m(\mathbf{X}_2)f(\boldsymbol{\beta}^T\mathbf{X}_2)]^T \\
&= \int [-\int m(\mathbf{X}_1)f(\boldsymbol{\beta}^T\mathbf{X}_1) d(e^{-\frac{\sigma_\omega^2 \parallel \mathbf{X}_1-\mathbf{X}_2 \parallel^2}{2}})] d[m(\mathbf{X}_2)f(\boldsymbol{\beta}^T\mathbf{X}_2)]^T \\
&= -\int\int m(\mathbf{X}_1)f(\boldsymbol{\beta}^T\mathbf{X}_1)[m(\mathbf{X}_2)f(\boldsymbol{\beta}^T\mathbf{X}_2)]^T d[e^{-\frac{\sigma_\omega^2 \parallel \mathbf{X}_1-\mathbf{X}_2 \parallel^2}{2}} \sigma_\omega^2(\mathbf{X}_1-\mathbf{X}_2)]^T d\mathbf{X}_1 \\
&= \int\int m(\mathbf{X}_1)f(\boldsymbol{\beta}^T\mathbf{X}_1)[m(\mathbf{X}_2)f(\boldsymbol{\beta}^T\mathbf{X}_2)]^T e^{-\frac{\sigma_\omega^2 \parallel \mathbf{X}_1-\mathbf{X}_2 \parallel^2}{2}} \sigma_\omega^2[\mathbf{I}_p - \sigma_\omega^2(\mathbf{X}_1-\mathbf{X}_2)(\mathbf{X}_1-\mathbf{X}_2)^T] d\mathbf{X}_1 d\mathbf{X}_2 \\
&= E_{(\mathbf{X}_1,Y_1),(\mathbf{X}_2,Y_2)} Y_1 Y_2 \frac{f(\boldsymbol{\beta}^T\mathbf{X}_1)}{f(\mathbf{X}_1)} \frac{f(\boldsymbol{\beta}^T\mathbf{X}_2)}{f(\mathbf{X}_2)} e^{-\frac{\sigma_\omega^2 \parallel \mathbf{X}_1-\mathbf{X}_2 \parallel^2}{2}} \sigma_\omega^2[\mathbf{I}_p - \sigma_\omega^2(\mathbf{X}_1-\mathbf{X}_2)(\mathbf{X}_1-\mathbf{X}_2)^T] \\
&= E_{(\mathbf{X}_1,Y_1),(\mathbf{X}_2,Y_2)} Y_1 Y_2 \frac{1}{f(\boldsymbol{\beta}_0^T\mathbf{X}_1)} \frac{1}{f(\boldsymbol{\beta}_0^T\mathbf{X}_2)} e^{-\frac{\sigma_\omega^2 \parallel \mathbf{X}_1-\mathbf{X}_2 \parallel^2}{2}} \sigma_\omega^2[\mathbf{I}_p - \sigma_\omega^2(\mathbf{X}_1-\mathbf{X}_2)(\mathbf{X}_1-\mathbf{X}_2)^T] \\
&\propto E_{(\mathbf{X}_1,Y_1),(\mathbf{X}_2,Y_2)} Y_1 Y_2 e^{-\frac{\sigma_\omega^2 \parallel \mathbf{X}_1-\mathbf{X}_2 \parallel^2}{2}} \sigma_\omega^2[\mathbf{I}_p - \sigma_\omega^2(\mathbf{X}_1-\mathbf{X}_2)(\mathbf{X}_1-\mathbf{X}_2)^T].
\end{aligned}
$$

□

Comment: This is the formulation for an univariate response Y. For a multivariate response $\mathbf{Y}$, all the steps are the same, except start from the ninth equation, we will have $\mathbf{Y}_1^T\mathbf{Y}_2$ instead of Y_1Y_2.

A1.2 Proofs

Proof of lemma 1

Given $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_d)$, a basis of $\mathcal{S}_{E(Y|\mathbf{X})}$, we have $m(\mathbf{X}) = E(Y|\mathbf{X}) = E(Y|\boldsymbol{\beta}^T\mathbf{X})$ $= m(\boldsymbol{\beta}^T\mathbf{X})$. Then,

$$\begin{aligned}
&\frac{d[m(\mathbf{X})f(\boldsymbol{\beta}^T\mathbf{X})]}{d\mathbf{X}} = \frac{d[m(\boldsymbol{\beta}^T\mathbf{X})f(\boldsymbol{\beta}^T\mathbf{X})]}{d\mathbf{X}} \\
&= \frac{d(\boldsymbol{\beta}^T\mathbf{X})^T}{d\mathbf{X}}(m'(\boldsymbol{\beta}^T\mathbf{X})f(\boldsymbol{\beta}^T\mathbf{X}) + m(\boldsymbol{\beta}^T\mathbf{X})f'(\boldsymbol{\beta}^T\mathbf{X})) \\
&= \boldsymbol{\beta}(m'(\boldsymbol{\beta}^T\mathbf{X})f(\boldsymbol{\beta}^T\mathbf{X}) + m(\boldsymbol{\beta}^T\mathbf{X})f'(\boldsymbol{\beta}^T\mathbf{X})).
\end{aligned}$$

Thus, the gradient of $[m(\mathbf{X})f(\boldsymbol{\beta}^T\mathbf{X})]$ is a linear combination of $\boldsymbol{\beta}$, and it is in $\mathcal{S}_{E(Y|\mathbf{X})}$. Let $\text{supp}(\mathbf{X}) = \{\mathbf{x} \in \mathbb{R}^p : f(\mathbf{x}) > 0\}$ be the support of $\mathbf{X}$. To show $\mathcal{S}_{E(Y|\mathbf{X})}$ is fully spanned by the collection of all of the gradient of $m(\mathbf{X})f(\boldsymbol{\beta}^T\mathbf{X})$ over $\mathbf{X} \in \text{supp}(\mathbf{X})$, we show that $\alpha^T\boldsymbol{\beta} = \mathbf{0}$ is equivalent to $\alpha^T\frac{d[m(\mathbf{X})f(\boldsymbol{\beta}^T\mathbf{X})]}{d\mathbf{X}} = 0$ for $\alpha \in \mathbb{R}^p$.

$\Rightarrow$ If $\alpha^T\boldsymbol{\beta} = \mathbf{0}$, then $\alpha^T\frac{d[m(\mathbf{X})f(\boldsymbol{\beta}^T\mathbf{X})]}{d\mathbf{X}} = \alpha^T\boldsymbol{\beta}\frac{d[m(\boldsymbol{\beta}^T\mathbf{X})f(\boldsymbol{\beta}^T\mathbf{X})]}{d\boldsymbol{\beta}^T\mathbf{X}} = 0$ for all $\mathbf{X} \in$ $\text{supp}(\mathbf{X})$.

$\Leftarrow$ We prove this direction by contradiction. Assume there is an α_0 such that $\alpha_0^T\frac{d[m(\mathbf{X})f(\boldsymbol{\beta}^T\mathbf{X})]}{d\mathbf{X}} = 0$ for all $\mathbf{X} \in \text{supp}(\mathbf{X})$ and $\alpha_0^T\boldsymbol{\beta} \neq \mathbf{0}$. We denote $\frac{\alpha_0^T\boldsymbol{\beta}}{||\alpha_0^T\boldsymbol{\beta}||} = \gamma_1^T \in$ $\mathbb{R}^d$ with $\gamma_1 \neq \mathbf{0}$. By the assumption, we have $\gamma_1^T\frac{d[m(\boldsymbol{\beta}^T\mathbf{X})f(\boldsymbol{\beta}^T\mathbf{X})]}{d\boldsymbol{\beta}^T\mathbf{X}} = 0$. It means the directional derivative of $m(\boldsymbol{\beta}^T\mathbf{X})f(\boldsymbol{\beta}^T\mathbf{X})$ is 0 along γ_1 and $m(\boldsymbol{\beta}^T\mathbf{X})f(\boldsymbol{\beta}^T\mathbf{X})$ is a constant along γ_1. Now, we can expand γ_1 to $\boldsymbol{\Gamma} = (\gamma_1, \cdots, \gamma_d)$ such that $\boldsymbol{\Gamma}$ forms an orthonormal basis for $\mathbb{R}^d$. Let $\mathbf{V} = \boldsymbol{\Gamma}^T\boldsymbol{\beta}^T\mathbf{X} = (\mathbf{v}_1, \cdots, \mathbf{v}_d)^T$. Then, we have $\boldsymbol{\beta}^T\mathbf{X} = \boldsymbol{\Gamma}\mathbf{V}$ and $m(\boldsymbol{\beta}^T\mathbf{X})f(\boldsymbol{\beta}^T\mathbf{X}) = m(\boldsymbol{\Gamma}\mathbf{V})f(\boldsymbol{\Gamma}\mathbf{V})$. Now, $\frac{d[m(\boldsymbol{\Gamma}\mathbf{V})f(\boldsymbol{\Gamma}\mathbf{V})]}{d\mathbf{v}_1} =$ $\frac{d[\boldsymbol{\Gamma}\mathbf{V}]}{d\mathbf{v}_1}\frac{d[m(\boldsymbol{\beta}^T\mathbf{X})f(\boldsymbol{\beta}^T\mathbf{X})]}{d\boldsymbol{\beta}^T\mathbf{X}} = \gamma_1^T\frac{d[m(\boldsymbol{\beta}^T\mathbf{X})f(\boldsymbol{\beta}^T\mathbf{X})]}{d\boldsymbol{\beta}^T\mathbf{X}} = 0$. It means $m(\boldsymbol{\Gamma}\mathbf{V})f(\boldsymbol{\Gamma}\mathbf{V})$ does not depend on $\mathbf{v}_1$ and we have $m(\boldsymbol{\beta}^T\mathbf{X})f(\boldsymbol{\beta}^T\mathbf{X}) = m(\boldsymbol{\Gamma}\mathbf{V})f(\boldsymbol{\Gamma}\mathbf{V}) = \tilde{m}(\mathbf{v}_2, \cdots, \mathbf{v}_d) =$ $\tilde{m}((\gamma_2^T, \cdots, \gamma_d^T)\boldsymbol{\beta}^T\mathbf{X})$. It means that $\boldsymbol{\beta}(\gamma_2, \cdots, \gamma_d)$ is also a dimension reduction basis matrix for $\mathcal{S}_{E(Y|\mathbf{X})}$ and has dimension at most $d-1$. It contradicts to the given con-

dition that $\boldsymbol{\beta}$ has d columns. Thus, given $\alpha_0^T \frac{d[m(\mathbf{X})f(\boldsymbol{\beta}^T\mathbf{X})]}{d\mathbf{X}} = 0$ for all $\mathbf{X} \in \text{supp}(\mathbf{X})$, we have $\alpha_0^T\boldsymbol{\beta} = \mathbf{0}$.

We have proved that $\mathcal{S}_{E(Y|\mathbf{X})} = \texttt{span}\{\frac{d[m(\mathbf{X})f(\boldsymbol{\beta}^T\mathbf{X})]}{d\mathbf{X}}, \mathbf{X} \in \texttt{supp}(\mathbf{X})\}$. By the facts that Fourier transformation is one-to-one and $\frac{d[m(\mathbf{X})f(\boldsymbol{\beta}^T\mathbf{X})]}{d\mathbf{X}} = (2\pi)^{-1}\int e^{-\mathbf{i}\omega^T\mathbf{X}}\psi(\omega)d\omega$, we have $\mathcal{S}_{E(Y|\mathbf{X})} = \texttt{span}\{\psi(\omega), \omega \in \mathbb{R}^p\}$.

□

Proof of lemma 2

Assume $\boldsymbol{\beta}$ is a basis of $\mathcal{S}_{E(Y|\mathbf{X})}$, $(\boldsymbol{\beta}, \boldsymbol{\beta}_0)$ forms an orthogonal matrix and $(Y, \boldsymbol{\beta}^T\mathbf{X})$ is independent to $\boldsymbol{\beta}_0^T\mathbf{X}$.

1. Given the above assumptions, we have

$$\begin{aligned}
\psi(\omega) &= -E(\mathbf{i}\omega e^{\mathbf{i}\omega^T\mathbf{X}}Y\frac{f(\boldsymbol{\beta}^T\mathbf{X})}{f(\mathbf{X})}) \\
&= -E(\mathbf{i}\omega e^{\mathbf{i}\omega^T P_{\boldsymbol{\beta}}\mathbf{X}+\mathbf{i}\omega^T P_{\boldsymbol{\beta}_0}\mathbf{X}}Y\frac{1}{f(\boldsymbol{\beta}_0^T\mathbf{X})}) \\
&= -E(\mathbf{i}\omega e^{\mathbf{i}\omega^T P_{\boldsymbol{\beta}}\mathbf{X}}Y)E(e^{\mathbf{i}\omega^T P_{\boldsymbol{\beta}_0}\mathbf{X}}\frac{1}{f(\boldsymbol{\beta}_0^T\mathbf{X})}) \\
&= -E(\mathbf{i}\omega e^{\mathbf{i}\omega^T P_{\boldsymbol{\beta}}\mathbf{X}}Y)E(e^{\mathbf{i}\omega^T P_{\boldsymbol{\beta}_0}\mathbf{X}})E(e^{\mathbf{i}\omega^T P_{\boldsymbol{\beta}_0}\mathbf{X}})^{-1}E(e^{\mathbf{i}\omega^T P_{\boldsymbol{\beta}_0}\mathbf{X}}\frac{1}{f(\boldsymbol{\beta}_0^T\mathbf{X})}) \\
&= -E(\mathbf{i}\omega e^{\mathbf{i}\omega^T\mathbf{X}}Y)E(e^{\mathbf{i}\omega^T P_{\boldsymbol{\beta}_0}\mathbf{X}})^{-1}E(e^{\mathbf{i}\omega^T P_{\boldsymbol{\beta}_0}\mathbf{X}}\frac{1}{f(\boldsymbol{\beta}_0^T\mathbf{X})}) \\
&= k_0 E(\omega e^{\mathbf{i}\omega^T\mathbf{X}}Y) \\
&\propto E(\omega e^{\mathbf{i}\omega^T\mathbf{X}}Y) \\
&= \omega E(e^{\mathbf{i}\omega^T\mathbf{X}}Y),
\end{aligned}$$

   where $k_0 = -\mathbf{i}E(e^{\mathbf{i}\omega^T P_{\boldsymbol{\beta}_0}\mathbf{X}})^{-1}E(e^{\mathbf{i}\omega^T P_{\boldsymbol{\beta}_0}\mathbf{X}}\frac{1}{f(\boldsymbol{\beta}_0^T\mathbf{X})})$ is a constant. Then, we can use the ωs that have large $|E(e^{\mathbf{i}\omega^T\mathbf{X}}Y)|$ to recover the CMS.

2. Let $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$ be notations defined as before. Then, there is a rotation matrix $\mathbf{Q}$ such that $\boldsymbol{\eta}\mathbf{Q} = (\boldsymbol{\beta}_a, \boldsymbol{\beta}_b)$ with $\mathcal{S}(\boldsymbol{\beta}_a) \subseteq \mathcal{S}(\boldsymbol{\beta})$ and $\mathcal{S}(\boldsymbol{\beta}_b) \subseteq \mathcal{S}(\boldsymbol{\beta}_0)$. Then, given the above assumptions, let $\mathbf{W}_1 = (\mathbf{X}^T\boldsymbol{\beta}_a, \mathbf{0})^T$, $V_1 = Y$, $\mathbf{W}_2 = (\mathbf{0}, \mathbf{X}^T\boldsymbol{\beta}_b)^T$, $V_2 = 0$ and we know $(\mathbf{W}_1, V_1) \perp (\mathbf{W}_2, V_2)$. By the following 3 claims, we have $\text{MDD}(Y|\boldsymbol{\eta}^T\mathbf{X})^2 = \text{MDD}(Y|\mathbf{Q}^T\boldsymbol{\eta}^T\mathbf{X})^2 < \text{MDD}(Y|\boldsymbol{\beta}_a^T\mathbf{X})^2 \leq \text{MDD}(Y|\boldsymbol{\beta}^T\mathbf{X})^2$.

The first equation holds because $\text{MDD}(Y|\mathbf{X})^2 = -E[(Y-E(Y))(Y'-E(Y'))|\mathbf{X}-\mathbf{X}'|_p]$ and it is easy to see that it's invariant under rotation transformation [51]. The first inequality is true due to claim 3. The last inequality is true because of claim 2.

Claim 1: Suppose $\boldsymbol{\beta}$ is a d dimensional basis of $\mathcal{S}_{E(Y|\mathbf{X})}$, and $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ is any partition of $\boldsymbol{\beta}$. Then, $\text{MDD}(Y|\boldsymbol{\beta}_i^T\mathbf{X})^2 < \text{MDD}(Y|\boldsymbol{\beta}^T\mathbf{X})^2$, $i = 1, 2$.

Proof: Let $\mathbf{X}_1 = \boldsymbol{\beta}_1^T\mathbf{X}$, $\mathbf{X}_2 = \boldsymbol{\beta}_2^T\mathbf{X}$, $G(a,b) = \text{MDD}(Y|(a\mathbf{X}_1^T, b\mathbf{X}_2^T)^T)^2$ with $a, b \in \mathbb{R}$ and $g_1(a,b) = \dfrac{\partial G(a,b)}{\partial a}, g_2(a,b) = \dfrac{\partial G(a,b)}{\partial b}$. After simple computation, we have $G(a,b) = ag_1(a,b) + bg_2(a,b)$, $g_1(ca,cb) = g_1(a,b)$ and $g_2(ca,cb) = g_2(a,b)$ for any $c > 0$. $G(a,b) = 0$ if and only if $E(Y) = E(Y|(a\mathbf{X}_1^T, b\mathbf{X}_2^T)^T)$ almost surely [51]. In addition, because $\text{MDD}(Y|c\mathbf{X})^2 = c\text{MDD}(Y|\mathbf{X})^2$, $G(ca,cb) = cG(a,b)$[51].

Given $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \in \mathcal{S}(\boldsymbol{\beta})$, we have $G(0,1) > 0$ and $G(1,0) > 0$. For any $\gamma \in [0,1)$, $G(1,\gamma) < G(1,1)$ and $G(\gamma,1) < G(1,1)$. If it is not true, then there must be a $\gamma_0 \in [0,1)$ such that at least one of $G(1,\gamma_0) \geq G(1,1)$ and $G(\gamma_0,1) \geq G(1,1)$ holds. Without loss of generality, we only need to show that the statement "there exists a $\gamma_0 \in [0,1)$ such that $G(1,\gamma_0) \geq G(1,1)$" is wrong. Because $G(1,\gamma) = \gamma G(1/\gamma,1)$, we have $G(1/\gamma,1) \to G(0,1) > 0$ and $G(1,\gamma) \to \infty$ as $\gamma \to \infty$. It means there exists a $\gamma_1 \in (\gamma_0, \infty)$ such that $G(1,\gamma_1)$ is minimized. Then, it's partial derivative $g_2(1,\gamma_1) = 0$. Because $G(1,\gamma) = \gamma G(1/\gamma,1)$, then $g_1(1,\gamma_1) = g_1(1/\gamma_1,1) = 0$. Thus $G(1,\gamma_1) = g_1(1,\gamma_1) + \gamma_1 g_2(1,\gamma_1) = 0$. It means that $E(Y) = E(Y|(\mathbf{X}_1^T, \gamma_1\mathbf{X}_2^T)^T)$ almost surely, which contradicts to the assumption that $\boldsymbol{\beta}$ forms a d dimensional basis for $\mathcal{S}_{E(Y|\mathbf{X})}$. Therefore, $G(0,1) < G(1,1)$ and $G(1,0) < G(1,1)$ and they are equivalent to $\text{MDD}(Y|\boldsymbol{\beta}_i^T\mathbf{X})^2 < \text{MDD}(Y|\boldsymbol{\beta}^T\mathbf{X})^2$, $i = 1, 2$.

Claim 2: Suppose $\boldsymbol{\beta}$ is a d dimensional basis of $\mathcal{S}_{E(Y|\mathbf{X})}$, and $\boldsymbol{\eta}$ is a $p \times d_2$ matrix with $d_2 \leq d$ and $\mathcal{S}(\boldsymbol{\eta}) \subseteq \mathcal{S}(\boldsymbol{\beta})$. Then, $\text{MDD}(Y|\boldsymbol{\eta}^T\mathbf{X})^2 \leq \text{MDD}(Y|\boldsymbol{\beta}^T\mathbf{X})^2$ and the equality holds when $\mathcal{S}(\boldsymbol{\eta}) = \mathcal{S}(\boldsymbol{\beta})$.

Proof: Given $\mathcal{S}(\boldsymbol{\eta}) \subseteq \mathcal{S}(\boldsymbol{\beta})$ and $d_2 \leq d$, there is a matrix $\mathbf{A}$ such that

$\boldsymbol{\eta} = \boldsymbol{\beta}\mathbf{A}$. We can then apply singular value decomposition to $\mathbf{A}$ and have $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ with $\mathbf{U}$ and $\mathbf{V}$ are $d \times d$ and $d_2 \times d_2$ orthogonal matrices and $\mathbf{D}$ is a $d \times d_2$ diagonal matrix with 1 along its diagonal. Then, $\text{MDD}(Y|\boldsymbol{\eta}^T\mathbf{X})^2 = \text{MDD}(Y|\mathbf{V}\mathbf{D}^T\mathbf{U}^T\boldsymbol{\beta}^T\mathbf{X})^2 = \text{MDD}(Y|\mathbf{D}^T\mathbf{U}^T\boldsymbol{\beta}^T\mathbf{X})^2$. We further let $\mathbf{U}^T\boldsymbol{\beta}^T\mathbf{X} = (\tilde{X}_1, \cdots, \tilde{X}_d)$. Then, $\mathbf{D}^T\mathbf{U}^T\boldsymbol{\beta}^T\mathbf{X} = (\tilde{X}_1, \cdots, \tilde{X}_{d_2})$ and $\text{MDD}(Y|\mathbf{D}^T\mathbf{U}^T\boldsymbol{\beta}^T\mathbf{X})^2 \leq \text{MDD}(Y|\mathbf{U}^T\boldsymbol{\beta}^T\mathbf{X})^2$ by claim 1. Thus $\text{MDD}(Y|\boldsymbol{\eta}^T\mathbf{X})^2 = \text{MDD}(Y|\mathbf{V}\mathbf{D}^T\mathbf{U}^T\boldsymbol{\beta}^T\mathbf{X})^2 \leq \text{MDD}(Y|\mathbf{U}\mathbf{U}^T\boldsymbol{\beta}^T\mathbf{X})^2 = \text{MDD}(Y|\boldsymbol{\beta}^T\mathbf{X})^2$. The equality holds when $\mathcal{S}(\boldsymbol{\eta}) = \mathcal{S}(\boldsymbol{\beta})$.

Claim 3: Suppose $(\mathbf{W}_1, \mathbf{V}_1) \perp (\mathbf{W}_2, \mathbf{V}_2)$, where $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^p$ and $\mathbf{V}_1, \mathbf{V}_2 \in \mathbb{R}^q$, then $MDD((\mathbf{V}_1+\mathbf{V}_2)|(\mathbf{W}_1+\mathbf{W}_2))^2 \leq MDD(\mathbf{V}_1|\mathbf{W}_1)^2 + MDD(\mathbf{V}_2|\mathbf{W}_2)^2$.

Proof:

$$
\begin{aligned}
&MDD((\mathbf{V}_1 + \mathbf{V}_2)|(\mathbf{W}_1 + \mathbf{W}_2))^2 \\
&= \frac{1}{c_p}\int_{\mathbb{R}^p} \frac{|E(e^{\mathbf{i}\omega^T(\mathbf{W}_1+\mathbf{W}_2)}(\mathbf{V}_1 + \mathbf{V}_2)) - E(e^{\mathbf{i}\omega^T(\mathbf{W}_1+\mathbf{W}_2)})E(\mathbf{V}_1 + \mathbf{V}_2)|_q^2}{|\omega|_p^{1+p}} d\omega \\
&= \frac{1}{c_p}\int_{\mathbb{R}^p} \frac{|T_1 + T_2 - T_3 - T_4|_q^2}{|\omega|_p^{1+p}} d\omega \\
&\leq \frac{1}{c_p}\int_{\mathbb{R}^p} \frac{|T_1 - T_3|_q^2 + |T_2 - T_4|_q^2}{|\omega|_p^{1+p}} d\omega \\
&= \frac{1}{c_p}\int_{\mathbb{R}^p} \frac{|[E(e^{\mathbf{i}\omega^T\mathbf{W}_1}\mathbf{V}_1) - E(e^{\mathbf{i}\omega^T\mathbf{W}_1})E(\mathbf{V}_1)]E(e^{\mathbf{i}\omega^T\mathbf{W}_2})|_q^2}{|\omega|_p^{1+p}} d\omega \\
&+ \frac{1}{c_p}\int_{\mathbb{R}^p} \frac{|[E(e^{\mathbf{i}\omega^T\mathbf{W}_2}\mathbf{V}_2) - E(e^{\mathbf{i}\omega^T\mathbf{W}_2})E(\mathbf{V}_2)]E(e^{\mathbf{i}\omega^T\mathbf{W}_1})|_q^2}{|\omega|_p^{1+p}} d\omega \\
&\leq \frac{1}{c_p}\int_{\mathbb{R}^p} \frac{|E(e^{\mathbf{i}\omega^T\mathbf{W}_1}V_1) - E(e^{\mathbf{i}\omega^T\mathbf{W}_1})E(\mathbf{V}_1)|_q^2}{|\omega|_p^{1+p}} d\omega \\
&+ \frac{1}{c_p}\int_{\mathbb{R}^p} \frac{|E(e^{\mathbf{i}\omega^T\mathbf{W}_2}\mathbf{V}_2) - E(e^{\mathbf{i}\omega^T\mathbf{W}_2})E(\mathbf{V}_2)|_q^2}{|\omega|_p^{1+p}} d\omega \\
&= MDD(\mathbf{V}_1|\mathbf{W}_1)^2 + MDD(\mathbf{V}_2|\mathbf{W}_2)^2
\end{aligned}
$$

where $T_1 = E(e^{\mathbf{i}\omega^T\mathbf{W}_1}\mathbf{V}_1)E(e^{\mathbf{i}\omega^T\mathbf{W}_2})$, $T_2 = E(e^{\mathbf{i}\omega^T\mathbf{W}_1})E(e^{\mathbf{i}\omega^T\mathbf{W}_2}\mathbf{V}_2)$, $T_3 = E(e^{\mathbf{i}\omega^T\mathbf{W}_1})E(e^{\mathbf{i}\omega^T\mathbf{W}_2})E(\mathbf{V}_1)$ and $T_4 = E(e^{\mathbf{i}\omega^T\mathbf{W}_1})E(e^{\mathbf{i}\omega^T\mathbf{W}_2})E(\mathbf{V}_2)$. The equality holds if and only if $E(\mathbf{V}_1|\mathbf{W}_1) = E(\mathbf{V}_1)$ and $E(\mathbf{V}_2|\mathbf{W}_2) = E(\mathbf{V}_2)$.

□

Proof of theorem 2.2.1

In order to prove that $\hat{\mathbf{M}}_e$ converges to $\mathbf{M}_e$ at rate of $\sqrt{n}$, it is equivalent to prove that $\text{vec}(\hat{\mathbf{M}}_e)$ converges to $\text{vec}(\mathbf{M}_e)$ at rate of $\sqrt{n}$, where vec is an operation to stack all the columns of a matrix to a single vector.

By the form of $\mathbf{M}_e$, it can be written:

$$\begin{aligned}\text{vec}(\mathbf{M}_e) &= \text{vec}(\sum_{k=1}^{m}(c_k^2\omega_k\omega_k^T + d_k^2\omega_k\omega_k^T)) \\ &= \sum_{k=1}^{m}((c_k^2\text{vec}(\omega_k\omega_k^T) + d_k^2\text{vec}(\omega_k\omega_k^T)) = \mathbf{AE},\end{aligned}$$

where $\mathbf{A} = (\text{vec}(\omega_1\omega_1^T), \text{vec}(\omega_1\omega_1^T), \cdots, \text{vec}(\omega_m\omega_m^T), \text{vec}(\omega_m\omega_m^T))$ is a $p^2 \times 2m$ matrix, $c_k = E(Y\sin(\omega_k^T\mathbf{X}))$, $d_k = E(Y\cos(\omega_k^T\mathbf{X}))$, and $\mathbf{E} = (c_1^2, d_1^2, \cdots, c_m^2, d_m^2)^T$ is a $2m \times 1$ vector. In the same manner, $\hat{\mathbf{M}} = \mathbf{A}\hat{\mathbf{E}}$ with $\hat{\mathbf{E}} = (\hat{c}_1^2, \hat{d}_1^2, \cdots, \hat{c}_m^2, \hat{d}_m^2)^T$.

Let $\mathbf{E}_1 = (c_1, d_1, \cdots, c_m, d_m)^T$ and $\hat{\mathbf{E}}_1 = (\hat{c}_1, \hat{d}_1, \cdots, \hat{c}_m, \hat{d}_m)^T$. As $n \to \infty$,

$$\begin{aligned}\sqrt{n}(\hat{\mathbf{E}}_1 - \mathbf{E}_1) &= \sqrt{n}(\begin{bmatrix} \frac{1}{n}\sum_{j=1}^{n} Y_j\sin(\omega_1^T\mathbf{X}_j) \\ \frac{1}{n}\sum_{j=1}^{n} Y_j\cos(\omega_1^T\mathbf{X}_j) \\ \vdots \\ \frac{1}{n}\sum_{j=1}^{n} Y_j\sin(\omega_m^T\mathbf{X}_j) \\ \frac{1}{n}\sum_{j=1}^{n} Y_j\cos(\omega_m^T\mathbf{X}_j) \end{bmatrix} - \begin{bmatrix} E(Y\sin(\omega_1^T\mathbf{X}) \\ E(Y\cos(\omega_1^T\mathbf{X}) \\ \vdots \\ E(Y\sin(\omega_m^T\mathbf{X}) \\ E(Y\cos(\omega_m^T\mathbf{X}) \end{bmatrix}) \\ &= \sqrt{n}(\frac{1}{n}\sum_{j=1}^{n}\begin{bmatrix} Y_j\sin(\omega_1^T\mathbf{X}_j) \\ Y_j\cos(\omega_1^T\mathbf{X}_j) \\ \vdots \\ Y_j\sin(\omega_m^T\mathbf{X}_j) \\ Y_j\cos(\omega_m^T\mathbf{X}_j) \end{bmatrix} - \begin{bmatrix} E(Y\sin(\omega_1^T\mathbf{X}) \\ E(Y\cos(\omega_1^T\mathbf{X}) \\ \vdots \\ E(Y\sin(\omega_m^T\mathbf{X}) \\ E(Y\cos(\omega_m^T\mathbf{X}) \end{bmatrix}) \xrightarrow{d} N_{2m}(\mathbf{0}, \Sigma_E),\end{aligned}$$

where

$$\mathbf{\Sigma}_E = \text{cov} \begin{bmatrix} Y \sin(\omega_1^T \mathbf{X}) \\ Y \cos(\omega_1^T \mathbf{X}) \\ \vdots \\ Y \sin(\omega_m^T \mathbf{X}) \\ Y \cos(\omega_m^T \mathbf{X}) \end{bmatrix}.$$

Then, by delta method for a random vector, it can be shown that

$$\sqrt{n}(\hat{\mathbf{E}} - \mathbf{E}) \xrightarrow{d} N_{2m}(\mathbf{0}, \mathbf{D}\mathbf{\Sigma}_E\mathbf{D}^T),$$

where

$$\mathbf{D} = \begin{bmatrix} 2c_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 2d_1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 2c_2 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ 0 & 0 & 0 & \cdots & 2c_m & 0 \\ 0 & 0 & 0 & \cdots & 0 & 2d_m \end{bmatrix}.$$

Thus, we have $\sqrt{n}(\mathbf{A}\hat{\mathbf{E}} - \mathbf{A}\mathbf{E}) \xrightarrow{d} N_{p^2}(\mathbf{0}, \mathbf{A}\mathbf{D}\mathbf{\Sigma}_E\mathbf{D}^T\mathbf{A}^T)$.

That is, $\sqrt{n}(\text{vec}(\hat{\mathbf{M}}_1) - \text{vec}(\mathbf{M}_1)) \xrightarrow{d} N_{p^2}(\mathbf{0}, \mathbf{\Sigma}_1)$, where $\mathbf{\Sigma}_1 = \mathbf{A}\mathbf{D}\mathbf{\Sigma}_E\mathbf{D}^T\mathbf{A}^T$. □

Proof of theorem 2.2.2

We have $\mathbf{M} = \mathbf{C}\mathbf{C}^T$, where the $p \times 2m$ matrix $\mathbf{C} = (\mathbf{a}_1, \mathbf{b}_1, \mathbf{a}_2, \mathbf{b}_2, \cdots, \mathbf{a}_m, \mathbf{b}_m) = (c\omega_1, d\omega_1, c\omega_2, d\omega_2, \cdots, c\omega_m, d\omega_m)$. The correspondent sample version is $\hat{\mathbf{M}} = \hat{\mathbf{C}}\hat{\mathbf{C}}^T$ with values c and d in $\mathbf{C}$ replaced by $\hat{c}$ and $\hat{d}$. Under this form, we will investigate the asymptotic distribution of the smallest $\min(p-d,\ 2m-d)$ singular values of $\hat{\mathbf{C}}$ with an approach proposed by Eaton and Tyler [21]. After that, the asymptotic distribution of the smallest $\min(p-d,\ 2m-d)$ singular values of $\hat{\mathbf{C}}\hat{\mathbf{C}}^T$ can be found easily. By singular value decomposition, $\mathbf{C}$ can be expressed as

$$\mathbf{C} = \mathbf{U} \begin{bmatrix} \mathbf{\Delta} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}^T,$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices with dimension $p \times p$ and $2m \times 2m$ accordingly and $\mathbf{\Delta}$ is a $d \times d$ diagonal matrix with non-zero singular values of $\mathbf{C}$ as its elements. We partition $\mathbf{U}$ into $\mathbf{U}_1$ and $\mathbf{U}_2$ with dimensions $p \times d$ and $p \times (p-d)$, and then partition $\mathbf{V}$ into $\mathbf{V}_1$ and $\mathbf{V}_2$ with dimensions $2m \times d$ and $2m \times (2m-d)$.

Let $\tilde{\mathbf{C}} = \sqrt{n}(\hat{\mathbf{C}} - \mathbf{C})$ and $\tilde{\mathbf{W}} = \sqrt{n}\mathbf{U}_2^T(\hat{\mathbf{C}} - \mathbf{C})\mathbf{V}_2$. Because $\mathbf{U}_2^T\mathbf{C}\mathbf{V}_2 = \mathbf{0}$, $\tilde{\mathbf{W}} = \sqrt{n}\mathbf{U}_2^T\hat{\mathbf{C}}\mathbf{V}_2$, which is a $(p-d) \times (2m-d)$ matrix. Eaton and Tyler[21] showed that the asymptotic distribution of the smallest $\min(p-d,\ 2m-d)$ singular values of $\tilde{\mathbf{C}}$ is identical to the asymptotic distribution of the singular values of $\tilde{\mathbf{W}}$. By multivariate version of central limit theorem, $\text{vec}(\tilde{\mathbf{C}})$ converges to a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{\Sigma}_c = \mathbf{D}_\omega \mathbf{\Sigma}_E \mathbf{D}_\omega^T$, where $\mathbf{D}_\omega = \begin{bmatrix} \omega_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \omega_1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \omega_2 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ 0 & 0 & 0 & \cdots & \omega_m & 0 \\ 0 & 0 & 0 & \cdots & 0 & \omega_m \end{bmatrix}$ is a $2mp \times 2m$ matrix.

Now, the asymptotic distribution of $\hat{\Lambda}_d$ is identical to the asymptotic distribution of the sum of the squared singular values of $\sqrt{n} \cdot \text{vec}(\mathbf{U}_2^T\hat{\mathbf{C}}\mathbf{V}_2)$. That is, the asymptotic distribution of $\hat{\Lambda}_d$ is the same as the asymptotic distribution of $trace\tilde{\mathbf{W}}\tilde{\mathbf{W}}^T = n \cdot trace(\mathbf{U}_2^T\hat{\mathbf{C}}\mathbf{V}_2)(\mathbf{U}_2^T\hat{\mathbf{C}}\mathbf{V}_2)^T = n \cdot \text{vec}(\mathbf{U}_2^T\hat{\mathbf{C}}\mathbf{V}_2)^T\text{vec}(\mathbf{U}_2^T\hat{\mathbf{C}}\mathbf{V}_2)$. Because of the asymptotic normality of $\text{vec}(\tilde{\mathbf{C}})$, as $n \to \infty$ we have

$$\sqrt{n}\text{vec}(\mathbf{U}_2^T\hat{\mathbf{C}}\mathbf{V}_2) \to N_{(2m-d)(p-d)}(\mathbf{0}, (\mathbf{V}_2^T \otimes \mathbf{U}_2^T)\Sigma_c(\mathbf{U}_2 \otimes \mathbf{V}_2)).$$

Based on the formulation above, we have $\text{vec}(\tilde{\mathbf{W}}) = \sqrt{n}\text{vec}(\mathbf{U}_2^T\hat{\mathbf{C}}\mathbf{V}_2)$ converges to multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance $\mathbf{\Omega}_T = (\mathbf{V}_2^T \otimes \mathbf{U}_2^T)\Sigma_c(\mathbf{V}_2 \otimes \mathbf{U}_2)$. That is, as $n \to \infty$,

$$\text{vec}(\tilde{\mathbf{W}}) \to \mathbf{\Omega}_T^{1/2}\mathbf{Z},$$

where $\mathbf{Z}$ denotes the multivariate standard normal distribution.

Then, $\hat{\Lambda}_d = \text{vec}(\tilde{\mathbf{W}})^T\text{vec}(\tilde{\mathbf{W}}) \to (\mathbf{\Omega}_T^{1/2}\mathbf{Z})^T(\mathbf{\Omega}_T^{1/2}\mathbf{Z}) = \mathbf{Z}^T\mathbf{\Omega}_T\mathbf{Z}$. By eigenvalue decomposition, the nonnegative definite and symmetric covariance matrix can be ex-

pressed as $\mathbf{\Omega}_T = \mathbf{P}\Delta\mathbf{P}^T$, where $\mathbf{P}$ is orthonormal matrix and Δ is a $(2m-d)(p-d) \times (2m-d)(p-d)$ matrix with δ_js, eigenvalues of $\mathbf{\Omega}_T$, along its diagonal and 0 along its off-diagonal. Thus,

$$\begin{aligned}\mathbf{Z}^T\mathbf{\Omega}_T\mathbf{Z} &= \sum_{j=1}^{(2m-d)(p-d)} \delta_j(\mathbf{p}_j^T\mathbf{Z})^T(\mathbf{p}_j^T\mathbf{Z}) = \sum_{j=1}^{(2m-d)(p-d)} \delta_j\Big(\sum_{i=1}^{(2m-d)(p-d)} (p_{ij}\mathbf{z})^2\Big) \\ &= \sum_{j=1}^{(2m-d)(p-d)} \delta_j\mathbf{z}^2 = \sum_{j=1}^{(2m-d)(p-d)} \delta_j\chi_1^2.\end{aligned}$$

□

A2. Additional Simulations

A2.1 SDR Simulations

M7 Non-linear model (Model 12, Zhu et al., 2010 [80]): $Y = 2\cos(\beta_1^T\mathbf{X}) + \cos(\beta_2^T\mathbf{X}) + 0.25\epsilon$.

Let $p = 10$, $n = 400$, $\mathbf{X} = (X_1, \cdots, X_{10})^T$, $\boldsymbol{\beta}_1 = (1, 0, \cdots, 0)^T$, $\boldsymbol{\beta}_2 = (0, 1, \cdots, 0)^T$ and $\epsilon, X_1, \cdots, X_{10}$ be iid $N(0, 1)$. Table A1 reports the results.

Table A1: Model M7 SDR results

method	r	Δ_m	Δ_f	m_1	m_2
SIR	0.419 (0.147)	0.975 (0.041)	1.787 (0.144)	0.882 (0.120)	0.898 (0.093)
SAVE	0.973 (0.012)	0.286 (0.073)	0.456 (0.094)	0.217 (0.059)	0.231 (0.066)
DR	0.973 (0.012)	0.284 (0.073)	0.450 (0.096)	0.214 (0.073)	0.223 (0.072)
PHD	0.976 (0.011)	0.274 (0.072)	0.429 (0.098)	0.205 (0.062)	0.217 (0.061)
Euler (Ours)	**0.983** (0.007)	**0.223** (0.058)	**0.354** (0.077)	**0.172** (0.064)	**0.171** (0.053)
Kernel (Ours)	0.982 (0.010)	0.237 (0.067)	0.370 (0.090)	0.180 (0.073)	0.177 (0.060)

M8 Non-linear model (Model 14, Zhu et al., 2010[80]): $Y = 2(\beta_1^T\mathbf{X})^2 + (\beta_2^T\mathbf{X})^2 + 0.25\epsilon$.

Let $p = 10$, $n = 400$, $\mathbf{X} = (X_1, \cdots, X_{10})^T$, the real direction $\beta_1 = (1, 0, \cdots, 0)^T$, $\beta_2 = (0, 1, 0, \cdots, 0)^T$ and $\epsilon, X_1, \cdots, X_{10}$ be iid $N(0, 1)$. Table A2 listed the results.

M9 Non-normal model (Model A, Sheng and Yin, 2016 [54]): $Y = (\beta_1^T\mathbf{X})^2 + (\beta_2^T\mathbf{X}) + 0.1\epsilon$.

Let $n = 500$, $p = 20$, $\beta_1 = (1, 0, \cdots, 0)^T$, $\beta_2 = (0, 1, 0, \cdots, 0)^T$, $\epsilon \sim N(0, 1)$ and $\frac{X_i + 2}{5} \sim Beta(0.75, 1)$. The results are in Table A3.

Table A2: M8 SDR results

method	r	Δ_m	Δ_f	m_1	m_2
SIR	0.435 (0.124)	0.983 (0.026)	1.780 (0.120)	0.888 (0.110)	0.883 (0.109)
SAVE	0.977 (0.009)	0.256 (0.055)	0.416 (0.079)	0.195 (0.052)	0.214 (0.057)
DR	0.978 (0.009)	0.253 (0.055)	0.411 (0.080)	0.203 (0.061)	0.199 (0.055)
PHD	0.972 (0.011)	0.287 (0.064)	0.461 (0.085)	0.221 (0.059)	0.231 (0.066)
Euler (Ours)	**0.982** (0.007)	**0.226** (0.056)	**0.366** (0.076)	**0.178** (0.059)	0.180 (0.049)
Kernel (Ours)	0.981 (0.009)	0.240 (0.060)	0.377 (0.081)	0.192 (0.067)	**0.173** (0.057)

Table A3: Mean and variance of Criteria for model M9

method	r	Δ_m	Δ_f	m_1	m_2
SIR	0.931 (0.023)	0.467 (0.085)	0.718 (0.117)	0.350 (0.066)	0.365 (0.067)
SAVE	0.700 (0.069)	0.927 (0.121)	1.413 (0.163)	0.707 (0.139)	0.685 (0.153)
DR	0.961 (0.011)	0.311 (0.043)	0.549 (0.073)	0.276 (0.046)	0.269 (0.055)
PHD	0.709 (0.020)	0.976 (0.030)	1.410 (0.042)	0.692 (0.045)	0.716 (0.045)
Euler (Ours)	0.956 (0.017)	0.362 (0.082)	0.577 (0.106)	0.278 (0.083)	0.282 (0.093)
Kernel (Ours)	**0.983** (0.005)	**0.207** (0.030)	**0.364** (0.049)	**0.181** (0.034)	**0.180** (0.033)

Results in Tables A1, A2 and A3 are from these additional SDR simulations, with models including non-linear and non-normal, support the same conclusion that the estimation from our approaches are consistent and have top performance.

A2.2 Additional Dimension Test Results

We present the dimension determination results by permutation method for models M7 and M8. In our simulation, we choose $\alpha = 0.05$ and $L = 100$. The proportion of correctly determined dimension d for each model is reported in Table A4. The results show that permutation test works quite well with large samples.

Table A4: Dimension estimation by permutation test

model	method	n=400	n=800	n=1200
M7	Kernel	0.96	0.96	0.98
	Euler	0.93	0.91	0.93
M8	Kernel	0.96	0.98	0.95
	Euler	0.98	0.93	0.93

A2.3 Comparisons to Other Methods

Comparison to MAVE

We present the simulation results of MAVE in Table A5. Compare this table to our results in section 5.1 of the paper, our methods outperform MAVE for all the models. The sparse MAVE results are reported in Table A6, where we use the method developed by Wang and Yin [61] and their **Matlab** code. For both Models M1 and M2, our two approaches have similar TPR but smaller FPR.

Table A5: MAVE SDR results

Models	r	Δ_m	Δ_f	m_1	m_2
M1	0.982 (0.007)	0.184 (0.039)	0.260 (0.055)	0.184 (0.039)	–
M2	0.860 (0.078)	0.659 (0.203)	0.972 (0.276)	0.484 (0.219)	0.431 (0.207)
M4	0.840 (0.032)	0.538 (0.055)	0.761 (0.077)	0.538 (0.055)	–
M6a	0.876 (0.141)	0.427 (0.176)	0.604 (0.247)	0.427 (0.176)	–
M7	0.877 (0.093)	0.601 (0.255)	0.877 (0.349)	0.372 (0.250)	0.416 (0.267)
M8	0.916 (0.070)	0.494 (0.212)	0.736 (0.283)	0.296 (0.190)	0.374 (0.219)
M9	0.970 (0.008)	0.282 (0.043)	0.481 (0.063)	0.234 (0.050)	0.241 (0.050)

Table A6: SparseMAVE Results

Model	d	Criteria	SparseMAVE
M1	1	TPR	1.000
		FPR	0.293
M2	2	TPR	1.000
		FPR	0.058

A2.4 Additional simulations on Euler Approach

SDR results with various m's and τ's

We compare the estimation accuracy for Euler Approach in different models with either fixing τ and varying m or fixing m and varying τ, where m is the total number of ω generated and τ is the percentage of ω selected. The estimation accuracy for models M2, M6a, M7 and M8 are reported in Table A7.

From Table A7, we can see that with a fixed percentage of ω, the estimation gets better as the number of ω gets larger. For a simple model ($d = 1$; M6a), the estimation is very stable after $m = 5000$; and for relatively complex models ($d = 2$; M2, M7 and M8), the results start to be stable after $m = 10000$. It is within our expectation that more ωs are needed to recover the true directions for a complex model. When m is fixed in a relative small number, the results are better with a

Table A7: SDR Estimation from Euler Approach in Various Settings

Model	m	τ	r	Δ_m	Δ_f	m_1	m_2
M6a	1000	0.20	0.962 (0.018)	0.264 (0.063)	0.373 (0.089)	0.264 (0.063)	
		0.10	0.964 (0.017)	0.256 (0.063)	0.362 (0.090)	0.256 (0.064)	
		0.05	0.965 (0.017)	0.253 (0.064)	0.357 (0.091)	0.253 (0.064)	
	5000	0.20	0.969 (0.014)	0.239 (0.054)	0.339 (0.076)	0.239 (0.054)	
		0.10	0.972 (0.012)	0.228 (0.051)	0.323 (0.072)	0.228 (0.051)	
		0.05	0.975 (0.011)	0.217 (0.048)	0.307 (0.068)	0.217 (0.048)	
	10000	0.20	0.970 (0.013)	0.239 (0.052)	0.337 (0.073)	0.239 (0.052)	
		0.10	0.973 (0.012)	0.227 (0.049)	0.320 (0.069)	0.227 (0.049)	
		0.05	0.976 (0.010)	0.211 (0.045)	0.298 (0.064)	0.211 (0.045)	
	20000	0.20	0.970 (0.013)	0.236 (0.050)	0.334 (0.070)	0.236 (0.050)	
		0.10	0.974 (0.011)	0.223 (0.047)	0.316 (0.066)	0.223 (0.047)	
		0.05	0.978 (0.010)	0.206 (0.044)	0.291 (0.062)	0.206 (0.044)	
M2	1000	0.20	0.951 (0.026)	0.361 (0.098)	0.598 (0.141)	0.286 (0.085)	0.300 (0.101)
		0.10	0.947 (0.049)	0.364 (0.135)	0.604 (0.189)	0.289 (0.123)	0.300 (0.109)
		0.05	0.950 (0.032)	0.367 (0.105)	0.602 (0.151)	0.275 (0.082)	0.312 (0.116)
	5000	0.20	0.966 (0.028)	0.297 (0.096)	0.495 (0.139)	0.236 (0.074)	0.249 (0.093)
		0.10	0.970 (0.027)	0.280 (0.088)	0.468 (0.128)	0.232 (0.085)	0.228 (0.070)
		0.05	0.970 (0.025)	0.281 (0.088)	0.467 (0.126)	0.228 (0.087)	0.230 (0.068)
	10000	0.20	0.969 (0.028)	0.283 (0.094)	0.472 (0.135)	0.220 (0.063)	0.243 (0.095)
		0.10	0.972 (0.027)	0.267 (0.090)	0.446 (0.129)	0.227 (0.094)	0.211 (0.057)
		0.05	0.974 (0.021)	0.258 (0.081)	0.433 (0.117)	0.214 (0.082)	0.212 (0.057)
	20000	0.20	0.964 (0.036)	0.309 (0.122)	0.503 (0.169)	0.251 (0.118)	0.237 (0.085)
		0.10	0.975 (0.011)	0.261 (0.058)	0.432 (0.086)	0.208 (0.058)	0.216 (0.061)
		0.05	0.977 (0.009)	0.255 (0.053)	0.423 (0.076)	0.210 (0.058)	0.204 (0.058)
M7	1000	0.20	0.967 (0.016)	0.307 (0.080)	0.493 (0.116)	0.220 (0.089)	0.254 (0.086)
		0.10	0.959 (0.020)	0.349 (0.094)	0.548 (0.131)	0.248 (0.107)	0.272 (0.110)
		0.05	0.929 (0.053)	0.456 (0.165)	0.697 (0.223)	0.265 (0.149)	0.372 (0.191)
	5000	0.20	0.981 (0.008)	0.238 (0.057)	0.384 (0.078)	0.178 (0.060)	0.194 (0.064)
		0.10	0.982 (0.008)	0.233 (0.055)	0.373 (0.076)	0.172 (0.064)	0.189 (0.058)
		0.05	0.978 (0.013)	0.259 (0.076)	0.408 (0.103)	0.180 (0.073)	0.206 (0.091)
	10000	0.20	0.982 (0.007)	0.226 (0.052)	0.367 (0.072)	0.167 (0.054)	0.187 (0.066)
		0.10	0.984 (0.007)	0.218 (0.053)	0.353 (0.072)	0.157 (0.057)	0.182 (0.061)
		0.05	0.982 (0.008)	0.232 (0.059)	0.367 (0.079)	0.163 (0.065)	0.185 (0.074)
	20000	0.20	0.983 (0.007)	0.221 (0.052)	0.360 (0.075)	0.160 (0.057)	0.188 (0.059)
		0.10	0.985 (0.006)	0.205 (0.049)	0.336 (0.072)	0.145 (0.047)	0.181 (0.057)
		0.05	0.986 (0.006)	0.200 (0.052)	0.325 (0.075)	0.152 (0.053)	0.162 (0.059)
M8	1000	0.20	0.966 (0.015)	0.312 (0.077)	0.502 (0.109)	0.231 (0.074)	0.253 (0.096)
		0.10	0.956 (0.025)	0.357 (0.107)	0.563 (0.148)	0.246 (0.106)	0.288 (0.121)
		0.05	0.923 (0.055)	0.478 (0.171)	0.729 (0.227)	0.290 (0.161)	0.375 (0.202)
	5000	0.20	0.980 (0.008)	0.237 (0.052)	0.386 (0.075)	0.178 (0.056)	0.198 (0.058)
		0.10	0.981 (0.008)	0.236 (0.058)	0.381 (0.081)	0.173 (0.059)	0.197 (0.063)
		0.05	0.977 (0.013)	0.260 (0.082)	0.412 (0.112)	0.177 (0.073)	0.215 (0.090)
	10000	0.20	0.982 (0.008)	0.228 (0.051)	0.373 (0.074)	0.177 (0.049)	0.187 (0.064)
		0.10	0.983 (0.008)	0.223 (0.055)	0.361 (0.078)	0.159 (0.054)	0.190 (0.061)
		0.05	0.981 (0.012)	0.236 (0.073)	0.376 (0.099)	0.157 (0.057)	0.201 (0.084)
	20000	0.20	0.983 (0.007)	0.221 (0.051)	0.362 (0.077)	0.173 (0.050)	0.179 (0.061)
		0.10	0.985 (0.007)	0.203 (0.049)	0.335 (0.073)	0.160 (0.054)	0.167 (0.049)
		0.05	0.986 (0.007)	0.197 (0.054)	0.323 (0.078)	0.152 (0.049)	0.162 (0.059)

relative large τ while the results are very stable across different τs when m is large. Thus, conservatively to have a stable estimation, we unify the $m = 20000$ and $\tau = 0.2$ for SDR study.

Permutation results with various m's

Table A8: Permutation Test with a fixed τ and different m's for Euler Approach

Model	m	n=400	n=800	n=1200
M6a	500	0.96	0.91	0.89
	1000	0.95	0.93	0.93
	2000	0.98	0.95	0.92

We fixed the percentage $\tau = 0.2$ as in SDR and work on a relative small number of ω, m. We compare the results from $m = 500, 1000$ and 2000. The results for model M6a are reported in Table A8. As we can see in the table, from $m = 500$ to $m = 1000$, there is a bit increment in the percentage of correct dimension determination while there is no significant difference between $m = 1000$ and $m = 2000$. Therefore, we use $m = 1000$ and $\tau = 0.2$ in permutation test.

SVS results with various m's

For sufficient variable selection, the primary goal is to identify the informative variables and a relative good direction estimation will be enough. Thus we use a relative small number of ωs to save the computation expense with $\tau = 0.2$. As in the permutation test, we compare the results from $m = 500$, 1000 and 2000. The results for model M6a are reported in Table A9, which clearly indicates that from $m = 500$ to $m = 1000$, the variable selection accuracy is a bit improved; From $m = 1000$ to $m = 2000$, the improvement is ignorable. Thus, we use $m = 1000$ in reported result.

Table A9: SVS with a fixed p and different m's for Euler Approach

Model	Method	m	TPR	FPR
M6a	Alasso	500	0.940	0.065
		1000	0.995	0.063
		2000	1.000	0.091
	CISE	500	1.000	0.075
		1000	1.000	0.024
		2000	1.000	0.020

SSVS Euler results with various m's

For the SSVS Euler approach, to reduce the computational cost, we also use a relative small m, the number of ω generated. The results with Model M5 for $m = 500$, $m = 1000$ and $m = 2000$ are reported in Table A10. From $m = 500$ to $m = 1000$, there is an increment while from $m = 1000$ to $m = 2000$, there is no significant difference.

Table A10: Sequential SVS for Euler Approach with Model M5 and different ms

Method	m	Δ_f	$\lvert r_1 \rvert$	TPR	FPR
SSVSEuler	500	0.940 (0.164)	0.861 (0.042)	0.991	0.128
	1000	0.713 (0.168)	0.917 (0.034)	0.994	0.095
	2000	0.708 (0.159)	0.913 (0.033)	0.996	0.149

B Supplementary Materials for Chapter 3

B1. Proofs

B1.1 Proof of theorem 3.2.1

Here we will prove that the sample $\hat{\boldsymbol{\xi}}$ converges to $\boldsymbol{\xi}$ at the rate of $\sqrt{n}$.

$$\sqrt{n}(vec(\hat{\boldsymbol{\xi}}) - vec(\boldsymbol{\xi})) = \sqrt{n}\Bigg(\begin{bmatrix} \frac{1}{n}\sum_{j=1}^{n} \omega_1 \tilde{y}_j \cos(\omega_1^T \tilde{\mathbf{x}}_j) \\ \frac{1}{n}\sum_{j=1}^{n} \omega_1 \tilde{y}_j \sin(\omega_1^T \tilde{\mathbf{x}}_j) \\ \vdots \\ \frac{1}{n}\sum_{j=1}^{n} \omega_m \tilde{y}_j \cos(\omega_1^T \tilde{\mathbf{x}}_j) \\ \frac{1}{n}\sum_{j=1}^{n} \omega_m \tilde{y}_j \sin(\omega_1^T \tilde{\mathbf{x}}_j) \end{bmatrix} - \begin{bmatrix} E(\omega_1 y \cos(\omega_1^T \mathbf{x}) \\ E(\omega_1 y \sin(\omega_1^T \mathbf{x}) \\ \vdots \\ E(\omega_m y \cos(\omega_m^T \mathbf{x}) \\ E(\omega_m y \sin(\omega_m^T \mathbf{x}) \end{bmatrix}\Bigg)$$

$$= \sqrt{n}\Big(\frac{1}{n}\sum_{j=1}^{n} \begin{bmatrix} \omega_1 \tilde{y}_j \cos(\omega_1^T \tilde{\mathbf{x}}_j) \\ \omega_1 \tilde{y}_j \sin(\omega_1^T \tilde{\mathbf{x}}_j) \\ \vdots \\ \omega_m \tilde{y}_j \cos(\omega_m^T \tilde{\mathbf{x}}_j) \\ \omega_m \tilde{y}_j \sin(\omega_m^T \tilde{\mathbf{x}}_j) \end{bmatrix} - \begin{bmatrix} E(\omega_1 y \cos(\omega_1^T \mathbf{x}) \\ E(\omega_1 y \sin(\omega_1^T \mathbf{x}) \\ \vdots \\ E(\omega_m y \cos(\omega_m^T \mathbf{x}) \\ E(\omega_m y \sin(\omega_m^T \mathbf{x}) \end{bmatrix}\Big) \xrightarrow{d} N(\mathbf{0}, \Gamma).$$

The conclusion is based on the multivariate version of central limit theorem and

$$\Gamma = \operatorname{cov}\begin{bmatrix} \omega_1 y \sin(\omega_1^T \mathbf{x}) \\ \omega_1 y \cos(\omega_1^T \mathbf{x}) \\ \vdots \\ \omega_m y \sin(\omega_m^T \mathbf{x}) \\ \omega_m y \cos(\omega_m^T \mathbf{x}) \end{bmatrix} = \begin{bmatrix} \omega_1 & 0 & \cdots & 0 & 0 \\ 0 & \omega_1 & \cdots & 0 & 0 \\ \vdots & & & & \\ 0 & 0 & \cdots & \omega_m & 0 \\ 0 & 0 & \cdots & 0 & \omega_m \end{bmatrix} \Sigma_{\boldsymbol{\xi}} \begin{bmatrix} \omega_1 & 0 & \cdots & 0 & 0 \\ 0 & \omega_1 & \cdots & 0 & 0 \\ \vdots & & & & \\ 0 & 0 & \cdots & \omega_m & 0 \\ 0 & 0 & \cdots & 0 & \omega_m \end{bmatrix}^T,$$

where $\Sigma_{\boldsymbol{\xi}} = \operatorname{cov}\begin{bmatrix} y \sin(\omega_1^T \mathbf{x}) \\ y \cos(\omega_1^T \mathbf{x}) \\ \vdots \\ y \sin(\omega_m^T \mathbf{x}) \\ y \cos(\omega_m^T \mathbf{x}) \end{bmatrix}$ is a $2m$ by $2m$ matrix and Γ is a $2mp$ by $2mp$ matrix.

□

B1.2 Proof of theorem 3.2.2 and 3.2.3

The proofs of theorem 3.2.2 and 3.2.3 are based on Shapiro's results[52] and Cook and Ni's lemmas $A.3$ and $A.4$[16]. They are labeled as Propositions B1.1, B1.2 and B1.3 here and they are directly adopted from Cook and Ni[16]. The whole proof follows the proof in the Appendix of Cook and Ni[16].

Proposition B1.1(Shapiro, 1986[52]) Assume $\boldsymbol{\eta}$ is a p dimensional parameter vector in a parameter space $\Theta \subseteq \mathbb{R}^p$ and $\boldsymbol{\eta}_0$ is the true value of $\boldsymbol{\eta}$. Let $\mathbf{f}(\boldsymbol{\eta}) = (f_1(\boldsymbol{\eta}), \cdots, f_m(\boldsymbol{\eta}))^T : \Theta \to \mathbb{R}^m$ be a vector of twice continuously differentiable functions on Θ and Δ be the corresponding Jacobian matrix. Δ can be rank deficiency and f will be overparameterized. In addition, suppose the following conditions hold.

1. The sample estimate $\hat{x}$ is asymptotically multivariate normal distributed. That is $\sqrt{n}(\hat{x} - \mathbf{f}(\boldsymbol{\eta}_0)) \xrightarrow{d} N(\mathbf{0}, \Gamma)$.

2. The discrepancy function $F(\hat{x}, \mathbf{f}(\boldsymbol{\eta})) = (\hat{x} - \mathbf{f}(\boldsymbol{\eta}))^T \mathbf{V} (\hat{x} - \mathbf{f}(\boldsymbol{\eta}))$ with a given $\mathbf{V}$ has the following conditions:

   - $F(x, y) \geq 0$ for any x, y;
   - $F(x, y) = 0$ if and only if $x = y$;

- F is twice continuously differentiable in x and y;
- There exist constants $\epsilon_1 > 0$ and $\epsilon_2 > 0$ such that $F(x, y) \geq \epsilon_2$ whenever $||x - y|| \geq \epsilon_1$.

3. $\boldsymbol{\eta}_0$ is regular;

4. $\mathbf{V}$ is positive definite on the linear space spanned by the columns of Δ. That is $\text{rank}(\Delta^T \mathbf{V} \Delta) = \text{rank}(\Delta)$ given $\Delta^T \mathbf{V} \Delta$ is nonnegative definite.

   With the above conditions hold, we have the following results:

   - Assume $\hat{F} = F(\hat{x}, \mathbf{f}(\hat{\boldsymbol{\eta}}))$ is the minimized value of $F(\hat{x}, \mathbf{f}(\boldsymbol{\eta}))$ over Θ. Then $n\hat{F}$ has the same distribution of $W^T U W$ with $W \sim N(0, \Gamma)$ and $U = \mathbf{V} - \mathbf{V}\Delta(\Delta^T \mathbf{V} \Delta)^{-} \Delta^T \mathbf{V} = \mathbf{V}^{1/2} Q_\Phi \mathbf{V}^{1/2}$, where $\Phi = \mathbf{V}^{1/2}\Delta$ and $Q_\Phi = I - \Phi(\Phi^T \Phi)^{-} \Phi^T$;
   - The estimate $\mathbf{f}(\hat{\boldsymbol{\eta}})$ which minimized $F(\hat{x}, \mathbf{f}(\boldsymbol{\eta}))$ is a consistent estimator and an asymptotically normal estimate of $\mathbf{f}(\boldsymbol{\eta}_0)$:
     $\sqrt{n}(\mathbf{f}(\hat{\boldsymbol{\eta}}) - \mathbf{f}(\boldsymbol{\eta})) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}^{-1/2} P_\Phi \mathbf{V}^{1/2} \Gamma \mathbf{V}^{1/2} P_\Phi \mathbf{V}^{-1/2})$;

If $\mathbf{V}$ is random instead of fixed, we need the following two additional propositions from Cook and Ni[16].

Proposition B1.2(Cook and Ni[16]) Assume $\{\mathbf{X}_n\} \in \mathbb{R}^s$ is a sequence of random vectors , and $\boldsymbol{\xi} \in \boldsymbol{\Xi} \subset \mathbb{R}^s$. In addition, let $\{\mathbf{V}_n > 0\}$ be a sequence of $s \times s$ matrices that converges in probability to $\mathbf{V} > 0$. If $n\hat{F}_{\mathbf{V}} = \min_{\boldsymbol{\xi} \in \boldsymbol{\Xi}} n(\mathbf{X}_n - \boldsymbol{\xi})^T \mathbf{V} (\mathbf{X}_n - \boldsymbol{\xi}) \xrightarrow{d} \Psi$, then $n\hat{F}_{\mathbf{V}_n} = \min_{\boldsymbol{\xi} \in \boldsymbol{\Xi}} n(\mathbf{X}_n - \boldsymbol{\xi})^T V_n (\mathbf{X}_n - \boldsymbol{\xi}) \xrightarrow{d} \Psi$ is also true and vice versa. Moreover, let $\hat{\boldsymbol{\xi}}_1$ and $\hat{\boldsymbol{\xi}}_2$ be the values of $\boldsymbol{\xi}$ that approach $n\hat{F}_{\mathbf{V}}$ and $n\hat{F}_{\mathbf{V}_n}$. If $\mathbf{V}^{1/2}\mathbf{X}_n \xrightarrow{p} \alpha$, then $\mathbf{V}^{1/2}\hat{\boldsymbol{\xi}}_1$ and $\mathbf{V}_n^{1/2}\hat{\boldsymbol{\xi}}_2$ converges to α in probability.

Proposition B1.3(Cook and Ni[16]) Let $\tilde{\mathbf{X}}_n$ be a simple random sample $\{\mathbf{X}_1, \cdots, \mathbf{X}_n\}$ with $\mathbf{X}_i$ being either scalar or vector. The distribution of $\mathbf{X}$ depends on parameters that include a vector $\boldsymbol{\eta} \in \boldsymbol{\Theta} \subset \mathbb{R}^k$. Assume $\boldsymbol{\eta}_0$ is the true value of $\boldsymbol{\eta}$ and the following conditions hold:

- $\boldsymbol{\Theta}$ is an open set;

- $\mathbf{g}(\boldsymbol{\eta}) : \boldsymbol{\Theta} \to \mathbb{R}^s$ is one-to-one, bi-continuous, and twice continuously differentiable. Denote $D(\boldsymbol{\eta}) = \dfrac{\partial \mathbf{g}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \in \mathbb{R}^{s\times k}$ and the true value $D_0 = D(\boldsymbol{\eta}_0)$;
- $Y_n = Y_n(\tilde{\mathbf{X}}_n) \in \mathbb{R}^s$ is a consistent estimate of $\mathbf{g}(\boldsymbol{\eta}_0)$ with $\sqrt{n}(Y_n - \mathbf{g}(\boldsymbol{\eta}_0)) \xrightarrow{d} N(\mathbf{0}, \Gamma)$;
- $\mathbf{V}_n = \mathbf{V}_n(\tilde{X}_n)$ is a positive-definite matrix that converges to a constant matrix $\mathbf{V}$ in probability.

Assume a discrepancy function is given by $H(Y_n, \mathbf{g}(\boldsymbol{\eta})) = (Y_n - \mathbf{g}(\boldsymbol{\eta}))^T \mathbf{V}_n (Y_n - \mathbf{g}(\boldsymbol{\eta}))$ and $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}(\tilde{\mathbf{X}}_n)$ is the value of $\boldsymbol{\eta}$ that minimizes H. Then

$$\sqrt{n}(\hat{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}_0) \xrightarrow{d} N(\mathbf{0}, (D_0^T \mathbf{V} D_0)^{-1} D_0^T \mathbf{V} \Gamma \mathbf{V} D_0 (D_0^T \mathbf{V} D_0)^{-1})$$

and

$$\sqrt{n}(\mathbf{g}(\hat{\boldsymbol{\eta}}) - \mathbf{g}(\hat{\boldsymbol{\eta}}_0)) \xrightarrow{d} N(\mathbf{0}, D_0 (D_0^T \mathbf{V} D_0)^{-1} D_0^T \mathbf{V} \Gamma \mathbf{V} D_0 (D_0^T \mathbf{V} D_0)^{-1} D_0).$$

Proof of Theorem 3.2.2

Given $\mathbf{V} = \mathbf{I}_{2m} \otimes \boldsymbol{\Sigma}$, the estimated version is $\mathbf{V}_n = \mathbf{I}_{2m} \otimes \hat{\boldsymbol{\Sigma}}$. Then by equation 3.2, we have the discrepancy function $F_d^{ccf}(\mathbf{B}, \mathbf{C}) = (\text{vec}(\hat{\boldsymbol{\xi}}) - \text{vec}(\mathbf{BC}))^T (\mathbf{I}_{2m} \otimes \hat{\boldsymbol{\Sigma}})(\text{vec}(\hat{\boldsymbol{\xi}}) - \text{vec}(\mathbf{BC}))$. Since $\mathbf{V}_n$ converges to $\mathbf{V}$ in probability, by proposition B1.2, we know the asymptotic distributions of $n\hat{F}_d^{ccf}$ and $n\hat{F}_{1d}$ are the same, where $n\hat{F}_{1d}$ is the minimized value of discrepancy function $F_{1d}(\mathbf{B}, \mathbf{C}) = (\text{vec}(\hat{\boldsymbol{\xi}}) - \text{vec}(\mathbf{BC}))^T \mathbf{V} (\text{vec}(\hat{\boldsymbol{\xi}}) - \text{vec}(\mathbf{BC}))$. In addition, we will use reparameterization and proposition B1.3 to show that the asymptotic distribution of $\text{vec}(\hat{\boldsymbol{\beta}}\hat{\nu})$ from $F_d^{ccf}(\mathbf{B}, \mathbf{C})$ and $F_{1d}(\mathbf{B}, \mathbf{C})$ are the same. First, $\boldsymbol{\beta}\nu = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T \nu = (I_d, (\boldsymbol{\beta}_2 \boldsymbol{\beta}_1^{-1})^T)^T \boldsymbol{\beta}_1 \nu$ with $\boldsymbol{\beta}_1$ is a $d \times d$ invertible matrix. Let $\boldsymbol{\beta}_1$ be identity, then we have new parameters $\boldsymbol{\beta}_2 \in \mathbb{R}^{(p-d)\times d}$ and $\nu \in \mathbb{R}^{d \times 2m}$. Now $\boldsymbol{\beta}_2$ together with ν is the $\boldsymbol{\eta}$ in proposition B1.3 which has a full rank Jacobian and an open parameter space in $\mathbb{R}^{d(2m+p-d)}$. In addition, the reparameterization will not influence the minimization and asymptotic properties. Thus, we can investigate F_{1d} directly. In the setting, $\boldsymbol{\eta} = (\text{vec}(\mathbf{B})^T, \text{vec}(\mathbf{C})^T)^T \in \mathbb{R}^{d(p+2m)}$, $\mathbf{f}(\boldsymbol{\eta}) = \text{vec}(\mathbf{BC}) \in \mathbb{R}^{2mp}$, $\hat{x} = \text{vec}(\hat{\boldsymbol{\xi}})$, and $\mathbf{f}(\boldsymbol{\eta}_0) = \boldsymbol{\beta}\nu$. Here, $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ is the population basis for $\mathcal{S}_{\boldsymbol{\xi}}$ and $\nu \in \mathbb{R}^{d\times 2m}$ is the fitting matrix. The Jacobian matrix is $\Delta = (\nu^T \otimes \mathbf{I}_p, \mathbf{I}_{2m} \otimes \boldsymbol{\beta})$ and

$\text{rank}(\Delta) = \text{rank}(\nu^T \otimes Q_{\boldsymbol{\beta}}, \mathbf{I}_{2m} \otimes \boldsymbol{\beta}) = d(p-d)+2md = d(2m+p-d)$. By theorem 3.2.1, proposition B1.1 condition 1 is satisfied; Given $\mathbf{V} = (\mathbf{I}_{2m} \otimes \boldsymbol{\Sigma})$, condition 2 is met; $\mathbf{f}(\boldsymbol{\eta})$ is analytic, condition 3 holds; $\mathbf{V}$ is positive definite, then condition 4 is true. Thus, $\text{vec}(\hat{\boldsymbol{\beta}}\hat{\nu})$ is a consistent estimator of $\text{vec}(\boldsymbol{\beta}\nu)$ and $\sqrt{n}(\text{vec}(\hat{\boldsymbol{\beta}}\hat{\nu}) - \text{vec}(\boldsymbol{\beta}\nu)) \xrightarrow{d} N(\mathbf{0}, \Gamma_c)$, where the covariance matrix $\Gamma_c = \Delta(\Delta^T\mathbf{V}\Delta)^-\Delta^T\mathbf{V}\Gamma\mathbf{V}\Delta(\Delta^T\mathbf{V}\Delta)^-\Delta^T$. Thus the first conclusion in theorem 3.2.2 is proved. What's more, $n\hat{F}_d^{ccf}$ has the same asymptotic distribution of $W^T\mathbf{V}^{1/2}Q_\Phi\mathbf{V}^{1/2}W$, where W has normal distribution with mean $\mathbf{0}$ and covariance matrix Γ and $\Phi = \mathbf{V}^{1/2}\Delta$. Thus, $n\hat{F}_d^{ccf}$ has an asymptotic weighted chi-square distribution $\Sigma_{i=1}^{2mp}\lambda_i\chi^2_{(1)}$, where $\chi^2_{(1)}$s are independent chi-square distribution with degrees of freedom 1 and λ_is are the ordered eigenvalues of $Q_\Phi\mathbf{V}^{1/2}\Gamma\mathbf{V}^{1/2}Q_\Phi$. Thus the second conclusion in theorem 3.2.2 is proved. Note, $\lambda_{2m+1} = \cdots = \lambda_{2mp} = 0$ because $\text{rank}(\Gamma) = 2m$. Conclusion 3 in theorem 3.2.2 follows directly from conclusion 1. □

Proof of Theorem 3.2.3

The proof of theorem 3.2.3 is similar to theorem 3.2.2. Here, $\mathbf{V} = \mathbf{I}$ and we do not need to estimate it. Thus, proposition B1.1 is enough for this proof. By taking $\mathbf{I}$ as $\mathbf{V}$, all the conditions in propositions are satisfied. Thus, $\text{vec}(\hat{\boldsymbol{\beta}}\hat{\nu})$ is a consistent estimator of $\text{vec}(\boldsymbol{\beta}\nu)$ and $\sqrt{n}(\text{vec}(\hat{\boldsymbol{\beta}}\hat{\nu}) - \text{vec}(\boldsymbol{\beta}\nu)) \xrightarrow{d} N(\mathbf{0}, \Gamma_I)$, where the covariance matrix Γ_I is expressed as $\Delta(\Delta^T\Delta)^-\Delta^T\Gamma\Delta(\Delta^T\Delta)^-\Delta^T$. Conclusion 1 in theorem 3.2.3 is proved. In addition, $n\hat{F}_d^{icf}$ has the same asymptotic distribution of $W^TQ_\Delta W$, where W has normal distribution with mean $\mathbf{0}$ and covariance matrix Γ. That is, $n\hat{F}_d^{icf}$ has an asymptotic weighted chi-square distribution $\Sigma_{i=1}^{2mp}\lambda_i\chi^2_{(1)}$, where $\chi^2_{(1)}$s are independent chi-square distribution with degrees of freedom 1 and λ_is are the ordered eigenvalues of $Q_\Delta\Gamma Q_\Delta$. Conclusion 2 in theorem 3.2.3 is proved. Note, $\lambda_{2m+1} = \cdots = \lambda_{2mp} = 0$ because $\text{rank}(\Gamma) = 2m$. Conclusion 3 of theorem 3.2.3 follows directly from conclusion 1. □

Proof of Theorem 3.3.1

We follow Qian et al. to prove this theorem[46]. Let $(\mathbf{B}_0, \mathbf{C}_0)$ and $(\hat{\mathbf{B}}, \hat{\mathbf{C}})$ be the minimizers of objective function 3.5 and 3.6 respectively, with $\mathbf{B}_0 = (B_{01}, \cdots, B_{0p})^T$

and $\hat{\mathbf{B}} = (\hat{B}_1, \cdots, \hat{B}_p)^T$. Then, we have $L(\hat{\mathbf{B}}, \hat{\mathbf{C}}) \leq L(\mathbf{B}_0, \mathbf{C}_0)$. That is,

$$\frac{1}{2}\text{tr}[(\hat{\boldsymbol{\xi}} - \hat{\mathbf{B}}\hat{\mathbf{C}})^T\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\xi}} - \hat{\mathbf{B}}\hat{\mathbf{C}})] + \lambda\sum_{j=1}^{p}\theta_j||\hat{B}_j||_2$$
$$\leq \frac{1}{2}\text{tr}[(\hat{\boldsymbol{\xi}} - \hat{\mathbf{B}}_0\hat{\mathbf{C}}_0)^T\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\xi}} - \hat{\mathbf{B}}_0\hat{\mathbf{C}}_0)] + \lambda\sum_{j=1}^{p}\theta_j||B_{0j}||_2.$$

After simplification, we have

$$T_1 + T_2 + \frac{1}{2}T_3 + \lambda\sum_{j=1}^{p}\theta_j||\hat{B}_j||_2 \leq \lambda\sum_{j=1}^{p}\theta_j||B_{0j}||_2, \tag{B.1}$$

where $T_1 = -\text{tr}[(\hat{\boldsymbol{\xi}} - \mathbf{B}_0\mathbf{C}_0)^T(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})(\hat{\mathbf{B}}\hat{\mathbf{C}} - \mathbf{B}_0\mathbf{C}_0)]$,

$T_2 = -\text{tr}[(\hat{\boldsymbol{\xi}} - \mathbf{B}_0\mathbf{C}_0)^T\boldsymbol{\Sigma}(\hat{\mathbf{B}}\hat{\mathbf{C}} - \mathbf{B}_0\mathbf{C}_0))]$, and

$T_3 = \text{tr}[(\hat{\mathbf{B}}\hat{\mathbf{C}} - \mathbf{B}_0\mathbf{C}_0)^T\hat{\boldsymbol{\Sigma}}(\hat{\mathbf{B}}\hat{\mathbf{C}} - \mathbf{B}_0\mathbf{C}_0)]$.

Given y is distributed as a sub-Gaussian (C1) and $\cos(\cdot)$, $\sin(\cdot)$ are bounded functions, both $y\cos(\omega^T\mathbf{x})$ and $y\sin(\omega^T\mathbf{x})$ follow sub-Gaussian. That is, there exist constants ν, $c_1 > 0$ such that for every $k > 2$, we have

$$E\{|yg(\omega^T\mathbf{x}) - E(yg(\omega^T\mathbf{x}))|^k\} \leq \frac{k!\nu^2c_1^{k-2}}{2},$$

where $g(\cdot)$ can be either $\cos(\cdot)$ or $\sin(\cdot)$. Then for every ϵ and large enough n, we have the following inequality based on the Bernstein inequality[6],

$$P(\frac{\sum_{k=1}^{n}[y_kg(\omega^T\mathbf{x}_k) - E(yg(\omega^T\mathbf{x}))]}{n} > \epsilon) \leq \exp(-\frac{n\epsilon^2}{2(\nu^2 + c_1\epsilon)}).$$

Take $\epsilon = 1$, with $p_n = \max\{p, n\}$ and with a large enough n, we have

$$P(\frac{\sum_{k=1}^{n}[y_kg(\omega^T\mathbf{x}_k) - E(yg(\omega^T\mathbf{x}))]}{n} > 1) \leq \exp(-\frac{n}{2(\nu^2 + c_1)}).$$

Thus, with a large n, we have the following inequality with probability greater than $1 - \exp(-\frac{n}{2(\nu^2 + c_1)})$,

$$\left|\frac{\sum_{k=1}^{n}[y_kg(\omega^T\mathbf{x}_k) - E(yg(\omega^T\mathbf{x}))]}{n}\right| \leq 1.$$

Given condition C2, there exist constants c_3, $c_4 > 0$ for every $1 \leq i, j \leq p$,

$$p(|\hat{\sigma}_{i,j} - \sigma_{i,j}| > \epsilon) \leq c_3\exp(-\frac{8n\epsilon^2}{c_4}),$$

where $\hat{\sigma}_{i,j} = (\hat{\mathbf{\Sigma}})_{i,j}$ and $\sigma_{i,j} = (\mathbf{\Sigma})_{i,j}$. Now, take $\epsilon = c_5(\frac{\log p_n}{n})^{1/2}$ with $c_5 > \sqrt{c_4}$, then

$$p(|\hat{\sigma}_{i,j} - \sigma_{i,j}| > c_5(\frac{\log p_n}{n})^{1/2}) \leq \frac{c_3}{p_n^8}.$$

By union bound, with probability greater than $1 - \frac{c_3}{p_n^6}$,

$$\max_{1\leq i,j\leq p} |\hat{\sigma}_{i,j} - \sigma_{i,j}| < c_5(\frac{\log p_n}{n})^{1/2}.$$

Thus, with probability greater than $1 - \frac{c_3}{p_n^6} - \exp(-\frac{n}{2(\nu^2 + c_1)})$, we have

$$\max_{1\leq i,j\leq p} |\hat{\sigma}_{i,j} - \sigma_{i,j}| \left| \frac{\sum_{k=1}^n [y_k g(\omega^T \mathbf{x}_k) - E(yg(\omega^T \mathbf{x}))]}{n} \right| \leq c_5(\frac{\log p_n}{n})^{1/2}.$$

Now, we consider

$$\begin{aligned}
&\left| \mathbf{e}_i^T (\hat{\boldsymbol{\xi}} - \mathbf{B}_0\mathbf{C}_0)^T (\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}) \mathbf{e}_j \right| \\
=&\left| \left[\frac{1}{n} \sum_{k=1}^n y_k g(\omega_i^T \mathbf{x}_k) - E(yg(\omega_i^T \mathbf{x})) \right] \sum_{k\in\mathcal{A}} \omega_{ki} (\hat{\sigma}_{kj} - \sigma_{kj}) \right| \\
\leq& \max_{1\leq i,j\leq p} |\hat{\sigma}_{ij} - \sigma_{ij}| \left| \frac{1}{n} \sum_{k=1}^n y_k g(\omega_i^T \mathbf{x}_k) - E(yg(\omega_i^T \mathbf{x})) \right|.
\end{aligned}$$

With above formulations, we have

$$\begin{aligned}
|T_1| \leq& \sum_{j=1}^p \left| \mathbf{e}_j^T (\hat{\mathbf{B}}\hat{\mathbf{C}} - \mathbf{B}_0\mathbf{C}_0)(\hat{\boldsymbol{\xi}} - \mathbf{B}_0\mathbf{C}_0)^T (\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}) \mathbf{e}_j \right| \\
\leq& \sum_{j=1}^p ||\mathbf{e}_j^T (\hat{\mathbf{B}}\hat{\mathbf{C}} - \mathbf{B}_0\mathbf{C}_0)||_2 ||(\hat{\boldsymbol{\xi}} - \mathbf{B}_0\mathbf{C}_0)^T (\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}) \mathbf{e}_j||_2 \\
\leq& c_5\sqrt{2m}\sqrt{\frac{\log p_n}{n}} \sum_{j=1}^p ||\mathbf{e}_j (\hat{\mathbf{B}}\hat{\mathbf{C}} - \mathbf{B}_0\mathbf{C}_0)||_2 \\
=& c_5\sqrt{2m}\sqrt{\frac{\log p_n}{n}} \sum_{j=1}^p ||\hat{\boldsymbol{\eta}}_j||_2,
\end{aligned}$$

where $\hat{\boldsymbol{\eta}}_j = \hat{\mathbf{C}}^T \hat{B}_j - \mathbf{C}_0^T B_{0j}$.

Next, with the condition C1 and C5, there exists a constant c_6 such that with high probability

$$\left| \frac{\sum_{k=1}^n [y_k g(\omega^T \mathbf{x}_k) - E(yg(\omega^T \mathbf{x}))]}{n} \right| \leq c_6(\frac{\log p_n}{n})^{1/2}.$$

$$
\begin{aligned}
|T_2| \leq & \sum_{j=1}^{p} \left| \mathbf{e}_j^T (\hat{\mathbf{B}}\hat{\mathbf{C}} - \mathbf{B}_0\mathbf{C}_0)(\hat{\boldsymbol{\xi}} - \mathbf{B}_0\mathbf{C}_0)^T \boldsymbol{\Sigma} \mathbf{e}_j \right| \\
\leq & \sum_{j=1}^{p} ||\mathbf{e}_j^T (\hat{\mathbf{B}}\hat{\mathbf{C}} - \mathbf{B}_0\mathbf{C}_0)||_2 ||(\hat{\boldsymbol{\xi}} - \mathbf{B}_0\mathbf{C}_0)^T \boldsymbol{\Sigma} \mathbf{e}_j||_2 \\
\leq & c_6 \sqrt{2m}\sigma_u \sqrt{\frac{\log p_n}{n}} \sum_{j=1}^{p} ||\mathbf{e}_j^T (\hat{\mathbf{B}}\hat{\mathbf{C}} - \mathbf{B}_0\mathbf{C}_0)||_2 \\
= & c_6 \sqrt{2m}\sigma_u \sqrt{\frac{\log p_n}{n}} \sum_{j=1}^{p} ||\hat{\boldsymbol{\eta}}_j||_2.
\end{aligned}
$$

Put the up bounds of $|T_1|$ and $|T_2|$ to equation (B.1), we have

$$
\begin{aligned}
& \frac{1}{2} T_3 + \lambda \sum_{j=1}^{p} \theta_j ||\hat{\boldsymbol{\eta}}_j||_2 \\
\leq & \lambda \sum_{j=1}^{p} \theta_j ||B_{0j}||_2 - \lambda \sum_{j=1}^{p} \theta_j ||\hat{B}_j||_2 + \lambda \sum_{j=1}^{p} \theta_j ||\hat{\boldsymbol{\eta}}_j||_2 + \tilde{c} \sqrt{\frac{\log p_n}{n}} \sum_{j=1}^{p} ||\hat{\boldsymbol{\eta}}_j||_2,
\end{aligned}
$$

where $\tilde{c} = c_7\sqrt{m}$ and $c_7 = \sqrt{2}(c_5 + c_6\sigma_u)$. Now, we know for every $j \in \mathcal{A}$, $\left| ||B_{0j}||_2 - ||\hat{B}_j||_2 \right| \leq ||\hat{\boldsymbol{\eta}}_j||$ and for every $j \in \mathcal{A}^c$, $||\hat{B}_j||_2 = ||\hat{\boldsymbol{\eta}}_j||$. Then

$$
\frac{1}{2} T_3 + \lambda \sum_{j=1}^{p} \theta_j ||\hat{\boldsymbol{\eta}}_j||_2 \leq 2\lambda \sum_{j \in \mathcal{A}} \theta_j ||\hat{\boldsymbol{\eta}}_j||_2 + \tilde{c} \sqrt{\frac{\log p_n}{n}} \sum_{j=1}^{p} ||\hat{\boldsymbol{\eta}}_j||_2,
$$

which equals to

$$
\frac{1}{2} T_3 + \sum_{j \in \mathcal{A}^c} (\lambda\theta_j - \tilde{c}\sqrt{\frac{\log p_n}{n}}) ||\hat{\boldsymbol{\eta}}_j||_2 \leq \sum_{j \in \mathcal{A}} (\lambda\theta_j + \tilde{c}\sqrt{\frac{\log p_n}{n}}) ||\hat{\boldsymbol{\eta}}_j||_2.
$$

Choosing $\lambda = 2^{1-\rho} c_7 C_\phi^{\rho/2} \sqrt{m} \sqrt{\frac{\log p_n}{n^{1+\rho\phi}}}$, and $2\rho(\eta - \phi/2) > 1 - 2\eta$, then under conditions C3 and C4, we have

$$
\lambda\theta_j \leq 2\tilde{c}\sqrt{\frac{\log p_n}{n}}, \forall j \in \mathcal{A} \quad \text{and} \quad \lambda\theta_j \geq 2\tilde{c}\sqrt{\frac{\log p_n}{n}}, \forall j \in \mathcal{A}^c,
$$

which implies

$$
\frac{1}{2} \sum_{j \in \mathcal{A}^c} \lambda\theta_j ||\hat{\boldsymbol{\eta}}_j||_2 \leq \frac{1}{2} \sum_{j \in \mathcal{A}^c} \lambda\theta_j ||\hat{\boldsymbol{\eta}}_j||_2 + \frac{1}{2} T_3 \leq 3 \sum_{j \in \mathcal{A}} \tilde{c} \sqrt{\frac{\log p_n}{n}} ||\hat{\boldsymbol{\eta}}_j||_2,
$$

and

$$\sum_{j\in\mathcal{A}^c}||\hat{\boldsymbol{\eta}}_j||_2 \leq 3\sum_{j\in\mathcal{A}}||\hat{\boldsymbol{\eta}}_j||_2.$$

Let $T_0 = \text{tr}[(\hat{\mathbf{B}}\hat{\mathbf{C}} - \mathbf{B}_0\mathbf{C}_0)^T\boldsymbol{\Sigma}(\hat{\mathbf{B}}\hat{\mathbf{C}} - \mathbf{B}_0\mathbf{C}_0)]$, we have

$$\begin{aligned}
|T_3 - T_0| \leq & |\sum_{j=1}^{p}\sum_{k=1}^{p}(\hat{\sigma}_{jk} - \sigma_{jk})\hat{\boldsymbol{\eta}}_j^T\hat{\boldsymbol{\eta}}_k| \\
\leq & c_5\sqrt{\frac{\log p_n}{n}}[|\sum_{j\in\mathcal{A}}\sum_{k\in\mathcal{A}}\hat{\boldsymbol{\eta}}_j^T\hat{\boldsymbol{\eta}}_k| + |\sum_{j\in\mathcal{A}^c}\sum_{k\in\mathcal{A}^c}\hat{\boldsymbol{\eta}}_j^T\hat{\boldsymbol{\eta}}_k| + 2|\sum_{j\in\mathcal{A}}\sum_{k\in\mathcal{A}^c}\hat{\boldsymbol{\eta}}_j^T\hat{\boldsymbol{\eta}}_k|] \\
\leq & c_5\sqrt{\frac{\log p_n}{n}}[(\sum_{j\in\mathcal{A}}||\hat{\boldsymbol{\eta}}_j||_2)^2 + (3\sum_{j\in\mathcal{A}}||\hat{\boldsymbol{\eta}}_j||_2)^2 + 6(\sum_{j\in\mathcal{A}}||\hat{\boldsymbol{\eta}}_j||_2)^2] \\
= & 16c_5\sqrt{\frac{\log p_n}{n}}(\sum_{j\in\mathcal{A}}||\hat{\boldsymbol{\eta}}_j||_2)^2.
\end{aligned}$$

Now, let $\hat{\mathcal{A}}_0$ denote the index set in $\mathcal{A}^c$ that corresponds to the s largest $||\hat{\boldsymbol{\eta}}_j||_2$ with $j \in \mathcal{A}^c$. Define $\tilde{\mathcal{A}} = \mathcal{A} \cup \hat{\mathcal{A}}_0$,

$$\begin{aligned}
T_0 \leq & |T_0 - T_3| + T_3 \\
\leq & 32c_5\sqrt{\frac{\log p_n}{n}}(\sum_{j\in\mathcal{A}}||\hat{\boldsymbol{\eta}}_j||_2)^2 + 6\tilde{c}\sqrt{\frac{\log p_n}{n}}\sum_{j\in\mathcal{A}}||\hat{\boldsymbol{\eta}}_j||_2 \\
\leq & 64c_5 s\sqrt{\frac{\log p_n}{n}}\sum_{j\in\tilde{\mathcal{A}}}||\hat{\boldsymbol{\eta}}_j||_2^2 + 6\tilde{c}(\frac{2s\log p_n}{n}\sum_{j\in\tilde{\mathcal{A}}}||\hat{\boldsymbol{\eta}}_j||_2^2)^{1/2}.
\end{aligned}$$

And it results in

$$(\sum_{j\in\tilde{\mathcal{A}}}||\hat{\boldsymbol{\eta}}_j||_2^2)^{1/2} \leq \frac{6\tilde{c}\sqrt{\frac{2s\log p_n}{n}}}{T_0/\sum_{j\in\tilde{\mathcal{A}}}||\hat{\boldsymbol{\eta}}_j||_2^2 - 64c_5 s\sqrt{\frac{\log p_n}{n}}} \leq \frac{12\tilde{c}}{\sigma_l}\sqrt{\frac{s\log p_n}{n}}.$$

Furthermore, we have $\sum_{j\in\mathcal{A}^c\setminus\hat{\mathcal{A}}_0}||\hat{\boldsymbol{\eta}}_j||_2^2 \leq 9\sum_{j\in\tilde{A}}||\hat{\boldsymbol{\eta}}_j||_2^2$ follow the exact same steps in Qian et al.'s [46] proof. Thus we have

$$||\hat{\mathbf{B}}\hat{\mathbf{C}} - \mathbf{B}_0\mathbf{C}_0||_F \leq \frac{48\tilde{c}}{\sigma_l}\sqrt{\frac{s\log p_n}{n}},$$

and by Wedin's theorem,

$$||P_{\mathcal{S}_{\hat{B}}} - P_{\mathcal{S}_{E(y|\mathbf{x})}}||_F = O_p(\sqrt{sm}\sqrt{\frac{\log p_n}{n}}).$$

The second statement in Theorem 3.3.1 can be proved similar to the proof of Qian et al. [46] and thus we omit it. □

Proof of Theorem 3.3.2

Let $(\hat{\mathbf{B}}, \hat{\mathbf{C}})$ be the minimizer of objective function 3.7 and $(\mathbf{B}_0, \mathbf{C}_0)$ be the minimizer of objective function without the penalty term, with $\mathbf{B}_0 = (B_{01}, \cdots, B_{0p})^T$ and $\hat{\mathbf{B}} = (\hat{B}_1, \cdots, \hat{B}_p)^T$. Then, we have $L_I(\hat{\mathbf{B}}, \hat{\mathbf{C}}) \leq L_I(\mathbf{B}_0, \mathbf{C}_0)$. That is,

$$\frac{1}{2}\text{tr}[(\hat{\boldsymbol{\xi}}-\hat{\mathbf{B}}\hat{\mathbf{C}})^T(\hat{\boldsymbol{\xi}}-\hat{\mathbf{B}}\hat{\mathbf{C}})]+\lambda\sum_{j=1}^{p}\theta_j||\hat{B}_j||_2 \leq \frac{1}{2}\text{tr}[(\hat{\boldsymbol{\xi}}-\hat{\mathbf{B}}_0\hat{\mathbf{C}}_0)^T(\hat{\boldsymbol{\xi}}-\hat{\mathbf{B}}_0\hat{\mathbf{C}}_0)]+\lambda\sum_{j=1}^{p}\theta_j||B_{0j}||_2.$$

After simplification, we have

$$T_{I1} + \frac{1}{2}T_{I3} + \lambda\sum_{j=1}^{p}\theta_j||\hat{B}_j||_2 \leq \lambda\sum_{j=1}^{p}\theta_j||B_{0j}||_2, \tag{B.2}$$

where $T_{I1} = -\text{tr}[(\hat{\boldsymbol{\xi}} - \mathbf{B}_0\mathbf{C}_0)^T(\hat{\mathbf{B}}\hat{\mathbf{C}} - \mathbf{B}_0\mathbf{C}_0)]$, and

$$T_{I3} = \text{tr}[(\hat{\mathbf{B}}\hat{\mathbf{C}} - \mathbf{B}_0\mathbf{C}_0)^T(\hat{\mathbf{B}}\hat{\mathbf{C}} - \mathbf{B}_0\mathbf{C}_0)].$$

Similar to the proof for Theorem 3.3.1, given condition C1, we have for a constant c_I

$$\begin{aligned}
|T_{I1}| \leq & \sum_{j=1}^{p}\left|\mathbf{e}_j^T(\hat{\mathbf{B}}\hat{\mathbf{C}} - \mathbf{B}_0\mathbf{C}_0)(\hat{\boldsymbol{\xi}} - \mathbf{B}_0\mathbf{C}_0)^T\mathbf{e}_j\right| \\
\leq & \sum_{j=1}^{p}||\mathbf{e}_j^T(\hat{\mathbf{B}}\hat{\mathbf{C}} - \mathbf{B}_0\mathbf{C}_0)||_2||(\hat{\boldsymbol{\xi}} - \mathbf{B}_0\mathbf{C}_0)^T\mathbf{e}_j||_2 \\
\leq & c_I\sqrt{2m}\sqrt{\frac{\log p_n}{n}}\sum_{j=1}^{p}||\mathbf{e}_j(\hat{\mathbf{B}}\hat{\mathbf{C}} - \mathbf{B}_0\mathbf{C}_0)||_2 \\
= & c_I\sqrt{2m}\sqrt{\frac{\log p_n}{n}}\sum_{j=1}^{p}||\hat{\boldsymbol{\eta}}_j||_2 \\
= & c^*\sqrt{\frac{\log p_n}{n}}\sum_{j=1}^{p}||\hat{\boldsymbol{\eta}}_j||_2.
\end{aligned}$$

Plug this inequality back to (B.2), we have

$$\begin{aligned}
&\frac{1}{2}T_{I2} + \lambda\sum_{j=1}^{p}\theta_j||\hat{\boldsymbol{\eta}}_j||_2 \\
&\leq \lambda\sum_{j=1}^{p}\theta_j||B_{0j}||_2 - \lambda\sum_{j=1}^{p}\theta_j||\hat{B}_j||_2 + \lambda\sum_{j=1}^{p}\theta_j||\hat{\boldsymbol{\eta}}_j||_2 + c^*\sqrt{\frac{\log p_n}{n}}\sum_{j=1}^{p}||\hat{\boldsymbol{\eta}}_j||_2.
\end{aligned}$$

Now under conditions C3 and C4, choose $\lambda = 2^{1-\rho}c^*C_\phi^{\rho/2}\sqrt{\frac{\log p_n}{n^{1+\rho\phi}}}$, and $2\rho(\eta - \phi/2) > 1 - 2\eta$, we have

$$\frac{1}{2}\sum_{j\in\mathcal{A}^c}\lambda\theta_j||\hat{\boldsymbol{\eta}}_j||_2 \leq \frac{1}{2}\sum_{j\in\mathcal{A}^c}\lambda\theta_j||\hat{\boldsymbol{\eta}}_j||_2 + \frac{1}{2}T_{I2} \leq 3\sum_{j\in\mathcal{A}}c^*\sqrt{\frac{\log p_n}{n}}||\hat{\boldsymbol{\eta}}_j||_2,$$

and it results in

$$\sum_{j\in\mathcal{A}^c}||\hat{\boldsymbol{\eta}}_j||_2 \leq 3\sum_{j\in\mathcal{A}}||\hat{\boldsymbol{\eta}}_j||_2.$$

Now, let $\hat{\mathcal{A}}_0$ denote the index subset from $\mathcal{A}^c$ that corresponds to the s largest $||\hat{\boldsymbol{\eta}}_j||_2$ with $j \in \mathcal{A}^c$. Define $\tilde{\mathcal{A}} = \mathcal{A} \cup \hat{\mathcal{A}}_0$, then

$$\sum_{j\in\tilde{\mathcal{A}}}||\hat{\boldsymbol{\eta}}_j||_2^2 \leq T_{I2} \leq 6\sum_{j\in\mathcal{A}}c^*\sqrt{\frac{\log p_n}{n}}||\hat{\boldsymbol{\eta}}_j||_2 \leq 6c^*(\frac{2s\log p_n}{n}\sum_{j\in\tilde{\mathcal{A}}}||\hat{\boldsymbol{\eta}}_j||_2^2)^{1/2}.$$

After simplification, we get

$$(\sum_{j\in\tilde{\mathcal{A}}}||\hat{\boldsymbol{\eta}}_j||_2^2)^{1/2} \leq 12c^*\sqrt{\frac{s\log p_n}{n}}.$$

Thus, we have

$$||\hat{\mathbf{B}}\hat{\mathbf{C}} - \mathbf{B}_0\mathbf{C}_0||_F \leq 48c^*\sqrt{\frac{s\log p_n}{n}}$$

and by Wedin's theorem,

$$||P_{\mathcal{S}_{\hat{B}}} - P_{\mathcal{S}_{E(y|\mathbf{x})}}||_F = O_p(\sqrt{m}\sqrt{\frac{s\log p_n}{n}}).$$

□

C Supplementary Materials for Chapter 4

C1 Derivatives of $\ln(\mathbf{A}, \mathbf{G})$

The first derivative of $\ln(\mathbf{A}, \mathbf{G})$ to $\text{vec}(\mathbf{A})$ is $D_{a1} = \frac{\partial \ln(\mathbf{A},\mathbf{G})}{\partial \text{vec}(\mathbf{A})} = \sum_{i=1}^n\sum_{j=1}^{n_i}\frac{\partial L_{ij}}{\partial \text{vec}(\mathbf{A})}$, where for $X_{i,j-1} = 1, 2, 3$,

$$\frac{\partial L_{ij}}{\partial \text{vec}(\mathbf{A})} = -\frac{\sum_{l=1,l\neq X_{i,j-1}}^7 e^{\mathbf{z}_i^T\mathbf{A}\mathbf{G}_b}(\mathbf{G}_b\otimes\mathbf{z}_i)}{1+\sum_{l=1,l\neq X_{i,j-1}}^7 e^{\mathbf{z}_i^T\mathbf{A}\mathbf{G}_b}} + (1-\delta_{X_{i,j-1},X_{ij}})(\mathbf{G}_b\otimes\mathbf{z}_i);$$

and for $X_{i,j-1} = 4$,

$$\frac{\partial L_{ij}}{\partial \text{vec}(\mathbf{A})} = -\frac{\sum_{l=5}^{7} e^{\mathbf{z}_i^T \mathbf{A}\mathbf{G}_{(14+l)}}(\mathbf{G}_{(14+l)} \otimes \mathbf{z}_i)}{1 + \sum_{l=5}^{7} e^{\mathbf{z}_i^T \mathbf{A}\mathbf{G}_{(14+l)}}} + (1 - \delta_{4,X_{ij}})(\mathbf{G}_{(14+l)} \otimes \mathbf{z}_i).$$

The correspondent second derivative is

$$D_{a2} = \frac{\partial^2 \ln(\mathbf{A}, \mathbf{G})}{\partial \text{vec}(\mathbf{A}) \partial \text{vec}(\mathbf{A})^T} = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \frac{\partial^2 L_{ij}}{\partial \text{vec}(\mathbf{A}) \partial \text{vec}(\mathbf{A})^T}.$$

Here, for $X_{i,j-1} = 1, 2, 3$,

$$\begin{aligned}
&\frac{\partial^2 L_{ij}}{\partial \text{vec}(\mathbf{A}) \partial \text{vec}(\mathbf{A})^T} \\
&= -\frac{\sum_{l=1, l \neq X_{i,j-1}}^{7} e^{\mathbf{z}_i^T \mathbf{A}\mathbf{G}_b}(\mathbf{G}_b \otimes \mathbf{z}_i)(\mathbf{G}_b^T \otimes \mathbf{z}_i^T)(1 + \sum_{l=1, l \neq X_{i,j-1}}^{7} e^{\mathbf{z}_i^T \mathbf{A}\mathbf{G}_b})}{(1 + \sum_{l=1,, l \neq X_{i,j-1}}^{7} e^{\mathbf{z}_i^T \mathbf{A}\mathbf{G}_b})^2} \\
&+ \frac{\sum_{l=1, l \neq X_{i,j-1}}^{7} e^{\mathbf{z}_i^T \mathbf{A}\mathbf{G}_b}(\mathbf{G}_b \otimes \mathbf{z}_i) \sum_{l=1, l \neq X_{i,j-1}}^{7} e^{\mathbf{z}_i^T \mathbf{A}\mathbf{G}_b}(\mathbf{G}_b^T \otimes \mathbf{z}_i^T)}{(1 + \sum_{l=1, l \neq X_{i,j-1}}^{7} e^{\mathbf{z}_i^T \mathbf{A}\mathbf{G}_b})^2};
\end{aligned}$$

for $X_{i,j-1} = 4$,

$$\begin{aligned}
&\frac{\partial^2 L_{ij}}{\partial \text{vec}(\mathbf{A}) \partial \text{vec}(\mathbf{A})^T} \\
&= -\frac{\sum_{l=5}^{7} e^{\mathbf{z}_i^T \mathbf{A}\mathbf{G}_{(14+l)}}(\mathbf{G}_{(14+l)} \otimes \mathbf{z}_i)(\mathbf{G}_{(14+l)}^T \otimes \mathbf{z}_i^T)(1 + \sum_{l=5}^{7} e^{\mathbf{z}_i^T \mathbf{A}\mathbf{G}_{(14+l)}})}{(1 + \sum_{l=5}^{7} e^{\mathbf{z}_i^T \mathbf{A}\mathbf{G}_{(14+l)}})^2} \\
&+ \frac{\sum_{l=5}^{7} e^{\mathbf{z}_i^T \mathbf{A}\mathbf{G}_{(14+l)}}(\mathbf{G}_{(14+l)} \otimes \mathbf{z}_i) \sum_{l=5}^{7} e^{\mathbf{z}_i^T \mathbf{A}\mathbf{G}_{(14+l)}}(\mathbf{G}_{(14+l)}^T \otimes \mathbf{z}_i^T)}{(1 + \sum_{l=5}^{7} e^{\mathbf{z}_i^T \mathbf{A}\mathbf{G}_{(14+l)}})^2};
\end{aligned}$$

where $b = 6(X_{i,j-1} - 1) + l$ for $X_{i,j-1} > l$ and $b = 6(X_{i,j-1} - 1) - u + l$ for $X_{i,j-1} < l$.

C2 Additional Results

C2.1: Additional results for RR-MLRfMC in the Application

In this section, we also reported the partial results when considering the intercept and the 8 adjusting covariates. In the model with 8 adjusting covariates in reduced rank and fixed intercept, the partial results are reported in Table C1 and rank 2 is the best. In the model with 9 adjusting covariates (including the intercept), the results are shown in Table C2 and rank 3 is the best. Since we are interested in interpreting the effect of APOE 4, we do not report further results for these two models.

C2.2: Additional results for PRR-MLRfMC in the Application

In PRR-MLRdMC model, if we included the intercept in the RR part instead of treating it as fixed, then we have the results reported in Table C3. Note that here, rather than rank 2, rank 3 is optimal since it fit the data best. We also reported the coefficient matrix along with **A** and **G** in Tables C4–C7. We can make the similar conclusion to the model with intercept as fixed. APOE 4 is found to be a significant predictor of a transition from normal cognition to dementia ($P = 0.024$) and from mixed MCI to Normal ($P = 0.011$).

Table C1: Fit statistics using 8 adjusting covariates, fixed intercept

Rank T	Log-likelihood	Number of Parameters	AIC
1	-6819.16	49	13736.31
2	-6787.23	75	**13724.45**
3	-6765.11	99	13728.23
4	-6752.56	121	13747.12
5	-6741.25	141	13764.50
6	-6733.18	159	13784.36
7	-6730.24	175	13810.48
8	-6728.33	189	13834.33

Table C2: Fit statistics using 9 adjusting covariates (include intercept)

Rank T	Log-likelihood	Number of Parameters	AIC
1	-6891.06	29	13840.11
2	-6813.53	56	13739.06
3	-6785.55	81	**13733.11**
4	-6763.43	104	13734.43
5	-6752.17	125	13754.34
6	-6741.08	144	13770.17
7	-6732.98	161	13787.97
8	-6730.08	176	13812.16
9	-6728.33	189	13834.66

Table C3: Fit statistics using 9 adjusting covariates, and fixed APOE4

Rank T	Log-likelihood	Number of Parameters	AIC
1	-6873.02	50	13846.05
2	-6795.72	77	13745.44
3	-6766.73	102	**13737.45**
4	-6744.23	125	13738.45
5	-6732.18	146	13756.37
6	-6722.11	165	13774.22
7	-6712.73	182	13789.46
8	-6710.69	197	13815.39
9	-6709.02	210	13838.03

Table C4: Parameter Estimates for Rank 3 Model with Normal as Prior State

Covariates	A-MCI	M-MCI	MCI	Dementia	Dropout	Death	$\mathbf{A}_1$	$\mathbf{A}_2$	$\mathbf{A}_3$
APOE4	-0.176	-0.228	0.440	1.185	-0.315	0.075			
se	0.109	0.119	0.328	0.526	0.249	0.375			
P	0.106	0.055	0.180	**0.024**	0.207	0.842			
intercept	-1.452	-1.597	-4.501	-5.266	-3.728	-5.227	-0.994	-0.074	0.000
family history	-0.064	-0.096	-0.127	-0.255	-0.174	-0.111	-0.032	-0.033	0.086
high BP	-0.031	-0.004	0.292	-0.487	0.436	0.782	0.042	-0.501	-0.101
$<$ 10 pack years	-0.151	-0.252	0.073	-0.880	-0.104	0.518	-0.015	-0.446	0.295
11-19 pack years	0.019	-0.038	0.262	-0.017	0.065	0.429	0.046	-0.110	0.202
$\geq$ 20 pack years	0.005	0.112	0.399	-0.464	0.782	1.011	0.067	-0.647	-0.360
low education	0.091	0.344	-0.320	0.482	0.422	-0.640	-0.036	0.194	-0.832
baseline age	0.033	0.073	0.023	0.134	0.123	-0.002	0.010	0.018	-0.123
head injury	-0.014	0.024	0.132	-0.253	0.282	0.387	0.019	-0.275	-0.131
	1.445	1.590	4.544	5.184	3.779	5.344			
	0.210	0.228	-0.226	1.502	-0.397	-1.163			
	-0.123	-0.429	0.137	-0.451	-0.761	0.270			
	$\mathbf{G}_1$	$\mathbf{G}_2$	$\mathbf{G}_3$	$\mathbf{G}_4$	$\mathbf{G}_5$	$\mathbf{G}_6$			

Table C5: Parameter Estimates for Rank 3 Model with A-MCI as Prior State

Covariates	Normal	M-MCI	MCI	Dementia	Dropout	Death	$\mathbf{A}_1$	$\mathbf{A}_2$	$\mathbf{A}_3$
APOE4	0.051	0.227	0.482	0.924	0.005	-1.381			
se	0.205	0.262	0.474	0.698	0.548	1.056			
P	0.803	0.386	0.309	0.186	0.992	0.191			
intercept	1.168	-0.454	-1.729	-3.304	-2.581	-2.645	-0.994	-0.074	0.000
family history	0.050	-0.024	-0.074	-0.129	-0.075	-0.058	-0.032	-0.033	0.086
high BP	-0.156	0.086	-0.467	-0.015	0.383	0.290	0.042	-0.501	-0.101
$<$ 10 pack years	0.003	-0.002	-0.442	-0.239	0.169	0.189	-0.015	-0.446	0.295
11-19 pack years	-0.032	0.004	-0.006	0.088	0.162	0.205	0.046	-0.110	0.202
$\geq$ 20 pack years	-0.262	0.150	-0.620	0.050	0.545	0.363	0.067	-0.647	-0.360
low education	-0.096	0.081	0.011	0.059	-0.123	-0.328	-0.036	0.194	-0.832
baseline age	-0.034	0.020	0.016	0.055	0.027	-0.004	0.010	0.018	-0.123
head injury	-0.096	0.056	-0.275	-0.011	0.206	0.134	0.019	-0.275	-0.131
	-1.188	0.465	1.661	3.298	2.635	2.690			
	0.169	-0.105	1.041	0.336	-0.524	-0.390			
	0.205	-0.142	0.158	-0.133	-0.086	0.188			
	$\mathbf{G}_7$	$\mathbf{G}_8$	$\mathbf{G}_9$	$\mathbf{G}_{10}$	$\mathbf{G}_{11}$	$\mathbf{G}_{12}$			

Table C6: Parameter Estimates for Rank 3 Model with M-MCI as Prior State

Covariates	Normal	A-MCI	MCI	Dementia	Dropout	Death	$\mathbf{A}_1$	$\mathbf{A}_2$	$\mathbf{A}_3$
APOE4	-0.381	-0.089	0.319	0.458	-0.343	0.302			
se	0.150	0.243	0.271	0.358	0.284	0.390			
P	**0.011**	0.716	0.239	0.202	0.227	0.440			
intercept	-0.395	-1.849	-2.279	-3.091	-2.408	-2.862	-0.994	-0.074	0.000
family history	0.022	-0.004	-0.077	-0.070	-0.076	-0.050	-0.032	-0.033	0.086
high BP	0.197	0.204	-0.033	0.197	0.386	-0.088	0.042	-0.501	-0.101
$<$ 10 pack years	0.253	0.283	-0.127	0.120	0.149	0.003	-0.015	-0.446	0.295
11-19 pack years	0.117	0.231	0.089	0.222	0.138	0.221	0.046	-0.110	0.202
$\geq$ 20 pack years	0.200	0.166	-0.028	0.226	0.566	-0.218	0.067	-0.647	-0.360
low education	-0.311	-0.570	-0.080	-0.384	-0.051	-0.546	-0.036	0.194	-0.832
baseline age	-0.035	-0.052	0.020	-0.008	0.035	-0.041	0.010	0.018	-0.123
head injury	0.087	0.065	-0.031	0.074	0.215	-0.105	0.019	-0.275	-0.131
	0.428	1.885	2.274	3.122	2.461	2.856			
	-0.410	-0.353	0.246	-0.187	-0.529	0.295			
	0.260	0.521	0.056	0.285	-0.167	0.603			
	$\mathbf{G}_{13}$	$\mathbf{G}_{14}$	$\mathbf{G}_{15}$	$\mathbf{G}_{16}$	$\mathbf{G}_{17}$	$\mathbf{G}_{18}$			

Table C7: Parameter Estimates for Rank 3 Model with MCI as Prior State

Covariates	Normal	A-MCI	M-MCI	Dementia	Dropout	Death	$\mathbf{A}_1$	$\mathbf{A}_2$	$\mathbf{A}_3$
APOE4	/	/	/	0.676	0.038	0.189			
se	/	/	/	0.346	0.435	0.578			
P	/	/	/	0.051	0.930	0.743			
intercept	/	/	/	-0.800	-2.065	-1.838	-0.994	-0.074	0.000
family history	/	/	/	-0.048	-0.066	-0.039	-0.032	-0.033	0.086
high BP	/	/	/	-0.603	0.336	-0.417	0.042	-0.501	-0.101
$<$ 10 pack years	/	/	/	-0.506	0.130	-0.252	-0.015	-0.446	0.295
11-19 pack years	/	/	/	-0.066	0.118	0.098	0.046	-0.110	0.202
$\geq$ 20 pack years	/	/	/	-0.810	0.494	-0.644	0.067	-0.647	-0.360
low education	/	/	/	0.066	-0.039	-0.355	-0.036	0.194	-0.832
baseline age	/	/	/	0.008	0.031	-0.035	0.010	0.018	-0.123
head injury	/	/	/	-0.346	0.187	-0.277	0.019	-0.275	-0.131
	/	/	/	0.713	2.111	1.783			
	/	/	/	1.228	-0.462	0.871			
	/	/	/	0.176	-0.151	0.553			
	/	/	/	$\mathbf{G}_{19}$	$\mathbf{G}_{20}$	$\mathbf{G}_{21}$			

Bibliography

[1] Abner, E. L., Nelson, P. T., Schmitt, F. A., Browning, S. R., Fardo, D. W., Wan, L., Jicha, G. A., Cooper, G. E., Smith, C. D., Caban-Holt, A. M., et al. Self-reported head injury and risk of late-life impairment and ad pathology in an ad center cohort. *Dement Geriatr Cogn Disord. 37*, 5-6 (2014), 294–306.

[2] Akaike, H. A new look at the statistical model identification. *IEEE Trans Automat Contr. 19*, 6 (1974), 716–723.

[3] Anderson, J. A. Regression and ordered categorical variables. *Journal of the Royal Statistical Society: Series B (Methodological) 46*, 1 (1984), 1–22.

[4] Anderson, T. W., et al. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics 22*, 3 (1951), 327–351.

[5] Aragon, Y. A gauss implementation of multivariate sliced inverse regression. *Computational Statistics 12*, 3 (1997), 355–372.

[6] Bernstein, S. On a modification of chebyshev's inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math 1*, 4 (1924), 38–49.

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

[8] Bura, E., and Cook, R. D. Extending sliced inverse regression: The weighted chi-squared test. *Journal of the American Statistical Association 96*, 455 (2001), 996–1003.

[9] Chen, X., Zou, C., Cook, R. D., et al. Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics 38*, 6 (2010), 3696–3723.

[10] Cole, B. F., Bonetti, M., Zaslavsky, A. M., and Gelber, R. D. A multistate markov chain model for longitudinal, categorical quality-of-life data subject to non-ignorable missingness. *Statist Med. 24*, 15 (2005), 2317–2334.

[11] Cook, R. D. On the interpretation of regression plots. *Journal of the American Statistical Association 89*, 425 (1994), 177–189.

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

[13] Cook, R. D. *Regression graphics: Ideas for studying regressions through graphics*, vol. 482. John Wiley & Sons, 2009.

[14] Cook, R. D., et al. Fisher lecture: Dimension reduction in regression. *Statistical Science 22*, 1 (2007), 1–26.

[15] Cook, R. D., Li, B., et al. Dimension reduction for conditional mean in regression. *The Annals of Statistics 30*, 2 (2002), 455–474.

[16] Cook, R. D., and Ni, L. Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association 100*, 470 (2005), 410–428.

[17] Cook, R. D., and Setodji, C. M. A model-free test for reduced rank in multivariate regression. *Journal of the American Statistical Association 98*, 462 (2003), 340–351.

[18] Cook, R. D., and Weisberg, S. Discussion of sliced inverse regression for dimension reduction. *Journal of the American Statistical Association 86*, 414 (1991), 328–332.

[19] Cook, R. D., and Yin, X. Theory & methods: special invited paper: dimension reduction and visualization in discriminant analysis (with discussion). *Australian & New Zealand Journal of Statistics 43*, 2 (2001), 147–199.

[20] Davison, A. C., and Hinkley, D. V. *Bootstrap methods and their application*. No. 1. Cambridge university press, 1997.

[21] Eaton, M. L., and Tyler, D. The asymptotic distribution of singular-values with applications to canonical correlations and correspondence analysis. *Journal of Multivariate Analysis 50*, 2 (1994), 238–264.

[22] Fan, J., and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association 96*, 456 (2001), 1348–1360.

[23] Fan, J., and Lv, J. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*, 5 (2008), 849–911.

[24] Fan, J., Samworth, R., and Wu, Y. Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research 10* (2009), 2013–2038.

[25] Fiocco, M., Putter, H., and van Houwelingen, H. C. Reduced-rank proportional hazards regression and simulation-based prediction for multi-state models. *Statistics in Medicine 27*, 21 (2008), 4340–4358.

[26] Fiocco, M., Putter, H., and Van Houwelingen, J. Reduced rank proportional hazards model for competing risks. *Biostatistics. 6*, 3 (2005), 465–478.

[27] Goldberg, Y., Lu, W., Fine, J., et al. Oracle estimation of parametric transformation models. *Electron J Stat. 10*, 1 (2016), 90–120.

[28] Goodman, L. A. Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *Journal of the American Statistical Association 76*, 374 (1981), 320–334.

[29] Heckman, J. J. *The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process and some Monte Carlo evidence*. Cambridge,MA: MIT Press; 1987.

[30] Hoerl, A. E., and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*, 1 (1970), 55–67.

[31] Hooper, J. W. Simultaneous equations and canonical correlation theory. *Econometrica: Journal of the Econometric Society* (1959), 245–256.

[32] Hotelling, H. Relations between two sets of variates. *Biometrika* (1936), 321–377.

[33] Huang, J., Horowitz, J. L., Ma, S., et al. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics 36*, 2 (2008), 587–613.

[34] Izenman, A. J. Reduced-rank regression for the multivariate linear model. *J Multivar Anal. 5*, 2 (1975), 248–264.

[35] Li, B., and Wang, S. On directional regression for dimension reduction. *Journal of the American Statistical Association 102*, 479 (2007), 997–1008.

[36] Li, B., Wen, S., and Zhu, L. On a projective resampling method for dimension reduction with multivariate responses. *Journal of the American Statistical Association 103*, 483 (2008), 1177–1186.

[37] Li, B., Zha, H., Chiaromonte, F., et al. Contour regression: a general approach to dimension reduction. *The Annals of Statistics 33*, 4 (2005), 1580–1616.

[38] Li, K.-C. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association 86*, 414 (1991), 316–327.

[39] Li, K.-C. On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association 87*, 420 (1992), 1025–1039.

[40] Li, L. Sparse sufficient dimension reduction. *Biometrika 94*, 3 (2007), 603–613.

[41] Lunt, M. Prediction of ordinal outcomes when the association between predictors and outcome differs between outcome levels. *Statist Med. 24*, 9 (2005), 1357–1369.

[42] Luo, W., and Li, B. Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika 103*, 4 (2016), 875–887.

[43] Luo, W., Li, B., Yin, X., et al. On efficient dimension reduction with respect to a statistical functional of interest. *Annals of Statistics 42*, 1 (2014), 382–412.

[44] Ni, L., and Cook, R. D. A robust inverse regression estimator. *Statistics & probability letters 77*, 3 (2007), 343–349.

[45] Park, T., Shao, X., Yao, S., et al. Partial martingale difference correlation. *Electronic Journal of Statistics 9*, 1 (2015), 1492–1517.

[46] Qian, W., Ding, S., and Cook, R. D. Sparse minimum discrepancy approach to sufficient dimension reduction with simultaneous variable selection in ultrahigh dimension. *Journal of the American Statistical Association 114*, 527 (2019), 1277–1290.

[47] Raber, J., Huang, Y., and Ashford, J. W. Apoe genotype accounts for the vast majority of ad risk and ad pathology. *Neurobiol Aging. 25*, 5 (2004), 641–650.

[48] Salazar, J. C., Schmitt, F. A., Yu, L., Mendiondo, M. M., and Kryscio, R. J. Shared random effects analysis of multi-state markov models: application to a longitudinal study of transitions to dementia. *Statist Med. 26*, 3 (2007), 568–580.

[49] Schmidli, H. *Reduced rank regression: with applications to quantitative structure-activity relationships*. Berlin/Heidelberg, Germany: Springer Science & Business Media; 2013.

[50] Setodji, C. M., and Cook, R. D. K-means inverse regression. *Technometrics 46*, 4 (2004), 421–429.

[51] Shao, X., and Zhang, J. Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association 109*, 507 (2014), 1302–1318.

[52] Shapiro, A. Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association 81*, 393 (1986), 142–149.

[53] Sheng, W., and Yin, X. Direction estimation in single-index models via distance covariance. *Journal of Multivariate Analysis 122* (2013), 148–161.

[54] Sheng, W., and Yin, X. Sufficient dimension reduction via distance covariance. *Journal of Computational and Graphical Statistics 25*, 1 (2016), 91–104.

[55] Song, C., Kuo, L., Derby, C. A., Lipton, R. B., and Hall, C. B. Multi-stage transitional models with random effects and their application to the einstein aging study. *Biom J. 53*, 6 (2011), 938–955.

[56] Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. Measuring and testing dependence by correlation of distances. *The annals of statistics 35*, 6 (2007), 2769–2794.

[57] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*, 1 (1996), 267–288.

[58] Tyas, S. L., Salazar, J. C., Snowdon, D. A., Desrosiers, M. F., Riley, K. P., Mendiondo, M. S., and Kryscio, R. J. Transitions to mild cognitive impairments, dementia, and death: findings from the nun study. *Am J Epidemiol. 165*, 11 (2007), 1231–1238.

[59] Wang, H., Li, R., and Tsai, C.-L. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika 94*, 3 (2007), 553–568.

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

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

[62] Wang, Q., Yin, X., and Critchley, F. Dimension reduction based on the hellinger integral. *Biometrika 102*, 1 (2015), 95–106.

[63] Wang, Pei, Y. X. Y. Q., and Kryscio, R. Feature filter for estimating central mean subspace and its sparse solution. *Submitted*.

[64] Weng, J., and Yin, X. Fourier transformation and a minimum discrepancy approach to sufficient dimension reduction and sufficient variable selection in ultrahigh dimension. *Submitted*.

[65] Wold, H. Estimation of principal components and related models by iterative least squares. *Multivariate analysis* (1966), 391–420.

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

[67] Xia, Y., Tong, H., Li, W. K., and Zhu, L.-X. An adaptive estimation of dimension reduction space. In *Exploration Of A Nonlinear World: An Appreciation of Howell Tong's Contributions to Statistics*. World Scientific, 2009, pp. 299–346.

[68] Yang, B., Yin, X., and Zhang, N. Sufficient variable selection using independence measures for continuous response. *Journal of Multivariate Analysis 173* (2019), 480–493.

[69] Ye, Z., and Weiss, R. E. Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association 98*, 464 (2003), 968–979.

[70] Yee, T. W., and Hastie, T. J. Reduced-rank vector generalized linear models. *Statistical modelling 3*, 1 (2003), 15–41.

[71] Yin, X., and Bura, E. Moment-based dimension reduction for multivariate response regression. *Journal of Statistical Planning and Inference 136*, 10 (2006), 3675–3688.

[72] Yin, X., and Cook, R. D. Dimension reduction for the conditional kth moment in regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*, 2 (2002), 159–175.

[73] Yin, X., and Cook, R. D. Direction estimation in single-index regressions. *Biometrika 92*, 2 (2005), 371–384.

[74] Yin, X., and Hilafu, H. Sequential sufficient dimension reduction for large p, small n problems. *Journal of the Royal Statistical Society: Series B: Statistical Methodology* (2015), 879–892.

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

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

[77] Yu, L., Griffith, W. S., Tyas, S. L., Snowdon, D. A., and Kryscio, R. J. A nonstationary markov transition model for computing the relative risk of dementia before death. *Statist Med. 29*, 6 (2010), 639–648.

[78] Zhang, N., and Yin, X. Direction estimation in single-index regressions via hilbert-schmidt independence criterion. *Statistica Sinica* (2015), 743–758.

[79] Zhu, L.-P., and Zhu, L.-X. Dimension reduction for conditional variance in regressions. *Statistica Sinica* (2009), 869–883.

[80] Zhu, L.-P., Zhu, L.-X., and Wen, S.-Q. On dimension reduction in regressions with multivariate responses. *Statistica Sinica* (2010), 1291–1307.

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

[82] Zhu, Y., and Zeng, P. Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association 101*, 476 (2006), 1638–1651.

[83] Zou, H. The adaptive lasso and its oracle properties. *Journal of the American statistical association 101*, 476 (2006), 1418–1429.

Vita

Pei Wang

Education

- **M.S. in Statistics** University of Texas at El Paso, El Paso, TX, 2014-2016
- **B.S. in Information & Computational Science** Zhejiang University of Technology, Zhejiang, China, 2009-2013

Teaching Experience

- **University of Kentucky**
  Instructor Fall 17 - Fall 20
  Teaching Assistant Fall 16 & Spring 17 & Spring 21
- **The University of Texas at El Paso, El Paso, TX**
  Teaching Assistant Summer 14 - Spring 16

Professional Experience

- **Graduate Research Assistant**
  Department of Statistics, UKY Summer 17, 20 & Spring 19, 20

Publications

- **Wang, P.**, Abner, E.L., Fardo, D.W., Schmitt,F.A., Jicha,G.A., Eldik, L.J.V and Kryscio, R.J. (2021) Reduced Rank Multinomial Logistic Regression in Markov Chains with Application to Cognitive Data. *Statistics in Medicine.* 1-15.
- **Wang, P.**, Yin, X, Yuan, Q and Kryscio, R.J. (2021) Feature Filter for Estimating Central Mean Subspace and Its Sparse Solution. *Computational Statistics and Data Analysis*, 107285.

- Su, X., Wonkye, Y., **Wang, P.** and Yin, X. (2019) Weighted Orthogonal Components Regression Analysis. *Journal of Data Science* **17(4)**, 674-695.

- Elliott, C., Lambert, J., Stromberg, A., **Wang, P.**, Zeng, T. and Thompson, K. (2020) Feasibility as a Mechanism for Model Identification and Validation. *Journal of Applied Statistics*, 1-20.

Honors & Awards

- **R.L. Anderson Outstanding Teaching Award** 2021
  Dr. Bing Zhang Department of Statistics, University of Kentucky

- **R.L. Anderson Outstanding Research Award** 2021
  Dr. Bing Zhang Department of Statistics, University of Kentucky