2019

TRANSFORMS IN SUFFICIENT DIMENSION REDUCTION AND THEIR APPLICATIONS IN HIGH DIMENSIONAL DATA

Jiaying Weng
University of Kentucky, 338gaga@gmail.com
Author ORCID Identifier:
https://orcid.org/0000-0002-9463-5714
Digital Object Identifier: https://doi.org/10.13023/etd.2019.231

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.

Jiaying Weng, Student

Dr. Xiangrong Yin, Major Professor

Dr. Constance L. Wood, Director of Graduate Studies

TRANSFORMS IN SUFFICIENT DIMENSION REDUCTION AND THEIR APPLICATIONS IN HIGH DIMENSIONAL DATA

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
Jiaying Weng
Lexington, Kentucky

Director: Dr. Xiangrong Yin, Professor of Statistics
Lexington, Kentucky
2019

ABSTRACT OF DISSERTATION

TRANSFORMS IN SUFFICIENT DIMENSION REDUCTION AND THEIR APPLICATIONS IN HIGH DIMENSIONAL DATA

The big data era poses great challenges as well as opportunities for researchers to develop efficient statistical approaches to analyze massive data. Sufficient dimension reduction is such an important tool in modern data analysis and has received extensive attention in both academia and industry.

In this dissertation, we introduce inverse regression estimators using Fourier transforms, which is superior to the existing SDR methods in two folds, (1) it avoids the slicing of the response variable, (2) it can be readily extended to solve the high dimensional data problem. For the ultra-high dimensional problem, we investigate both eigenvalue decomposition and minimum discrepancy approaches to achieve optimal solutions and also develop a novel and efficient optimization algorithm to obtain the sparse estimates. We derive asymptotic properties of the proposed estimators and demonstrate its efficiency gains compared to the traditional estimators. The oracle properties of the sparse estimates are derived. Simulation studies and real data examples are used to illustrate the effectiveness of the proposed methods.

Wavelet transform is another tool that effectively detects information from time-localization of high frequency. Parallel to our proposed Fourier transform methods, we also develop a wavelet transform version approach and derive the asymptotic properties of the resulting estimators.

KEYWORDS: Central subspace; Fourier transform; Predictors hypothesis tests; Sufficient dimension reduction; Sufficient variable selection; Wavelet transform.

Author's signature: Jiaying Weng

Date: May 27, 2019

TRANSFORMS IN SUFFICIENT DIMENSION REDUCTION AND THEIR APPLICATIONS IN HIGH DIMENSIONAL DATA

By

Jiaying Weng

Director of Dissertation: Xiangrong Yin

Director of Graduate Studies: Constance L. Wood

Date: May 27, 2019

ACKNOWLEDGMENTS

I want to express my sincerest appreciation to my PhD advisor, Dr. Xiangrong Yin. I have been very lucky to be one of his students. His enthusiasm and diligence for research encourage me for my future career and motivate me to move further. His persistent guidance and consistent support encouraged me to overcome difficulties during my PhD study and job searching.

I greatly enjoy the help of all my dissertation committee members: Dr. Arnold Stromberg, Dr. Derek Young, Dr. Solomon Harrar, Dr. David Fardo, and the outside examiner Dr. David Royster. Every one of them has contributed lots of insights, advice, and assistance for my research.

I am grateful to the Department of Statistics for providing precious teaching opportunities and consistent financial support during my 5 years at the University of Kentucky. I would like to appreciate Dr. Arnold Stromberg and Dr. Kristen McQuerry, the valuable experience at the Applied Statistics Lab will have a notable influence on my future career. I am also thankful to all the faculty, staff, peers and friends, who are always ready to help and are pleasant to work with.

I would like to thank my family members who have provided me through moral and emotional support in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

1.1 Introduction

In the era of big data, big volume and high dimensional data are often collected less expensively through electronic device and internet. Storing and accessing massive data are conceivable for local computer with the help of cloud computing technique. Plenty of statistical learning methods and efficient optimization algorithms have been developed to reduce structure dimension, re-sample massive data, build feasible models, interpret statistical significance, and visualize data in different ways. Statisticians have developed various novel methodologies and algorithms to handle big data to achieve desirable goals, which include obtaining more accurate estimates, easier interpretation, less computational cost, and greater efficiency. Especially, many researchers studied parametric, semi-parametric, and non-parametric models in order to capture the unknown relationship between predictors and responses. In this dissertation, we employ two transformation approaches: Fourier and wavelet transform, to conduct sufficient dimension reduction and sufficient variable selection for high dimensional data.

1.2 Sufficient dimension reduction

Sufficient dimension reduction (SDR) is a statistical method to reduce dimension with the concept of sufficiency. SDR (Li, 1991; Cook, 1996) aims to find a few linear combinations of predictors so that using such linear combinations will preserve the regression information. In the regression problem, suppose that $Y \in \mathbb{R}$ is the response variable and $\mathbf{X} \in \mathbb{R}^p$ is the predictor vector. If $F(Y|\mathbf{X}) = F(Y|\boldsymbol{\eta}^T\mathbf{X})$, where F is a density function of Y and $\boldsymbol{\eta} \in \mathbb{R}^{p\times d}, d \leq p$, then the subspace spanned by the columns of $\boldsymbol{\eta}$ is called a dimension reduction subspace (DRS). We are interested in the central subspace (CS), $\mathcal{S}_{Y|\mathbf{X}}$, which is defined as the intersection of all DRSs if the intersection itself is still a DRS. Under mild conditions (Cook, 1996; Yin et al.,

2008), CS has been shown to exist and is unique. Therefore, we assume the existence of CS in this dissertation. Let $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$ be the dimension of CS, and $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ be a basis of CS. Then $Y \perp\!\!\!\perp \mathbf{X}$ given $\boldsymbol{\beta}^T\mathbf{X}$, where $\perp\!\!\!\perp$ indicates independence. In other work, the conditional distribution of Y given $\mathbf{X}$ is the same as the conditional distribution of Y given $\boldsymbol{\beta}^T\mathbf{X}$. Along with this idea of CS, some specific subspaces has been developed based on regression mean, variance and quantile (Cook and Li, 2002; Yin and Cook, 2002; Zhu and Zhu, 2009; Luo et al., 2014). Many SDR methods have been developed over the past 30 years. Sliced inverse regression (SIR) (Li, 1991) and sliced average variance estimation (SAVE) (Cook and Weisberg, 1991) are the most well-known methods.

1.3 Sufficient variable selection

Variable selection aims to identify important predictors in a large pool of variables. There are extensive studies on variable selection, such as non-negative garrotte (Breiman, 1995), lasso (Tibshirani, 1996), lars (Efron et al., 2004), elastic net (Zou and Hastie, 2005), and SCAD (Fan and Li, 2001). Specifically, Yuan and Lin (2006) proposed group lasso. The sparse-group lasso was proposed for group-wise and within group sparsity (Simon et al., 2013). Sufficient variable selection (SVS)(Yin and Hilafu, 2015), different from variable selection, is to find a subset of relevant variables but without losing any regression information, this is also why it is called *sufficient* variable selection. In the field of sufficient dimension reduction, SVS is devoted to seek a few sparse linear combination to perform dimension reduction, these approaches includes shrinkage SIR (Ni et al., 2005), sparse SIR (Li and Nachtsheim, 2006), sparse SDR (Li, 2007), coordinate-independent sparse estimation (CISE) (Chen et al., 2010), and sequential SDR (Yin and Hilafu, 2015).

1.4 Fourier and Wavelet transform

Fourier transform approach has been introduced by Zhu and Zeng (2006) and Zhu et al. (2010c) to the study of SDR. In Chapter 2, we provide further developments

for Fourier transform (FT) in inverse regression and focus on multivariate responses. Differing from the forward motivation of Zhu et al. (2010c), our approach gives more detailed illustration of their inverse regression link and significantly illustrate the idea.

In Chapter 3, motivated by the work of Cook and Ni (2005), we employ the Fourier transform approach into quadratic discrepancy function. Several versions of different discrepancy are used to achieve robust estimation. In order to achieve SVS, we propose to add coordinate-independent penalty to the quadratic discrepancy functions. A novel coordinate descent algorithm and Stiefel manifold optimization are used to obtain the sparse estimates. We also conduct the conditional and marginal hypothesis tests for identifying the structural dimension and significance of predictors.

In Chapter 4, wavelet transform is employed to estimate the central subspace in sufficient dimension reduction. In parallel, we also develop a minimum discrepancy approach based on wavelet transform. The asymptotic normality property of the estimator is established as well. For determining the structural dimension, we applied a consistent order-determination procedure by Luo and Li (2016). At last, sufficient variable selection is investigated under the framework of minimizing the discrepancy function with penalty. The coordinate descent algorithm and Stiefel manifold optimization is used to minimize the objective function.

Chapter 2 Fourier transform approach for inverse dimension reduction method

2.1 Introduction

[1]For SIR or SAVE, the number of slices has to be chosen, and the choice of this number could be problematic. Hsing and Carroll (1992) derived asymptotic properties for a special case where each slice had only two observations, which was generalized by Zhu and Ng (1995). The result of Zhu and Ng (1995) can be interpreted as the number of observations per slice must be large enough to yield efficient estimates, but still relatively small when compared with the sample size. This suggests that slicing schemes with too many slices that have too few observations per slice should be avoided. However, empirically it is hard to establish a useful rule for selecting the number of slices. To avoid such difficulties, Zhu et al. (2010b) developed the cumulative mean estimation, which uses a weighted average of SIR kernel matrices from all possible slicing schemes with two slices. Furthermore, Cook and Zhang (2014) proposed fused estimators by cumulating different number of slices: fused inverse regression estimator (FIRE) and degenerated inverse regression estimator (DIRE). Another improvement for SIR is to use Fourier transform (Zhu et al., 2010c). Fourier transform was first introduced by Zhu and Zeng (2006) in SDR to recover the dimensions in central mean subspace and CS.

The concept of SDR for multivariate response $\mathbf{Y} \in \mathbb{R}^q$ is simply to replace univatiate Y by $\mathbf{Y}$. The majority of SDR methods focuses on the univariate response, however, many methods have been developed for multivariate regression as well. [For instance, slicing the multi-dimensional $\mathbf{Y}$ into hypercubes similar to intervals in one-dimension, k-nearest neighborhood mean approaches (Aragon, 1997; Hsing, 1999; Setodji and Cook, 2004), and approaches combining all the marginal SDR for each

[1]This is an original manuscript of an article published by Taylor & Francis in Journal of Nonparametric Statistics on August 20, 2018, available online:
https://www.tandfonline.com/doi/abs/10.1080/10485252.2018.1515432?journalCode=gnst20.

component of $\mathbf{Y}$ to estimate the multivariate CS (Cook and Setodji, 2003; Saracco, 2005; Yin and Bura, 2006).] Li et al. (2008) proposed a projective resampling (PR) method to avoid multivariate slicing while effectively estimating the CS. When data have categorical variables, but SDR is only on continuous predictors, then such an SDR approach leads to partial SDR (Chiaromonte et al., 2002) and (Li et al., 2003). SDR is quite useful for reducing predictors and helping to build a better model. However, it is still difficult to interpret the predictors in the model as the linear combination consists of all the original variables. To this end, SDR with penalization can help to select important variables, leading to sufficient variable selection (SVS). One of the approaches is a general procedure by Li (2007), which developed a sparse SDR estimator for a general dimension reduction kernel matrix by transforming the eigenvalue-decomposition approach to a regression-type optimization problem. Then a penalty term (such as a L^1 penalty) is added to shrink the number of parameters. Recently, Yin and Hilafu (2015) developed a sequential SDR and SVS procedure to deal with the large p, small n data with two effective algorithms, combining the techniques of SDR methods for the univariate response, multivariate responses, partial SDR and penalization.

In this chapter, we provide further developments for Fourier transform (FT) in inverse regression and focus on multivariate response. Differing from the forward motivation of Zhu et al. (2010c), our approach gives more detailed illustration of their inverse regression link and significantly develops the idea. We have the following main contributions: We provide a result regarding the choice of the number of FTs, only a finite number, much less than the sample size as suggested by Zhu et al. (2010c), which is sufficient enough. Indeed, empirically, 50 FTs are sufficient, and the results are quite stable. This will not only save computational time, but also ensure the accuracy of the estimate. We obtain the asymptotic tests for determining dimensions for FT. We develop a partial SDR for FT and obtain the respective asymptotic tests for estimating dimensions. We further propose SVS in two useful cases: For $n > p$, we use the idea of Li (2007) to develop a sparse SDR version of FT, which produces sparse and more accurate estimate. Using the sequential SDR and SVS of Yin and

Hilafu (2015), we develop a procedure of FT to deal with large p, small n data.

The chapter is organized as follows: Section 2.2 provides the theoretical reasons for the FT estimate, comparison between SIR and FT in population and sample sense, properties of choice of the number of Fourier transforms, and algorithms for estimating CS, along with the test for dimension. Section 2.3 develops a method for estimating the partial SDR using the FT approach. Section 2.4 proposes the sufficient variable selection for two situations: large n, small p and large p, small n data. Section 2.5 presents simulation studies and a real data analysis. Section 2.6 summarizes our conclusion. All proofs are included in the Appendix.

2.2 Methodology

Estimation Method

This section introduces FT estimator. To facilitate our discussion, we use standardized predictor $\mathbf{Z}$ of $\mathbf{X}$, due to the equivalence of the CS of $\mathbf{Y}|\mathbf{X}$ and the CS of $\mathbf{Y}|\mathbf{Z}$ (Cook, 1998). Let $\mathbf{Z} = \Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$, where $\boldsymbol{\mu}$ and Σ are the mean and covariance matrix of $\mathbf{X}$. Under the well-known linearity condition, $m(\mathbf{y}) = \mathrm{E}(\mathbf{Z}|\mathbf{Y} = \mathbf{y}) \in \mathcal{S}_{\mathbf{Y}|\mathbf{Z}}$ (Cook, 1998). Thus, estimating the space spanned by $m(\mathbf{y})$ ($\mathcal{S}_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})}$) is to recover part of the CS . Let $f_{\mathbf{Y}}(\mathbf{y})$ be the marginal density distribution of $\mathbf{Y}$. Then, FT of the density-weighted conditional mean $m(\mathbf{y})f_{\mathbf{Y}}(\mathbf{y})$ is $\psi(\boldsymbol{\omega}) = \int e^{i\boldsymbol{\omega}^T\mathbf{y}} m(\mathbf{y})f_{\mathbf{Y}}(\mathbf{y})d\mathbf{y} = a(\boldsymbol{\omega}) + ib(\boldsymbol{\omega}), \boldsymbol{\omega} \in \mathbb{R}^q$, where $a(\boldsymbol{\omega})$, $b(\boldsymbol{\omega})$ are the real, imaginary part of $\psi(\boldsymbol{\omega})$, respectively.

We claim that $\psi(\boldsymbol{\omega}) = \mathrm{E}(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\mathbf{Z})$ and $\mathcal{S}_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})} = \mathrm{Span}\{\psi(\boldsymbol{\omega}), \boldsymbol{\omega} \in \mathbb{R}^q\}$. The first assertion is due to

$$\begin{aligned}
\psi(\boldsymbol{\omega}) &= \int e^{i\boldsymbol{\omega}^T\mathbf{y}} m(\mathbf{y})f_{\mathbf{Y}}(\mathbf{y})d\mathbf{y} &&= \int e^{i\boldsymbol{\omega}^T\mathbf{y}} \mathrm{E}(\mathbf{Z}|\mathbf{y})f_{\mathbf{Y}}(\mathbf{y})d\mathbf{y} \\
&= \int \mathrm{E}(e^{i\boldsymbol{\omega}^T\mathbf{y}}\mathbf{Z}|\mathbf{y})f_{\mathbf{Y}}(\mathbf{y})d\mathbf{y} &&= \mathrm{E}[\mathrm{E}(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\mathbf{Z}|\mathbf{Y})] \\
&= \mathrm{E}(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\mathbf{Z}).
\end{aligned}$$

Note that $\mathcal{S}_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})} = \mathrm{Span}\{m(\mathbf{y}), \mathbf{y} \in \mathrm{supp}(f_{\mathbf{Y}})\} = \mathrm{Span}\{m(\mathbf{y})f_{\mathbf{Y}}(\mathbf{y}), \mathbf{y} \in \mathrm{supp}(f_{\mathbf{Y}})\} \subseteq \mathcal{S}_{\mathbf{Y}|\mathbf{Z}}$, under the linearity condition. By its inverse transform of $\psi(\boldsymbol{\omega})$, then $m(\mathbf{y})f_{\mathbf{Y}}(\mathbf{y}) = (2\pi)^{-1}\int e^{-i\boldsymbol{\omega}^T\mathbf{y}}\psi(\boldsymbol{\omega})d\boldsymbol{\omega}$. Thus, $\mathcal{S}_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})} = \mathrm{Span}\{\psi(\boldsymbol{\omega}), \boldsymbol{\omega} \in \mathbb{R}^q\} = \mathrm{Span}\{a(\boldsymbol{\omega}), b(\boldsymbol{\omega}), \boldsymbol{\omega} \in \mathbb{R}^q\}$, so the second assertion holds. Note that above derivation differs from the for-

ward illustration of Zhu and Zeng (2006), but does agree with their comment on inverse regression approach (right above Proposition 1, p 1295, Zhu et al., 2010). Although we give more details, both lead to the same estimator.

FT estimates the CS just as SIR does, but they might be different in estimation. In the population sense, SIR and FT estimate the space spanned by $\mathrm{E}(\mathbf{Z}|\mathbf{Y} = \mathbf{y})$, regardless of continuous or categorical $\mathbf{Y}$. That is,

$$\mathrm{Span}\{\psi(\boldsymbol{\omega}), \boldsymbol{\omega} \in \mathbb{R}^q\} = \mathrm{Span}\{\mathrm{E}(\mathbf{Z}|\mathbf{Y} = \mathbf{y}), \mathbf{y} \in \mathbb{R}^q\}.$$

When $\mathbf{Y}$ is a categorical variable, in the sample sense, SIR and FT are also equivalent (See the appendix). That is, for categorical $\mathbf{Y}$, the left-hand side of the above equation does not gain any useful information by changing the number of $\boldsymbol{\omega}$, comparing with the right-hand side of the above equation. However, for continuous response $\mathbf{Y}$, empirical estimates for these two methods are different in accuracy, mainly due to the limited sample size. Note that the left-hand side (using FT) needs to choose the number of $\boldsymbol{\omega}$, while the right-hand side (using slices) needs to select the number of slices. The right-hand side has uncertainty for selecting the number of slices. Theoretically, it should choose a large number of slices due to its conditional mean, but practically it should use a small number of slices due to limited sample size. It is also well-known that the number of slices will greatly affect the accuracy of estimates. However, it seems that FT is quite stable for choosing the number of $\boldsymbol{\omega}$, as long as it is large enough.

Property of Covering and Choice of $\boldsymbol{\omega}$

In the previous discussion, we achieved the estimate given a sequence of $\boldsymbol{\omega}$. Hence, we need information about the number and the value of $\boldsymbol{\omega}$. Note that $\boldsymbol{\omega} \in \mathbb{R}^q$, but practically we cannot take the entire $\mathbb{R}^q$. Proposition 1 below, however, indicates that a finite number of $\boldsymbol{\omega} \in \mathbb{R}^q$ will be enough to recover the entire $\mathcal{S}_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})}$. Yin and Li (2011) used a general dense class of functions of $\mathbf{Y}$ to estimate CS. FT is one of such dense classes, so the proof of Proposition 1 is similar to that of Theorem 2.2

(Yin and Li, 2011). Hence, we omit its proof. Because we discuss the partial SDR, we don't need the linearity condition for the Proposition 1.

Proposition 1. *1. There exists a finite sequence of* $\boldsymbol{\omega}_j \in \mathbb{R}^q, j = 1, \ldots, t$, *such that* $\mathcal{S}_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})} = \mathrm{Span}\{a(\boldsymbol{\omega}_1), b(\boldsymbol{\omega}_1), \ldots, a(\boldsymbol{\omega}_t), b(\boldsymbol{\omega}_t)\}$.

2. Consider a random sequence $\boldsymbol{\omega}_j$, $j = 1, 2, \ldots$, *there exist an integer* t_0 *such that for all* $t \geq t_0$, $\mathcal{S}_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})} = \mathrm{Span}\{a(\boldsymbol{\omega}_1), b(\boldsymbol{\omega}_1), \ldots, a(\boldsymbol{\omega}_t), b(\boldsymbol{\omega}_t)\}$.

Part 1 of Proposition 1 indicates that the finite number of $\boldsymbol{\omega}$ is enough to recover $\mathcal{S}_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})}$ and one could choose as small as half of the dimension of $\mathcal{S}_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})}$. But typically, we do not know the dimension of $\mathcal{S}_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})}$. Part 2 of Proposition 1 indicates that if the number of selected $\boldsymbol{\omega}$ is large enough, we can then recover $\mathcal{S}_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})}$. This again in practice does not provide a useful rule. However, our simulations later show that when the number of $\boldsymbol{\omega}$ is large enough, the results are quite stable. Indeed, we find that 50 $\boldsymbol{\omega}$s is enough for capturing the structure of CS, as well as testing the dimension.

Another related issue is how to select $\boldsymbol{\omega}$. Zhu et al. (2010c) provide an argument to choose $\boldsymbol{\omega}$. For a multivariate $\mathbf{Y}$, we choose a small s, say $s = 0.1$, with $P(|\boldsymbol{\omega}^T\mathbf{Y}| > \pi) \leq s$, then randomly generate $\boldsymbol{\omega} \sim N(\mathbf{0}, \frac{s\pi^2}{\mathrm{E}(\mathbf{Y}^T\mathbf{Y})}I)$. Our limited simulations indicate that such a method performed very stable.

Algorithm

In this section, we summarize what we have discussed above and show the algorithm for the estimate using sample. Let $\Psi = (a(\boldsymbol{\omega}_1), b(\boldsymbol{\omega}_1), \ldots, a(\boldsymbol{\omega}_t), b(\boldsymbol{\omega}_t))$, for some $t > 0$, and $V = \Psi\Psi^T$ as the population kernel matrix. Recall, $\psi(\boldsymbol{\omega}) = a(\boldsymbol{\omega}) + ib(\boldsymbol{\omega})$. Let $(\mathbf{y}_i, \mathbf{x}_i)$, $i = 1, \ldots, n$ be a random sample, and assume that the dimension of $\mathcal{S}_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})}$ is known as d. The algorithm of FT, similar to that of Zhu et al. (2010c), is the following:

1. Standardize $\mathbf{x}_i$: $\hat{\mathbf{z}}_i = \hat{\Sigma}_{\mathbf{X}}^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}}), i = 1, \ldots, n$, where $\bar{\mathbf{x}}$ is the sample mean, and $\hat{\Sigma}_{\mathbf{X}}$ is the sample covariance matrix of $\mathbf{X}$.

2. Choose $\{\boldsymbol{\omega}_j\}_{j=1}^t$ as in Section 2.2 and for each $\boldsymbol{\omega}_j$, calculate sample version of $\psi(\boldsymbol{\omega}_j)$: $\hat{\psi}(\boldsymbol{\omega}_j) = \frac{1}{n}\sum_{k=1}^n e^{i\boldsymbol{\omega}_j^T \mathbf{y}_k}\hat{\mathbf{z}}_k$, and $\hat{a}(\boldsymbol{\omega}_j) = \text{Real}(\hat{\psi}(\boldsymbol{\omega}_j))$ and $\hat{b}(\boldsymbol{\omega}_j) = \text{Image}(\hat{\psi}(\boldsymbol{\omega}_j))$.

3. Form $\hat{\Psi}$ and $\hat{V}$ as $\hat{\Psi} = \{\hat{a}(\boldsymbol{\omega}_j), \hat{b}(\boldsymbol{\omega}_j)\}_{j=1}^t$, $\hat{V} = \hat{\Psi}\hat{\Psi}^T$, where $\hat{\Psi}$ is a $p \times 2t$ matrix and $\hat{V}$ is a $p \times p$ sample kernel matrix.

4. The first d eigenvectors $(\hat{\eta}_i, i = 1, \ldots, d)$ of $\hat{V}$ corresponding to the first d largest eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_d$ are the estimated directions of $\mathcal{S}_{\text{E}(\mathbf{Z}|\mathbf{Y})}$. Transform back to the $\mathbf{X}$ scale, $\hat{\beta}_i = \hat{\Sigma}_{\mathbf{X}}^{-\frac{1}{2}}\hat{\eta}_i, i = 1, \ldots, d$.

Testing methods for dimension

Previously, we obtained the estimate by assuming the dimensions of CS with the inference required for real data. Hence, we develop the test statistics and associated asymptotic distribution in this section. We construct the statistic

$$\hat{\Lambda}_m = n \sum_{j=m+1}^{p} \hat{\lambda}_j$$

to test the hypothesis of the form $d = m$ versus $d > m$. The value of m begins with 0, so we test $d = m$ by comparing sample $\hat{\Lambda}_m$ with the quantile of the asymptotic distribution of $\hat{\Lambda}_m$ under the null hypothesis $d = m$. If we fail to reject, then $d = m$, otherwise we increase m by 1 and continue the same process until we fail to reject. The asymptotic distribution of $\hat{\Lambda}_d$ is stated the below Proposition 2, of which proof is in the appendix. As the Proposition 2 is stated in terms of the partial SDR, we do not need the linearity condition.

Proposition 2. *Let $d = \dim[\mathcal{S}_{\text{E}(\mathbf{Z}|\mathbf{Y})}]$ and assume that $2t > d + 1$ and $p > d$. Then the asymptotic distribution of $\hat{\Lambda}_d$ is the same as the distribution of*

$$C = \sum_{i=1}^{(p-d)(2t-d)} \lambda_i C_i,$$

where the C_is are independent chi-square random variables each with one degree of freedom and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{(p-d)(2t-d)}$ are eigenvalues of the covariance matrix Ω, where Ω is defined in the Appendix.

One can directly obtain the distribution of the weighted Chi-square Statistic C, however, simplification is possible. Following Bentler and Xie (2000), we consider two types of simplified test statistics.

Scaled Statistic: $\overline{T}_m = [\text{trace}(\hat{\Omega}_n)/p^*]^{-1} n \sum_{j=m+1}^{p} \hat{\lambda}_j \sim \chi^2_{p^*}$, where $\hat{\Omega}_n$ is a consistent estimator of Ω and $p^* = (p-m)(2t-m)$.

Adjusted Statistic: $\widetilde{T}_m = [\text{trace}(\hat{\Omega}_n)/d^*]^{-1} n \sum_{j=m+1}^{p} \hat{\lambda}_j \sim \chi^2_{d^*}$, where $d^* = \frac{[\text{trace}(\hat{\Omega}_n)]^2}{\text{trace}(\hat{\Omega}_n^2)}$.

Sparse Eigen-Decomposition estimation (SED) (Zhu et al., 2010a) is another method to estimate d. We sketch SED here. Let $\hat{V}$ be the sample kernel matrix in Section 2.2 Algorithm. The SED procedure is the following:

$$(\hat{\boldsymbol{\lambda}}, \hat{\alpha}, \hat{\beta}) = \underset{\lambda,\alpha,\beta}{\arg\min}\, n||\hat{V} - \sum_{i=1}^{p} \lambda_i \boldsymbol{\alpha}_i \boldsymbol{\beta}_i^T||^2 + l_n \sum_{i=1}^{p} \hat{w}_i |\lambda_i|,$$

subject to $\beta^T \beta = I_p$, where $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_p)^T$ be a $p \times 1$ vector, $\alpha = (\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_p)$ and $\beta = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_p)$ be $p \times p$ matrices, $\hat{\boldsymbol{w}} = (\hat{w}_1, ...\hat{w}_p)^T$ be a known weight vector. The tuning parameter, l_n, is select by typical AIC and BIC as suggested by Zhu et al. (2010a). The number of dimensions is equal to the number of nonzero values of $\hat{\boldsymbol{\lambda}}$.

2.3 Partial Central Subspace

When predictors consist of both continuous and categorical variables, we focus on the partial SDR (Chiaromonte et al., 2002) which is only on continuous variables. In this section, we extend FT to partial SDR. Without loss of generality, let W be the categorical variable with K levels. Chiaromonte et al. (2002) defined the partial CS to be the intersection of all subspaces spanned by $\boldsymbol{\eta} \in \mathbb{R}^{p \times d}$ such that $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | (\boldsymbol{\eta}^T \mathbf{X}, W)$, if the intersection itself also satisfies such a condition. Let $\mathcal{S}^W_{\mathbf{Y}|\mathbf{X}}$ be the partial CS, then $\mathcal{S}^W_{\mathbf{Y}|\mathbf{X}} = \bigoplus_{k=1}^{K} \mathcal{S}_{\mathbf{Y}_k|\mathbf{X}_k}$, where $\mathcal{S}_{\mathbf{Y}_k|\mathbf{X}_k}$ is the CS conditioning on level k.

Suppose that for each group, the mean and covariance matrix of $\mathbf{X}_k$ are $\boldsymbol{\mu}_k$ and Σ_k. To facilitate the discussion, we further assume that the covariance structures are the same across each level, that is, $\Sigma_k = \Sigma_{pool},\ k = 1, \ldots, K$. Let $\mathbf{Z}_k = \Sigma_{pool}^{-1/2}(\mathbf{X}_k - \boldsymbol{\mu}_k)$, then $\mathcal{S}^W_{\mathbf{Y}|\mathbf{X}} = \Sigma_{pool}^{-1/2} \bigoplus_{k=1}^{K} \mathcal{S}_{\mathbf{Y}_k|\mathbf{Z}_k}$. For each level, we construct the kernel matrix V_k and combine them into an overall kernel matrix: the partial kernel matrix $V^W =$

$\sum_{k=1}^{K} P(W=k)V_k$. Suppose that the linearity and coverage conditions for each level hold:

- Linearity: $\mathrm{E}(\mathbf{Z}_k|P_{\mathcal{S}_{\mathbf{Y}_k|\mathbf{Z}_k}}\mathbf{Z}_k) = P_{\mathcal{S}_{\mathbf{Y}_k|\mathbf{Z}_k}}\mathbf{Z}_k$, for $k = 1, \ldots, K$.
- Coverage: $\mathrm{Span}(V^W) = \bigoplus_{k=1}^{K} \mathrm{Span}(V_k) = \bigoplus_{k=1}^{K} \mathcal{S}_{\mathbf{Y}_k|\mathbf{Z}_k}$.

Assume that the dimension of $\mathcal{S}^W_{\mathbf{Y}|\mathbf{X}}$, d, is known, we have the following algorithm for estimating $\mathcal{S}^W_{\mathbf{Y}|\mathbf{X}}$. The algorithm is similar to Chiaromonte et al. (2002) except applying our new partial kernel matrix. The estimate from the following steps is referred as the partial Fourier transform (PFT).

1. For each level k, $\bar{\boldsymbol{x}}_k$ and $\hat{\Sigma}_k$ are the sample mean and covariance matrix of $\mathbf{X}_k$, the common covariance matrix is $\hat{\Sigma}_{pool} = \sum_{k=1}^{K} \frac{n_k}{n}\hat{\Sigma}_k$ and $\hat{\mathbf{z}}_{ik} = \hat{\Sigma}_{pool}^{-1/2}(\mathbf{x}_{ik} - \bar{\mathbf{x}}_k)$, $i = 1, \ldots, n_k$ and $k = 1, \ldots, K$.

2. Apply the algorithm in Section 2.2 to obtain the sample kernel matrix for each level k: $\hat{V}_k$, and then $\hat{V}^W = \sum_{k=1}^{K} \frac{n_k}{n}\hat{V}_k$.

3. The first d eigenvectors $(\hat{\eta}_i, i = 1, \ldots, d)$ of $\hat{V}^W$ corresponding to the first d largest eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_d$ are the estimates. Transform back to the $\mathbf{X}$ scale, $\hat{\beta}_i = \hat{\Sigma}_{pool}^{-\frac{1}{2}}\hat{\eta}_i, i = 1, \ldots, d$.

To estimate d of PFT, we construct a test statistic

$$\hat{\Lambda}^W_m = n \sum_{j=m+1}^{p} \hat{\lambda}_j.$$

Proposition 3. *Under the linearity and coverage conditions for partial SDR, let* $d = \dim[\mathcal{S}^W_{\mathbf{Y}|\mathbf{X}}]$ *and assume that* $2\sum t_k > Kd + 1$ *and* $p > d$. *Then the asymptotic distribution of* $\hat{\Lambda}^W_d$ *is the same as the distribution of*

$$C = \sum_{i=1}^{(p-d)(2\sum t_k - Kd)} \lambda_i C_i$$

where the C_i*s are independent chi-square random variables each with one degree of freedom and* $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{(p-d)(2\sum t_k - Kd)}$ *are eigenvalues of the covariance matrix* Ω^W*, where* Ω^W *is defined in the appendix.*

2.4 Sufficient Variable Selection

In some cases, not all predictors contribute for the estimate so SDR with penalization is helpful to choose significant variables. Variable selection is an essential step especially for a sparse model. In this section, we extend FT for sufficient variable selection via the penalized approach. We consider two different cases: the traditional large n, small p data, and the modern large p, small n data.

Large n, Small p: We adopt a general sparse SDR via penalty approach developed by Li (2007): $\tilde{V}\tilde{\eta}_i = \rho_i \Sigma \tilde{\eta}_i$, for $i = 1, ..., p$, where $\tilde{V} = \Sigma^{1/2} V \Sigma^{1/2}$ is a symmetric kernel SDR matrix; Σ is the covariance matrix; vector $\tilde{\eta}_1, \cdots, \tilde{\eta}_p$ are eigenvectors satisfying $\tilde{\eta}_i^T \Sigma \tilde{\eta}_j = 1$ if $i = j$, and 0 if $i \neq j$; and $\rho_1 \geq \cdots \geq \rho_p \geq 0$ are corresponding eigenvalues. Then the eigenvalue-decomposition approach via penalty term becomes an optimization problem for sparse SDR as follows:

$$\min_{\alpha,\beta}(\sum_{i=1}^{p} ||\Sigma^{-1} v_i - \alpha\beta^T v_i||_{\Sigma}^2 + \lambda_2 \text{trace}(\beta^T \Sigma \beta) + \sum_{j=1}^{d} \lambda_{1j} |\beta_j|_1),$$

subject to $\alpha^T \Sigma \alpha = I_d$, where $\lambda_{1j} \geq 0$,$j = 1, ..., d$ are the tuning parameters, and v_i, $i = 1, ..., p$ are the columns of $\tilde{V}^{1/2}$.

The algorithm of Li (2007) can be summarized as below:

1. Initialize α and β using the sample kernel matrix in Section 2.2.

2. Given α, update β as below:

$$\hat{\beta}_{\alpha j} = \arg\min_{\beta_j}(||y^* - x^* \beta_j||^2 + \lambda_{1j} |\beta_j|_1),$$

where $x^* = \begin{bmatrix} \tilde{V}^{1/2} \\ \sqrt{\lambda_2}\Sigma^{1/2} \end{bmatrix}_{2p \times p}$, $y^* = \begin{bmatrix} \tilde{V}^{1/2}\alpha_j \\ 0 \end{bmatrix}_{2p \times 1}$.

3. Given β, let U_α, D_α, and V_α denote the matrices from the singular value decomposition of the matrix $\Sigma^{-1/2}\tilde{V}\beta$, then $\hat{\alpha} = \Sigma^{-1/2} U_\alpha V_\alpha^T$.

4. Continue steps 2 and 3 until β converges.

Typically, we need to fix λ_{1j} and λ_2 in the above algorithm. The final selection of tuning parameters of λ_{1j} and λ_2 can be determined by AIC and BIC (Li, 2007). For our purpose, we simply use FT kernel matrix to replace $\tilde{V}$, and denote such a procedure as S-FT.

Large p, Small n: Yin and Hilafu (2015) proposed a sequential SDR (SSDR) for such a problem. We extend FT in their algorithm. Note that the algorithm of Yin and Hilafu (2015) is based on the following result.

Proposition 4. *(Yin and Hilafu, 2015) If $\mathbf{X}_1$ and $\mathbf{X}_2$ are random vectors, $B^T\mathbf{X}_1$ is a linear combination of $\mathbf{X}_1$, where B is a matrix, then either (a) or (b) implies (c) below:*

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

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

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

Statement (c) is very important, if it is true, then $p(\mathbf{Y}|\mathbf{X}_1, \mathbf{X}_2) = p(\mathbf{Y}|\mathbf{X}_1, \mathbf{X}_2, B^T\mathbf{X}_1) = p(\mathbf{Y}|\mathbf{X}_2, B^T\mathbf{X}_1)$. Thus, if the dimension of $B^T\mathbf{X}_1$ is less than $\mathbf{X}_1$, we achieved dimension reduction without loss of any information. To force statement (c), we may use statement (a) or statement (b). Write $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, and choose $\mathbf{X}_1$ with dimension $p_1 < n$. Then reduce $\mathbf{X}_1$ to $B^T\mathbf{X}_1$, and replace $\mathbf{X}$ with $(B^T\mathbf{X}_1, \mathbf{X}_2)$ as new $\mathbf{X}$. Keep doing this until there is no more reduction. To find $B^T\mathbf{X}_1$, Path I procedure (Yin and Hilafu, 2015) uses statement (a) when the response variable is continuous. This procedure needs to construct $B^T\mathbf{X}_1$ using regression $(\mathbf{X}_2, \mathbf{Y})$ on $\mathbf{X}_1$. On the other hand, when dealing with the categorical response, statement (b) is the choice which is called Path II procedure by Yin and Hilafu (2015). Path II conducts the partial SDR for regression $\mathbf{X}_2$ on $\mathbf{X}_1$ given $\mathbf{Y}$, and the usual SDR of $\mathbf{Y}$ on $\mathbf{X}_1$. Because of the categorical response, FT is equivalent to SIR, we only use Path I to construct an estimate. For clarity, we illustrate the algorithm of Path I of Yin and Hilafu for FT below.

1. Order the predictors using the distance correlation in Li et al. (2012).

2. Decompose $\mathbf{X} \in \mathbb{R}^p$ into $\mathbf{X}^T = (\mathbf{X}_1^T, \mathbf{X}_2^T)$, where $\mathbf{X}_1$ is a $p_1 \times 1$ vector such that $n > p_1$, and consider the problem of $\mathbf{X}_1 \perp (\mathbf{X}_2, Y)|\beta_1^T\mathbf{X}_1$.

3. For SDR solution, apply the method in Section 2.2 to new response $\mathbf{Y}_{new}^T = (\mathbf{X}_2^T, \mathbf{Y})$ given $\mathbf{X}_1$, and find the reduced variable $\beta_1^T\mathbf{X}_1$; For SVS solution, apply multivariate regression with penalization to the problem of $\mathbf{Y}_{new}^T|\mathbf{X}_1$, and find the reduced variable $\beta_1^T\mathbf{X_1}$.

4. Replace predictors $\mathbf{X}$ by $(\beta_1^T\mathbf{X}_1, \mathbf{X}_2)$ and go back to step 1, until there is no further reduction.

We will compare the original SSDR using SIR (SSDR-SIR) and SSDR using FT (SSDR-FT) in the simulation for Path I.

2.5 Numerical Study

Suppose that $\hat{B} = (\hat{\beta}_1, ..., \hat{\beta}_d)$ is the estimate of a $p \times d$ matrix B, and both $\hat{B}$ and B are orthogonal matrices. We use following criteria to measure the accuracy of the estimates.

1. Let ρ_i^2's be the eigenvalues of matrix $\hat{B}^T B B^T \hat{B}$ for $i = 1, \cdots, d$: the vector correlation coefficient is $r_1 = \sqrt{|\hat{B}^T B B^T \hat{B}|} = |\prod_{i=1}^d \rho_i|$ and the trace correlation is $r_2 = \sqrt{\sum_{i=1}^d \rho_i^2/d}$ (Ye and Weiss, 2003). The bigger the r_1 or r_2, the better the estimate.

2. Define $\Delta(B, \hat{B}) = ||\hat{B}\hat{B}^T - BB^T||$ (Li et al., 2005). We use two ways to calculate $||.||$: (a) $\Delta_m(A) = ||A||$ is the maximum singular value of A, and (b) $\Delta_f(A) = ||A||$ is the Frobenius norm as $\Delta_f(A) = \sqrt{\text{trace}(AA^T)}$. The smaller the $\Delta_m(A)$ or $\Delta_f(A)$, the better the estimate.

For SVS, we use true positive rate (TPR) and false positive rate (FPR): TPR is the number of correctly identified active predictors to the number of truly active predictors, and FPR is the number of falsely identified active predictors to the total

number of inactive predictors to compare different methods. Better estimates have bigger TPRs and smaller FPRs.

Simulations

In this section, we illustrate the advantages of FT with six models. Each model has a different purpose. We use Model 1 to assess if the number of $\boldsymbol{\omega}$'s in FT could affect estimate accuracy and Model 2 to compare FT with SIR, IRE (Cook and Ni, 2005), FIRE and DIRE and, further to compare S-FT with S-SIR. We use Model 3 to estimate the dimension using the Weighted Chi-square, Scaled, Adjusted Statistic and SED and Model 4 for multivariate regression. We use Model 5 to compare partial SDR using SIR (PSIR) (Chiaromonte et al., 2002) and PFT. Finally, Model 6 is used for a large p, small n problem.

Model 1. $Y = X_1 + 0.5X_2^2$, with $p = 5$, $n = 800$ and $d = 2$. Predictors $X_1, X_3, X_5 \overset{iid}{\sim} N(0,1)$, and $X_2 = X_1 + Z$ where $Z \sim N(0,1)$ and $X_4 = (1 + X_2)Z$. Let $\{e_i\}$ be $p \times 1$ vectors whose i^{th} entry is 1 and other entries are 0. Then $B = (e_1, e_2)$.

Figure 2.1 plots mean values of respective r_1, r_2, Δ_m and Δ_f over 100 simulated data vs sizes of $\boldsymbol{\omega}$: $\{5, 10, 15, ..., 100\}$. It shows that FT has high accuracy, and its estimates keep the same magnitude for the different number of $\boldsymbol{\omega}$'s. This seems consistent with the result of Proposition 1. Hence, as long as the size of $\boldsymbol{\omega}$ is large enough, estimates of the CS are accurate and stable.

Model 2. This is the first example of Cook and Zhang (2014). $Y = |\sin X_1| + 0.2\epsilon$, with $d = 1$ and $B = e_1$. Predictors $\mathbf{X}_i \sim \frac{1}{4}N_p(\boldsymbol{\mu}_1, \Sigma_1) + \frac{1}{2}N_p(\boldsymbol{\mu}_2, \Sigma_2) + \frac{1}{4}N_p(\boldsymbol{\mu}_3, \Sigma_3)$, where $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_3 = (1, 0, ..., 0)^T$, $\boldsymbol{\mu}_2 = (2, 0, ..., 0)^T$, $\Sigma_1 = \Sigma_2 = \sqrt{0.1}I_p$ and $\Sigma_3 = \sqrt{10}I_p$. Let $p = 15$, $n = 400$, and ϵ is a uniform (0,1).

We compare SIR, IRE, FIRE, DIRE, and FT for this model. The number of slices for SIR and IRE are $\{3, 4, ..., 15\}$. For FIRE and DIRE, we fuse $H = \{3, 4, ..., 15\}$, while for FT, the size of $\boldsymbol{\omega}$ is 50. Figure 2.2 plots mean values of respective r_1, r_2, Δ_m and Δ_f over 100 simulated data vs the number of slices from 3 to 15. We see that

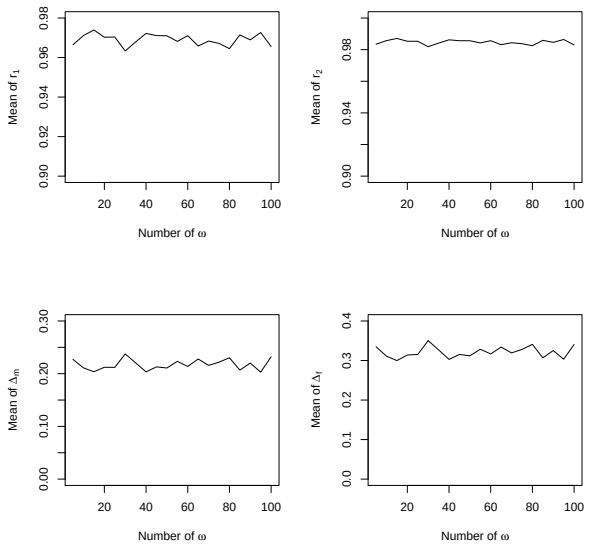

Figure 2.1: Mean values of respective r_1, r_2, Δ_m and Δ_f over 100 simulated data vs sizes of $\boldsymbol{\omega}$: $\{5, ..., 100\}$ in Model 1.

Table 2.1: Means and Standard Deviations of TPR and FPR, respectively, over 100 simulated data in Model 2.

	S-FT	S-SIR
TPR	0.8700(0.3380)	0.6300(0.4852)
FPR	0.0650(0.2418)	0.1207(0.3240)

the results of SIR and IRE change with different slices, indicating that the choice of number of slices is important. FIRE and DIRE combine different slices together, thus they are constant lines. Regardless, FT has the largest values of r_1 and r_2, and the smallest Δ_m and Δ_f compared with the other four methods, indicating that FT is the best method for this model. We also conduct SVS for S-FT and S-SIR and report the respective TPR and FPR over 100 simulated data. The number of slices for S-SIR is 5 and the number of $\boldsymbol{\omega}$ for S-FT is 50. Table 2.1 shows that S-FT has larger TPR and smaller FPR compared to these of S-SIR, thus better results for S-FT than those of S-SIR.

Model 3. This model is similar to example 4.1 of Bentler and Xie (2000). $Y = X_1 + 0.5\epsilon$, with $p = 4$, $d = 1$ and $B = e_1$. Predictor vector $\mathbf{X}_i$ follows multivariate normal

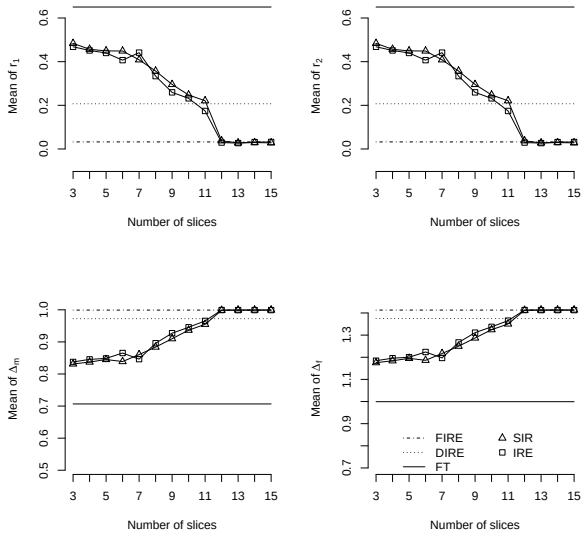

Figure 2.2: Mean values of respective r_1, r_2, Δ_m and Δ_f over 100 simulated data vs the number of slices, $3 \cdots 15$, in Model 2.

distribution with the mean $(1, 2, 3, 4)$ and equi-correlation matrix with a variance of 1 and a correlation of 0.5, and $\epsilon \sim N(0, 1)$.

We check the three asymptotic dimension tests: the Weighted Chi-square Statistic, Scaled Statistic, and Adjusted Statistic, as well as the SED method (Zhu et al., 2010a) for this model. The size of $\boldsymbol{\omega}$ is 50, and we use three sample sizes of $n = 400, 600, 800$. The percentages of correctly detected dimensions among 100 simulated data are reported in Table 2.2, which shows that the Scaled Statistic performs better than Weighted Chi-square Statistic and Adjusted Statistic. This is consistent with example 4.1 of Bentler and Xie (2000). The proportions of the correctly identified dimensions for the three asymptotic tests are 100% when the sample size is 800, resulting in more accurate estimates for larger sample sizes. Nevertheless, the Scaled test statistic is the best among all four tests. Moreover, we report TPR and FPR for S-FT and S-SIR, respectively, over 100 simulated data in Table 2.2, and the results are optimal.

Model 4. This is Example 3 of Zhu et al. (2010c). $Y_1 = 1+\beta_1^T\mathbf{X}+\sin(\beta_2^T\mathbf{X})+\epsilon_1$, $Y_2 = \frac{\beta_2^T\mathbf{X}}{0.5+(\beta_1^T\mathbf{X}+1)^2} + \epsilon_2$, $Y_3 = |\beta_1^T\mathbf{X}|\epsilon_3$, $Y_4 = \epsilon_4$, $Y_5 = \epsilon_5$, with $p = 20$, $d = 2$, and $\beta_1 = e_1$ and

Table 2.2: Percentages of correctly detected dimensions in Model 3.

Sample Size	Weighted χ^2	Scale Stat.	Adj Stat.	SED
400	0.0000	1.0000	0.0000	0.9500
600	0.1800	1.0000	0.1700	1.0000
800	1.0000	1.0000	1.0000	1.0000

Table 2.3: TPR and FPR over 100 simulated data in Model 3.

	$n = 400$		$n = 600$		$n = 800$	
	S-FT	S-SIR	S-FT	S-SIR	S-FT	S-SIR
TPR	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
FPR	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

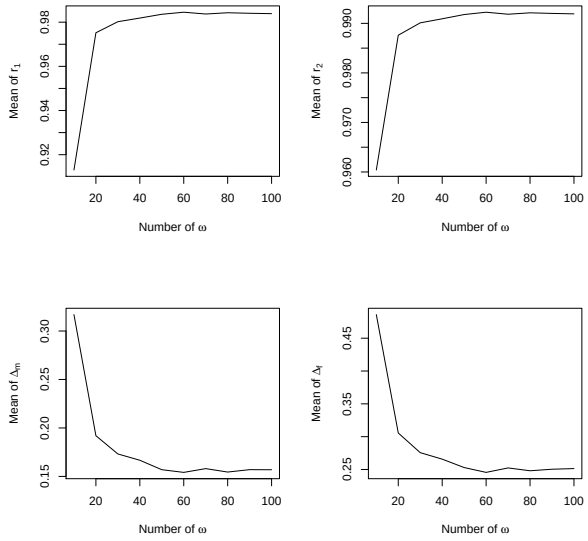

Figure 2.3: Mean values of respective r_1, r_2, Δ_m and Δ_f over 100 simulated data vs sizes of $\boldsymbol{\omega}$: $\{10, 20, ..., 100\}$ in Model 4.

$\beta_2 = e_2 + e_3$. Predictor $\mathbf{X}_i \sim N(0, I)$, $n = 2000$ and $\boldsymbol{\epsilon}_i = (\epsilon_1, \epsilon_2, ..., \epsilon_5)^T \sim N_5(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix}$, $A = \begin{pmatrix} 1 & -1/2 \\ -1/2 & 1/2 \end{pmatrix}$ and D =diag$(1/2, 1/3, 1/4)$.

This is a multivariate model. Figure 2.3 plots mean values of respective r_1, r_2, Δ_m and Δ_f over 100 simulated data vs sizes of $\boldsymbol{\omega}$: $\{10, 20, \cdots, 100\}$. All four criteria show that FT estimates tend to improve as the size of $\boldsymbol{\omega}$ increases and then become stable. On the other hand, Zhu et al. (2010c) has demonstrated the advantage of FT for multivariate regression over other methods: the projective resampling method (Li

et al., 2008), the K-means inverse regression (Setodji and Cook, 2004), alternative SIR (Li et al., 2005), nearest neighbor inverse regression (Hsing, 1999) and moment approach (Yin and Bura, 2006). We omit the related comparisons here.

The left panel of Figure 2.4 shows the Weighted Chi-square Statistic, Scaled Statistic, Adjusted Statistic and SED test. Compared with the Weighted Chi-square Statistic and Adjusted Statistic, the Scaled Statistic is better. (Actually, we also use sample sizes $n = 1000$, but not reported here. The Scaled Statistic still performs well, which indicates the Scaled Statistic converges more quickly.) If the size of $\boldsymbol{\omega}$ is large enough, the performance of the Scaled Statistic becomes stable, and the proportion of correct decisions gets closer to 1, which agrees with the estimation accuracy. The Scaled Statistic is better than SED when the size of $\boldsymbol{\omega}$ is large enough. However, SED is not stable. When the size of $\boldsymbol{\omega}$ is large enough (over 60 as in Figure 2.4), its result is worse, contradicting the accuracy of the estimate. The middle and right panels of Figure 2.4 show TPR and FPR, respectively, for S-FT and S-SIR. TPR values are similar for the two methods with smaller FPR for S-FT when the size of $\boldsymbol{\omega}$ is between 20 to 80 compared to S-SIR. Regardless, FPRs are all relatively small using either S-SIR or S-FT.

Additionally, we change the number of predictors to be $p = \{10, 20\}$ and use sample sizes $n = \{1000, 2000\}$, with the number of response variables to be $q = \{5, 10, 15\}$ (not reported here). The number of predictors and sample size affect the asymptotic results in testing the dimension. As sample size increases, the performance of the Weighted Chi-square Statistic, Scaled Statistic, and Adjusted Statistic improve, especially for the Scaled Statistic. If the number of predictors increases, a larger sample size is needed for asymptotic results to converge. While adding some noise response variables and changing the number of response variables does not significantly affect the results.

Model 5. For $W = 0$, let $Y = X_1 + 0.1\epsilon$, with $B_1 = e_1$. Predictors $\mathbf{X}_{i1} \sim \frac{1}{4}N_p(\boldsymbol{\mu}_1, \Sigma_1) + \frac{1}{2}N_p(\boldsymbol{\mu}_2, \Sigma_2) + \frac{1}{4}N_p(\boldsymbol{\mu}_3, \Sigma_3)$, where $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_3 = (1, 0, ..., 0)^T$, $\boldsymbol{\mu}_2 = (2, 0, ..., 0)^T$, $\Sigma_1 = \Sigma_2 = \sqrt{0.1}I_p$ and $\Sigma_3 = \sqrt{10}I_p$. Let $p = 10$, and ϵ is a uniform (0,1), with 1000 observations. For $W = 1$, let Y be the Y_2 in the model 4 with

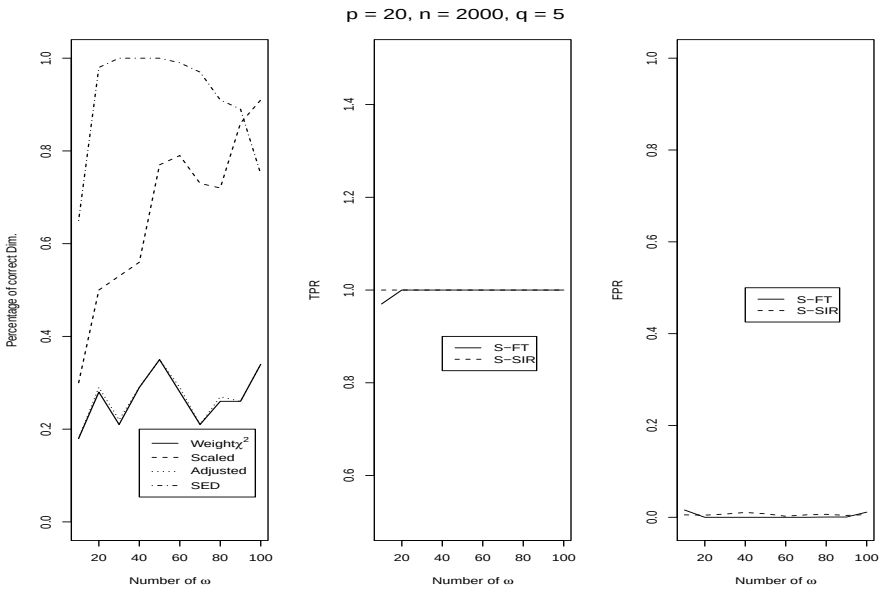

Figure 2.4: Left panel: percentages of correctly detected dimension over 100 simulated data vs sizes of $\boldsymbol{\omega}$: $\{10, 20, ..., 100\}$ in Model 4; Middle and Right panel: TPR and FPR over 100 simulated data vs sizes of $\boldsymbol{\omega}$.

Table 2.4: Mean values of respective r_1, r_2, Δ_m and Δ_f over 100 simulated data for PSIR and PFT and proportion of correctly detected dimension in Model 5.

	r_1	r_2	Δ_m	Δ_f	$Proportion_c$
PSIR	0.9889	0.9944	0.1249	0.2081	0.9700
PFT	0.9930	0.9965	0.1066	0.1640	1.0000

$B_2 = e_2 + e_3$ and 1000 observations.

This example compares PSIR and PFT. The number of slices for PSIR is 10, and the size of $\boldsymbol{\omega}$ for PFT is 50. We replicate 100 times for the model and then calculate the averages of respective r_1, r_2, Δ_m and Δ_f and the proportion of correctly detected dimensions using the Scaled Statistics, say, $Proportion_c$. Table 2.4 shows that PFT performs consistently better than PSIR does in every criterion.

Model 6. This is Model 4, except: $\beta_1 = e_1 + e_2 + e_3 + e_4$, $\beta_2 = e_5 + e_6 + \cdots + e_{12}$, $p = 1000$, and $n = 400$. This is a large p, small n problem.

We use path I algorithm in Section 2.4. We use 10 slices for SIR, 50 sizes of $\boldsymbol{\omega}$ for FT and $p_1 = 15$ as the number of predictors in each step for both SIR and FT. The

asymptotic Scaled Statistic test is used in each step for estimating the dimensionality. Table 2.5 reports the average values for each criterion over 100 simulated data. SSDR-FT performs consistently better than that of SSDR-SIR, except that both TPR and FPR are the same for the two methods.

Table 2.5: Accuracy for large p, small n data in Model 6

	Corr1	Corr2	Δ_f	Δ_m	TPR	FPR
SSDR-SIR	0.8764	0.8024	0.7264	0.4567	0.9783	0.0134
	(0.2034)	(0.1852)	(0.1058)	(0.0788)	(0.1062)	(0.0646)
SSDR-FT	0.9565	0.9106	0.6335	0.3912	0.9783	0.0134
	(0.0218)	(0.0350)	(0.1018)	(0.0705)	(0.1062)	(0.0646)

Data analysis

The data set is the"2015 Planning Database" (PDB) with 2010 Census and 2009-2013 American Community Survey data, which is publicly available (http://goo.gl/LlcwY7). PDB assembles information from housing, demographic, socioeconomic, and Census operational data, and accumulates at the block-group level. A census block is the smallest geographic unit used by the Census Bureau, and a block-group comprises multiple blocks, usually containing between 600 and 3,000 people. The PDB comprises approximately 220,000 block groups.

The response variable is the number of people with one type of health insurance coverage (Y). A total of 15 variables are identified as relevant candidate predictor variables. Because most of the variables are count numbers with a large range of values, we treat all of them as continuous variables.

We focus on the block groups in Rhode Island, which have 4270 blocks. We first excluded any observations where the variables had missing values. There were 4098 blocks left for Rhode Island. We then used Box-Cox transformation for the predictors to ensure that the linearity condition was approximately satisfied.

Using the Scaled Statistic for all the blocks in Rhode Island, the estimated dimension (using 50 as the size of ω) is one. In addition, if we plot the scatter plot (Figure 2.5) of response variable versus the first reduced variable, we can see the

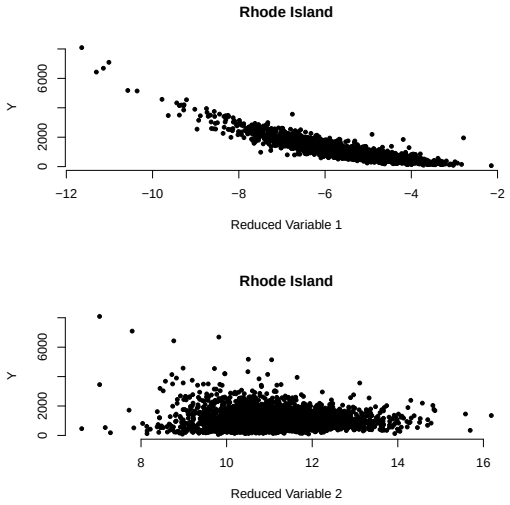

Figure 2.5: Scatter Plots: response variable versus the first and the second reduced variable.

strong association. The second reduced variable does not contribute much. Hence, one dimension is sufficient to capture the CS. Thus, we used one dimension for the following analysis.

To illustrate the advantages of FT, we used five datasets: the first 100 blocks, the first 200 blocks, the first 400 blocks, the first 800 blocks, and all blocks of Rhode Island. For each data, we estimated the vector $\hat{\beta}$ (of the CS). We then bootstrapped 100 samples from that data and obtained the bootstrap estimate $\hat{\beta}^b$ for each bootstrap sample. Then we compare means of r_2 between the bootstrap estimate $\hat{\beta}^b$ and $\hat{\beta}$ using the following methods: FT, SIR, SAVE, PHD, FIRE, and DIRE. For SIR and SAVE, we fix the number of slices to be 5, which is typically what researchers suggested. For FIRE and DIRE, the sequence of slices is $\{3, 4, 5, \cdots, 15\}$, which is what Cook and Zhang (2014) suggested. Table 2.6 shows the results. It indicates that when sample size increases, every method performs better, which is expected. However, none of them is comparable with FT, until sample size reaches to 4098. On the other hand, FT approach provides the most accurate and stable estimates among all these methods and across all sample sizes. Even in the small sample size of 100, FT still provides an accurate estimate with $r_2 = 0.9840$. It indicates that its estimates converge much faster than all other methods.

Table 2.6: Means and standard deviations of r_2 for each method: FT, SIR, SAVE, PHD, FIRE and DIRE over sample sizes: {100, 200, 400, 800, 4098}

r_2	Rhode (n=100)	Rhode (n=200)	Rhode (n=400)	Rhode (n=800)	Rhode (n=4098)
FT	0.9840	0.9879	0.9926	0.9956	0.9991
	(0.0079)	(0.0058)	(0.0034)	(0.002)	(4e-04)
SIR	0.5543	0.7072	0.8591	0.892	0.9754
	(0.2334)	(0.2209)	(0.1233)	(0.059)	(0.0147)
SAVE	0.4136	0.6017	0.7417	0.7319	0.9629
	(0.2676)	(0.2834)	(0.2612)	(0.2984)	(0.0281)
PHD	0.7665	0.6128	0.7787	0.7944	0.8597
	(0.2437)	(0.3054)	(0.1926)	(0.2355)	(0.1156)
FIRE	0.4857	0.5056	0.8296	0.9133	0.9869
	(0.2424)	(0.2954)	(0.1392)	(0.0500)	(0.0082)
DIRE	0.3816	0.3669	0.6911	0.9002	0.9882
	(0.2096)	(0.2157)	(0.1600)	(0.0561)	(0.0066)

2.6 Discussion

Using FT, we develop a complete package for estimating CS. We provide its estimator, algorithm and asymptotic properties. It is important to note that FT approach avoids the trouble of selecting the number of slices in inverse regression and provides a natural solution for multivariate response. We further extended this approach to partial SDR, SVS, and large p, small n data. Given the current FT approach, a general discussion about inverse regression family may be developed. Such an investigation is our on-going project.

Chapter 3 A minimum discrepancy approach for Fourier transform inverse regression in sufficient dimension reduction

3.1 Introduction

Sliced inverse regression (SIR; Li 1991) and sliced average variance estimation (SAVE; Cook and Weisberg 1991) are the first two methods proposed for SDR. The key steps in SIR and SAVE approach is to firstly find a kernel matrix, then use the column space of its first d eigenvectors to estimate the central subspace. Cook (2004) developed a procedure to test predictor contributions in SDR by reformulating eigen-decomposition as an ordinal least square problem. Along with this idea, Cook and Ni (2005) also investigated the hypothesis tests via minimum discrepancy function and developed inverse regression estimators (IRE). In the small sample size setting, Ni and Cook (2007) introduced robust IRE and corresponding hypothesis tests. Then Cook and Zhang (2014) proposed fused estimators (FIRE and DIRE) based on an optimal inverse regression estimator.

Fourier transform idea was first been introduced to SDR in Zhu and Zeng (2006). And Zhu et al. (2010c) further developed a unified method to recover the central dimension reduction subspace in regressions with multivariate responses on high-dimensional predictors. Weng and Yin (2018) recently investigated Fourier transform (FT) in testing structural dimension, partial SDR, and SVS (S-FT). In this chapter, we employ the Fourier transform approach into quadratic discrepancy function with four different inner product matrices, which leads to "degenerate", "special", "robust", and "partial" estimators. The degenerate and special estimation have less computational cost since they simplify the calculation of the inverse covariance matrix. In addition, the robust estimation has advantage on small sample size as it only uses second moments of the predictor for estimation and inference. The partial estimation applies Fourier transform in partial sufficient dimension reduction (Chiaromonte et al., 2002). To perform SVS, we propose to add a coordinate-independent penalty

to the quadratic discrepancy functions. A coordinate descent algorithm and Stiefel manifold optimization (Qian et al., 2018) is given to solve the optimization problem. We also conduct the conditional and marginal hypothesis tests for identifying structural dimensions and significance of predictors. And the simulation results show that the power of our proposed tests is close to 1.

This chapter is organized as follows: Section 3.2 provides inverse regression estimators and asymptotic results in four situations: general, degenerate, robust and partial. Section 3.3 develops hypothesis tests for structural dimension and predictors, while Section 3.4 proposes sufficient variable selection approach. In Section 3.5, simulation studies and real data analysis are presented to support our theoretical analysis. We conclude this chapter by summarizing the pros and cons of the proposed approaches in Section 3.6. All proofs are included in the Appendix.

3.2 Methodology and Estimation

Review Fourier Transform Approach

Let $\mathbf{Z} = \Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$, where $\boldsymbol{\mu}$ and Σ are the mean and covariance matrix of $\mathbf{X}$. Let $m(\mathbf{y}) = \mathrm{E}(\mathbf{Z}|\mathbf{Y} = \mathbf{y})$. Under the linearity condition: $\mathrm{E}(\mathbf{Z}|\mathbf{P}_{\mathcal{S}_{\mathbf{Y}|\mathbf{Z}}}\mathbf{Z}) = \mathbf{P}_{\mathcal{S}_{\mathbf{Y}|\mathbf{Z}}}\mathbf{Z}$, $m(\mathbf{y}) \in \mathcal{S}_{\mathbf{Y}|\mathbf{Z}}$ (Cook, 1998), where $\mathbf{P}_{\mathcal{S}_{\mathbf{Y}|\mathbf{Z}}}$ is the orthogonal projection onto $\mathcal{S}_{\mathbf{Y}|\mathbf{Z}}$ in the usual inner product. Thus, the space spanned by $m(\mathbf{y})$, $\mathcal{S}_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})}$, recovers part of the CS. Zhu et al. (2010c) developed a unification of inverse regression and forward regression method, which was further developed by Weng and Yin (2018), called the Fourier transform (FT) approach.

Fourier transform is applied on the conditional mean $m(\mathbf{y})$, that is, $\psi(\boldsymbol{\omega}) = \int e^{i\boldsymbol{\omega}^T\mathbf{y}} m(\mathbf{y}) f(\mathbf{y}) d\mathbf{y} = a(\boldsymbol{\omega}) + ib(\boldsymbol{\omega}), \boldsymbol{\omega} \in \mathbb{R}^q$, where $a(\boldsymbol{\omega})$, $b(\boldsymbol{\omega})$ are the respective real and imaginary parts of $\psi(\boldsymbol{\omega})$, and $f(\mathbf{y})$ is the marginal density function of $\mathbf{y}$. Zhu et al. (2010c) and Weng and Yin (2018) proved that $\psi(\boldsymbol{\omega}) = \mathrm{E}(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\mathbf{Z})$, and $\mathcal{S}_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})} = \mathrm{Span}\{\psi(\boldsymbol{\omega}), \boldsymbol{\omega} \in \mathbb{R}^q\} = \mathrm{Span}\{a(\boldsymbol{\omega}), b(\boldsymbol{\omega}), \boldsymbol{\omega} \in \mathbb{R}^q\}$. Zhu et al. (2010c) provided an argument to choose $\boldsymbol{\omega}$, which satisfies $P(|\boldsymbol{\omega}^T\mathbf{Y}| > \pi) \leq s$ with $s = 0.1$. Then $\boldsymbol{\omega}$ is randomly generated from $N(\mathbf{0}, \frac{s\pi^2}{\mathrm{E}(\mathbf{Y}^T\mathbf{Y})}I)$. Weng and Yin (2018) showed

that a finite number of $\boldsymbol{\omega}$ is enough to recover $\mathcal{S}_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})}$. In our limited simulations, 50 Fourier transforms ($\boldsymbol{\omega}$'s) are sufficient, and estimates are quite stable and accurate. Instead of using the sample mean of $\mathrm{E}(\mathbf{Y}^T\mathbf{Y})$, the median provides more robust estimate. For instance, the model $Y = \exp(\mathbf{X}^T\beta) + \epsilon$ and $Y = \frac{1}{\mathbf{X}^T\beta} + \epsilon$ have some extreme Y values or outliers. The rule of thumb is that if the ratio of sample mean over median is over 100, then the median is preferable.

Due to the equivalence between $\mathcal{S}_{\mathbf{Y}|\mathbf{Z}}$ and $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ (Cook, 1998), we can operate in either scale of $\mathbf{X}$ and $\mathbf{Z}$. The following sections are developed in $\mathbf{X}$ scale.

Fourier Transform Inverse Regression Estimators

We employ Fourier transform approach by minimizing a quadratic discrepancy function following Cook and Ni (2005). Assume that the number of $\boldsymbol{\omega}$ is m. The working meta-parameter is defined as $\mathcal{S}_\xi = \sum_{j=1}^m \mathrm{Span}(\boldsymbol{\xi}_j^R, \boldsymbol{\xi}_j^I)$, where $\boldsymbol{\xi}_j = \Sigma^{-1}[\mathrm{E}(e^{i\boldsymbol{\omega}_j^T\mathbf{Y}}\mathbf{X}) - \mathrm{E}(e^{i\boldsymbol{\omega}_j^T\mathbf{Y}})\mathrm{E}(\mathbf{X})] \in \mathbb{C}^p$, and the indexes R and I represent the real and imaginary parts of a vector. If the linearity condition holds, then $\mathcal{S}_\xi \subseteq \mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. We further assume that the coverage condition holds, that is $\mathcal{S}_\xi = \mathcal{S}_{\mathbf{Y}|\mathbf{X}}$. Let $d = \dim(\mathcal{S}_\xi)$ and $\beta \in \mathbb{R}^{p\times d}$ be a basis of CS, and there exists a vector $\boldsymbol{\gamma}_j \in \mathbb{C}^p$ such that $\boldsymbol{\xi}_j = \beta\boldsymbol{\gamma}_j$ for each j. Define $\xi = (\boldsymbol{\xi}_1^R, \boldsymbol{\xi}_1^I, ..., \boldsymbol{\xi}_m^R, \boldsymbol{\xi}_m^I) = \beta\nu$, where $\nu = (\boldsymbol{\gamma}_1^R, \boldsymbol{\gamma}_1^I, ..., \boldsymbol{\gamma}_m^R, \boldsymbol{\gamma}_m^I)$. Suppose $\{\mathbf{y}_i, \mathbf{x}_i\}$,$i = 1, ..., n$ are iid samples of $(\mathbf{Y}, \mathbf{X})$. Let $\bar{\mathbf{x}}$ be the sample mean of $\mathbf{X}$. The sample version of $\boldsymbol{\xi}_j$ is $\hat{\boldsymbol{\xi}}_j = \hat{\Sigma}^{-1}(\frac{1}{n}\sum_{k=1}^n e^{i\boldsymbol{\omega}_j^T\mathbf{y}_k}\mathbf{x}_k - \frac{1}{n}\sum_{k=1}^n e^{i\boldsymbol{\omega}_j^T\mathbf{y}_k}\bar{\mathbf{x}})$. Let $\hat{\xi} = (\hat{\boldsymbol{\xi}}_1^R, \hat{\boldsymbol{\xi}}_1^I, ..., \hat{\boldsymbol{\xi}}_m^R, \hat{\boldsymbol{\xi}}_m^I) \in \mathbb{R}^{p\times 2m}$. Define a random vector $\boldsymbol{\epsilon} = (\epsilon_1^R, \epsilon_1^I, ..., \epsilon_m^R, \epsilon_m^I)^T$, where ϵ_j^R and ϵ_j^I are the real and imaginary parts of $e^{i\boldsymbol{\omega}_j^T\mathbf{Y}} - \mathrm{E}e^{i\boldsymbol{\omega}_j^T\mathbf{Y}} - \mathbf{Z}^T\mathrm{E}(e^{i\boldsymbol{\omega}_j^T\mathbf{Y}}\mathbf{Z})$ for $j = 1, ..., m$. They are the population residuals from ordinary least squares fit of $e^{i\boldsymbol{\omega}_j^T\mathbf{Y}}$ on $\mathbf{Z}$.

Fourier transform quadratic discrepancy function with the inner product matrix V is defined as

$$F_d(B, C; V) = [\mathrm{vec}(\hat{\xi}) - \mathrm{vec}(BC)]^T V [\mathrm{vec}(\hat{\xi}) - \mathrm{vec}(BC)]. \tag{3.1}$$

where the columns of $B \in \mathbb{R}^{p\times d}$ estimate an orthogonal basis of the CS, and $C \in \mathbb{R}^{d\times 2m}$ represents the coordinates of ξ relative to B.

Theorem 1. *Assume that* $\{\mathbf{y}_k, \mathbf{x}_k\}$,$k = 1, ..., n$ *are random samples of* $(\mathbf{Y}, \mathbf{X})$ *with finite fourth moments. Then*

$$\sqrt{n}[\text{vec}(\hat{\xi}) - \text{vec}(\beta\nu)] \xrightarrow{D} N(0, \Gamma),$$

where $\Gamma = \text{Cov}\{\text{vec}[\Sigma^{-1}(\mathbf{X} - \boldsymbol{\mu})\boldsymbol{\epsilon}^T]\} \in \mathbb{R}^{2pm \times 2pm}$.

Theorem 1 provides an asymptotic covariance matrix of the random vector $\hat{\xi}$ by using the central limit theorem. If m is too large, the asymptotic covariance matrix contains considerable noise. This noise will not affect the accuracy of the estimate of the CS, but it will deteriorate the ability to detect dimensions of the CS (see simulation results in Section 3.5).

Theorem 2. *Assume that* $\{\mathbf{y}_k, \mathbf{x}_k\}$,$k = 1, ..., n$ *are random samples of* $(\mathbf{Y}, \mathbf{X})$ *with finite fourth moments. Let* $(\hat{\beta}, \hat{\nu}) = \arg\min_{B,C} F_d(B, C; \hat{\Gamma}^{-1})$*, where* $\hat{\Gamma}$ *is a consistent estimate of* Γ*. Then the following results hold:*

1. $\text{vec}(\hat{\beta}\hat{\nu})$ *is asymptotically efficient, and* $\sqrt{n}[\text{vec}(\hat{\beta}\hat{\nu}) - \text{vec}(\beta\nu)]$ *is asymptotically normal with mean 0 and covariance matrix* $\Delta(\Delta^T\Gamma^{-1}\Delta)^-\Delta^T$*, where* $\Delta = (v^T \otimes I_p, I_{2m} \otimes \beta)$ *with* $2mp \times d(p + 2m)$ *dimensions.*

2. $n\hat{F}_d$ *has an asymptotic chi-square distribution with degrees of freedom* $(p - d)(2m - d)$.

3. $\text{Span}(\hat{\beta})$ *is a consistent estimator of* $\mathcal{S}_\xi$.

Part 1 of Theorem 2 indicates that $\text{vec}(\hat{\beta}\hat{\nu})$ from optimizing quadratic discrepancy function (3.1) is asymptotically close to $\text{vec}(\beta\nu)$. Part 2 of Theorem 2 provides an asymptotic distribution of $n\hat{F}_d$, which can be employed in the dimension and hypothesis tests. Part 3 of Theorem 2 indicates that when the sample size is large enough, the subspace spanned by $\hat{\beta}$ will approach CS. The estimate $\hat{\beta}$ minimizing the quadratic discrepancy function (3.1) is denoted as a Fourier transform inverse regression estimator (FT-IRE). Even the discrepancy function approach achieves an optimal estimator, they are computationally expensive because of algorithm iterates until convergence. In order to solve this problem, we develop degenerated and robust

quadratic discrepancy functions. The degenerate version changes the inner product matrix V into diagonal block matrices with fewer parameters, resulting in less computation time. The robust version changes $\boldsymbol{\epsilon}$ into a simplified version to handle small sample size.

Degenerate and Special Estimators

In the previous section, choosing different inner product matrices V leads to different estimators. One disadvantage of FT-IRE is that the asymptotic covariance matrix has $O(4p^2m^2)$ parameters to estimate. Alternatively, diagonal block inner product matrices are used to reduce the number of parameters in covariance matrices. We conduct m Fourier transforms, then they are divided into K parts $\{\omega_j^{(l)}\}_{j=1}^{m_l}$ with $m_l > 0, l = 1, ..., K$ and $\sum_{l=1}^{K} m_l = m$. For each part of $\{\omega_j^{(l)}\}_{j=1}^{m_l}$, $V_l \in \mathbb{R}^{2pm_l \times 2pm_l}$ are the inner product matrices. Degenerated quadratic discrepancy functions combine the quadratic discrepancy functions (3.1) from each part, that is

$$F_d(B, C; \{V_l\}) = \sum_{l=1}^{K} [\text{vec}(\hat{\xi}_l) - \text{vec}(BC_l)]^T V_l [\text{vec}(\hat{\xi}_l) - \text{vec}(BC_l)], \tag{3.2}$$

which is equivalent to replace V in function (3.1) with new inner product matrix $\text{diag}(\{V_l\})$. Compared to FT-IRE, the inner product matrix of the degenerated version has fewer parameters to estimate, which makes it structurally simpler and computationally cheaper.

Theorem 1 is employed to find the asymptotic covariance matrix Γ_l for each part l, then the inner product matrix of function (3.2) is constructed by $\Gamma_D^{-1} = \text{diag}\{\Gamma_1^{-1}, ..., \Gamma_K^{-1}\}$. Using the diagonal block inner product matrix, the objective quadratic function (3.2) is to sum up quadratic discrepancy functions (3.1) for each part, pretending each part is independent of each other. We denoted $\hat{\beta}$ by minimizing function (3.2) as a Fourier transform degenerated inverse regression estimator (FT-DIRE). The following theorem states the consistency of estimate and asymptotic distribution of test statistics.

Theorem 3. *Assume that* $\{\mathbf{y}_k, \mathbf{x}_k\}$, $k = 1, ..., n$ *are random samples on* $(\mathbf{Y}, \mathbf{X})$ *with finite fourth moments. Let* $(\hat{\beta}, \hat{\nu}) = \arg\min_{B,C} F_d(B, C; \hat{\Gamma}_D^{-1})$. *Then the following results hold:*

1. $\mathrm{Span}(\hat{\beta})$ *is a consistent estimator of* $\mathcal{S}_\xi$.
2. *As* $n \to \infty$, $\hat{\Lambda}_d = n\hat{F}_d(B, C; \hat{\Gamma}_D^{-1}) \xrightarrow{D} \sum_{k=1}^{(p-d)(2m-d)} \lambda_k C_k$ *where the* C_k*s are independent chi-square random variables each with one degree of freedom and* $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{(p-d)(2m-d)}$ *are eigenvalues of the covariance matrix* $Q_\Phi \Omega Q_\Phi$, *where* Φ *and* Ω *are defined in the Appendix.*

We also introduce a special case, which is equivalent to FT (a spectral decomposition approach) (Weng and Yin, 2018). If $V_l = \Sigma$, the estimate using the degenerated discrepancy approach would be equivalent to the spectral decomposition estimate using kernel matrix $\sum_{j=1}^m \Re[\hat{\boldsymbol{\psi}}_{\omega_j} \bar{\hat{\boldsymbol{\psi}}}_{\omega_j}^T]$, where $\Re$ means the real part of complex value. In fact,

$$\begin{aligned} F_d(B, C; \mathrm{diag}\{\hat{\Sigma}\}) &= \sum_{l=1}^K F_d(B, C_l; \hat{\Sigma}) \\ &= [\mathrm{vec}(\hat{\xi}) - \mathrm{vec}(BC)]^T \mathrm{diag}\{\hat{\Sigma}\}[\mathrm{vec}(\hat{\xi}) - \mathrm{vec}(BC)]. \end{aligned} \tag{3.3}$$

Let $(\hat{\beta}, \hat{\nu})$ be a minimizer of discrepancy function (3.3). We denote $\hat{\beta}$ as Fourier transform special inverse regression estimator (FT-SIRE). We summarize our discussion as the following lemma and theorem.

Lemma 1. *Assume that* $\{\mathbf{y}_k, \mathbf{x}_k\}$,$k = 1, \cdots, n$ *are random samples on* $(\mathbf{Y}, \mathbf{X})$ *with finite fourth moments, let* $\hat{u}_1, \cdots, \hat{u}_p$ *be the eigenvectors of* $\hat{M}_{ft} = \sum_{j=1}^m \Re[\hat{\boldsymbol{\psi}}_{\omega_j} \bar{\hat{\boldsymbol{\psi}}}_{\omega_j}^T]$ *corresponding to eigenvalues* $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$. *Then the minimizer* $\mathrm{Span}(\hat{\beta})$ *is equal to* $\mathrm{Span}(\hat{u}_1, \cdots, \hat{u}_d)$.

Theorem 4. *Assume that* $\{\mathbf{y}_k, \mathbf{x}_k\}$,$k = 1, \cdots, n$ *are random samples on* $(\mathbf{Y}, \mathbf{X})$ *with finite fourth moments. Let* $(\hat{\beta}, \hat{\nu}) = \arg\min_{B,C} F_d(B, C; \mathrm{diag}\{\hat{\Sigma}\})$. *Then the following results hold:*

1. $\mathrm{Span}(\hat{\beta})$ *is a consistent estimator of* $\mathcal{S}_\xi$.

2. *As* $n \to \infty$, $\hat{\Lambda}_d = n\hat{F}_d(B, C; \text{diag}\{\hat{\Sigma}\}) \xrightarrow{D} \sum_{k=1}^{(p-d)(2m-d)} \lambda_k C_k$ *where the* C_k*s are independent chi-square random varibles each with one degree of freedom and* $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{(p-d)(2m-d)}$ *are eigenvalues of the covariance matrix* $Q_\Phi \Omega Q_\Phi$*, where* Φ *and* Ω *are defined in the Appendix.*

Both Theorem 3 and Theorem 4 indicate that the corresponding test statistic $\hat{\Lambda}_d$ follows a weighted chi-square distribution instead of a chi-square distribution as in Theorem 2. The test statistic can be applied in tests of dimension and hypothesis. Reducing the number of parameters of the inner product matrix improves not only accuracy but also computationally efficiency.

Robust Estimators

We introduce a robust version of FT-IRE, called FT-RIRE in which only second-order moments of predictors is required. The robust version is to replace the previous FT-IRE residual variables $e^{i\boldsymbol{\omega}_j^T \mathbf{Y}} - \text{E}e^{i\boldsymbol{\omega}_j^T \mathbf{Y}} - \mathbf{Z}^T \text{E}(e^{i\boldsymbol{\omega}_j^T \mathbf{Y}} \mathbf{Z})$ with $e^{i\boldsymbol{\omega}_j^T \mathbf{Y}} - \text{E}e^{i\boldsymbol{\omega}_j^T \mathbf{Y}}$ for $j = 1, ..., m$. Let $\widetilde{\boldsymbol{\epsilon}} = (\widetilde{\epsilon}_1^R, \widetilde{\epsilon}_1^I, ..., \widetilde{\epsilon}_m^R, \widetilde{\epsilon}_m^I)^T$, where $\widetilde{\epsilon}_j^R, \widetilde{\epsilon}_j^I$ are the real and imaginary parts of $e^{i\boldsymbol{\omega}_j^T \mathbf{Y}} - \text{E}e^{i\boldsymbol{\omega}_j^T \mathbf{Y}}$ for $j = 1, ..., m$. Let $\widetilde{\boldsymbol{\xi}}_j = \Sigma^{-1}(\frac{1}{n}\sum_{k=1}^n e^{i\boldsymbol{\omega}_j^T \mathbf{y}_k} \mathbf{x}_k - \frac{1}{n}\sum_{k=1}^n e^{i\boldsymbol{\omega}_j^T \mathbf{y}_k} \bar{\mathbf{x}})$ and $\widetilde{\xi} = (\widetilde{\boldsymbol{\xi}}_1^R, \widetilde{\boldsymbol{\xi}}_1^I, ..., \widetilde{\boldsymbol{\xi}}_m^R, \widetilde{\boldsymbol{\xi}}_m^I) \in \mathbb{R}^{p \times 2m}$.

Theorem 5. *Assume that* $\{\mathbf{y}_k, \mathbf{x}_k\}$*,* $k = 1, ..., n$ *are random samples on* $(\mathbf{Y}, \mathbf{X})$ *with finite second moments. Then*

$$\sqrt{n}[\text{vec}(\widetilde{\xi}) - \text{vec}(\beta\nu)] \xrightarrow{D} N(0, \widetilde{\Gamma}),$$

where $\widetilde{\Gamma} = (I \otimes \Sigma^{-1/2})\text{Cov}[\text{vec}(\mathbf{Z}\widetilde{\boldsymbol{\epsilon}}^T)](I \otimes \Sigma^{-1/2})$.

Theorem 5 indicates that $\widetilde{\xi}$ is an asymptotic estimate of βv. Then, we can define robust quadratic discrepancy function as

$$F_d(B, C; \widetilde{G}^{-1}) = [\text{vec}(\hat{\xi}) - \text{vec}(BC)]^T \widetilde{G}^{-1} [\text{vec}(\hat{\xi}) - \text{vec}(BC)],$$

where $\widetilde{G} = (I \otimes \hat{\Sigma}^{-1/2})\widehat{\text{Cov}}[\text{vec}(\mathbf{Z}\widetilde{\boldsymbol{\epsilon}}^T)](I \otimes \hat{\Sigma}^{-1/2})$.

Theorem 6. *Assume that* $\{\mathbf{y}_k, \mathbf{x}_k\}$*,*$k = 1, ..., n$ *are random samples on* $(\mathbf{Y}, \mathbf{X})$ *with finite second moments and let* $(\hat{\beta}, \hat{\nu}) = \arg\min_{B,C} F_d(B, C; \widetilde{G}^{-1})$*. Then*

1. *$n\hat{F}_d(B, C; \widetilde{G}^{-1})$ has an asymptotic chi-square distribution with degrees of freedom $(p-d)(2m-d)$.*

2. *$\text{Span}(\hat{\beta})$ is a consistent estimator of $\mathcal{S}_\xi$.*

The proof is very similar to the proof of Theorem 2. Hence, we omit it. We also define a diagonal block inner product matrix for FT-RIRE that is $\widetilde{G}_D^{-1} = \text{diag}\{\widetilde{G}_1^{-1}, ..., \widetilde{G}_K^{-1}\}$ following the notations in Section 3.2. The estimator minimizing $F_d(B, C; \widetilde{G}_D^{-1})$ denotes as Fourier transform degenerated robust inverse regression estimator (FT-DRIRE).

Theorem 7. *Assume that $\{\mathbf{y}_k, \mathbf{x}_k\}$,$k = 1, ..., n$ are random samples on $(\mathbf{Y}, \mathbf{X})$ with finite second moments. Let $(\hat{\beta}, \hat{\nu}) = \arg\min_{B,C} F_d(B, C; \widetilde{G}_D^{-1})$. Then the following results hold:*

1. *$\text{Span}(\hat{\beta})$ is a consistent estimator of $\mathcal{S}_\xi$.*

2. *As $n \to \infty$, $\hat{\Lambda}_d = n\hat{F}_d(B, C; \widetilde{G}_D^{-1}) \xrightarrow{D} \sum_{k=1}^{(p-d)(2m-d)} \lambda_k C_k$ where the C_ks are independent chi-square random variables each with one degree of freedom and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{(p-d)(2m-d)}$ are eigenvalues of the covariance matrix $Q_\Phi \Omega Q_\Phi$, where Φ and Ω are defined in the Appendix.*

Partial Estimator

Without loss of generality, let W be a categorical variable with K levels. Chiaromonte et al. (2002) defined partial CS relative to $\mathbf{X}$ to be the intersection of all subspace spanned by $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ such that $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | (\boldsymbol{\beta}^T\mathbf{X}, W)$, denoted by $\mathcal{S}^W_{\mathbf{Y}|\mathbf{X}}$. The relationship between partial CS and conditional CS is $\mathcal{S}^W_{\mathbf{Y}|\mathbf{X}} = \bigoplus_{k=1}^K \mathcal{S}_{\mathbf{Y}_k|\mathbf{X}_k}$, where $\mathcal{S}_{\mathbf{Y}_k|\mathbf{X}_k}$ is the CS conditioning on level k and $\bigoplus$ means direct sum. For each level, we generate $\{\boldsymbol{\omega}_j^{(l)}\}_{j=1}^m, l = 1, \cdots, K$. Then we calculate $\hat{\xi}_l$ and $\hat{\Gamma}_l^{-1}$ in Section 3.2 for each level to construct partial quadratic discrepancy function

$$\begin{aligned} F_d^K(B, C; \text{diag}\{\hat{\Gamma}_l^{-1}\}) &= \sum_{l=1}^K F_d(B, C_l; \hat{\Gamma}_l^{-1}) \\ &= \sum_{l=1}^K [\text{vec}(\hat{\xi}_l) - \text{vec}(BC_l)]^T \hat{\Gamma}_l^{-1} [\text{vec}(\hat{\xi}_l) - \text{vec}(BC_l)], \end{aligned} \tag{3.4}$$

which is the same as function (3.1) with $V = \text{diag}(\hat{\Gamma}_l^{-1})$. We denote the estimate by minimizing equation (3.4) as Fourier transform partial inverse regression estimator (FT-PIRE). Note that the quadratic discrepancy function for each level is independent of each other, we can prove the next result, Theorem 8, using Theorem 2. Hence, its proof is omitted.

Theorem 8. *Assume that* $\{\mathbf{y}_k, \mathbf{x}_k\}$,$k = 1, ..., n$ *are random samples on* $(\mathbf{Y}, \mathbf{X})$ *with finite fourth moments. Let* $(\hat{\beta}, \hat{\nu}) = \arg\min_{B,C} F_d^K(B, C; \text{diag}\{\hat{\Gamma}_l^{-1}\})$. *Then the following results hold:*

1. *Span*$(\hat{\beta})$ *is a consistent estimator of* $\mathcal{S}_\xi$.
2. $n\hat{F}_d^K$ *has an asymptotic chi-square distribution with degrees of freedom* $K(p - d)(2m - d)$.

Algorithm

The columns of $\hat{\boldsymbol{\xi}}$ are the real and imaginary parts of $\hat{\boldsymbol{\xi}}_j = \hat{\Sigma}^{-1}(\frac{1}{n}\sum_{k=1}^n e^{i\boldsymbol{\omega}_j^T \mathbf{y}_k}\mathbf{x}_k - \frac{1}{n}\sum_{k=1}^n e^{i\boldsymbol{\omega}_j^T \mathbf{y}_k}\bar{\mathbf{x}})$, and the sample version of $\boldsymbol{\varepsilon}$ is $\hat{\boldsymbol{\varepsilon}}_k = (\hat{\varepsilon}_{1k}^R, \hat{\varepsilon}_{1k}^I, ..., \hat{\varepsilon}_{mk}^R, \hat{\varepsilon}_{mk}^I)^T$, where $\hat{\varepsilon}_{jk}^R$ and $\hat{\varepsilon}_{jk}^I$ are the real and imaginary parts of $\hat{\varepsilon}_{jk} = e^{i\boldsymbol{\omega}_j^T \mathbf{y}_k} - \frac{1}{n}\sum_{k=1}^n e^{i\boldsymbol{\omega}_j^T \mathbf{y}_k}\mathbf{x}_k - (\mathbf{x}_k - \bar{\mathbf{x}})^T\hat{\boldsymbol{\xi}}_j$, $k = 1, ..., n$ and $j = 1, ..., m$. Then the sample version of inner product matrix Γ is to replace Σ with $\hat{\Sigma}$ and $\boldsymbol{\varepsilon}$ with $\hat{\boldsymbol{\varepsilon}}$. While doing FT-RIRE, the sample version of $\widetilde{\boldsymbol{\varepsilon}}$ is $\hat{\widetilde{\boldsymbol{\varepsilon}}}_k = (\hat{\widetilde{\varepsilon}}_{1k}^R, \hat{\widetilde{\varepsilon}}_{1k}^I, ..., \hat{\widetilde{\varepsilon}}_{mk}^R, \hat{\widetilde{\varepsilon}}_{mk}^I)^T$, where $(\hat{\widetilde{\varepsilon}}_{1k}^R, \hat{\widetilde{\varepsilon}}_{1k}^I)$ are the real and imaginary parts of $\hat{\widetilde{\varepsilon}}_{jk} = e^{i\boldsymbol{\omega}_j^T \mathbf{y}_k} - \frac{1}{n}\sum_{k=1}^n e^{i\boldsymbol{\omega}_j^T \mathbf{y}_k}\mathbf{x}_k$. The Moore-Penrose generalized inverse of $\hat{\Gamma}$ is used because $\hat{\Gamma}$ might be singular. We present the algorithm below:

1. Choose an initial value for $B \in \mathbb{R}^{p\times d}$. An initial choice will affect the speed of convergence. One choice could be $\boldsymbol{e}_i = (0, \cdots, 0, 1, 0, \cdots, 0)^T$ with i^{th} place 1 and other places 0s. Alternatively, we use the spectral decomposition result from FT (Weng and Yin, 2018).
2. Fixed B, update C by minimizing $F_d(B, C; V)$. Actually, vec(C) can be constructed by fitting linear regression $V^{1/2}\text{vec}(\hat{\xi})$ on $V^{1/2}(I_{2m}\otimes B)$, that is $\text{vec}(C) = [(I_{2m} \otimes B^T)V(I_{2m} \otimes B)]^{-1}(I_{2m} \otimes B^T)V\text{vec}(\hat{\xi})$. Assign err to be $F_d(B, C; V)$.

3. Fixed C, minimize $F_d(B, C; V)$ with respect to one column of B, subject to unit norm and orthogonal to other columns (keeping them constants). For this partial minimization problem, the quadratic discrepancy function is $F(b) = (\boldsymbol{\alpha}_k - (\boldsymbol{c}_k^T \otimes I_p)Q_{B_{(-k)}}\boldsymbol{b})^T V(\boldsymbol{\alpha}_k - (\boldsymbol{c}_k^T \otimes I_p)Q_{B_{(-k)}}\boldsymbol{b})$, where $\boldsymbol{\alpha}_k = \text{vec}(\hat{\xi} - B_{(-k)}C_{(-k)})$, $\boldsymbol{c}_k$ is kth column of C, $C_{(-k)}$ (or $B_{(-k)}$) are deleting k^{th} column from C (or B) and $Q_{B_{(-k)}}$ is orthogonal complement of Span$(B_{(-k)})$.

   a) For $k = 1, ..., d$:

      i. Denote $B = (\boldsymbol{b}_1, ..., \boldsymbol{b}_{k-1}, \boldsymbol{b}_k, \boldsymbol{b}_{k+1}, ..., \boldsymbol{b}_d)$ and update $\hat{\boldsymbol{b}}_k = Q_{B_{(-k)}}[Q_{B_{(-k)}}(\boldsymbol{c}_k^T \otimes I_p)V(\boldsymbol{c}_k^T \otimes I_p)Q_{B_{(-k)}}]^- Q_{B_{(-k)}}(\boldsymbol{c}_k^T \otimes I_p)V\boldsymbol{\alpha}_k$, then normalize $\hat{\boldsymbol{b}}_k$ using $\hat{\boldsymbol{b}}_k/||\hat{\boldsymbol{b}}_k||$.

      ii. Update B by replace $\boldsymbol{b}_k$ with $\hat{\boldsymbol{b}}_k$ and update C like step 2.

   b) Update err with $F_d(B, C; V)$.

4. Return to step 3 until err less than 10^{-6}.

5. The resulting estimates are $\hat{\beta} = (\hat{\boldsymbol{b}}_1, \hat{\boldsymbol{b}}_2, \cdots, \hat{\boldsymbol{b}}_d)$.

For FT-DIRE, FT-RIRE, and FT-SIRE, we only need to change the V matrix correspondingly. Furthermore, we need to know d before applying this algorithm. We employ the marginal dimension hypotheses: $d = t$ versus $d > t$. The marginal dimension hypothesis uses asymptotic distributions of $n\hat{F}_d$. Based on those theorems, sequential hypothesis of the form $d = t$ versus $d > t$ are constructed to test dimension d, where the value of t beginning with zero. The p-value for each hypothesis is calculated from asymptotic distribution under the null hypothesis $d = t$. If we fail to reject the null hypothesis, we say $d = t$. Otherwise, we increase t by 1 and continue the same process until we fail to reject.

3.3 Hypothesis Tests

Following Cook (2004), Cook and Ni (2005), and Ni and Cook (2007), we develop hypothesis tests for Fourier transform. Let $\mathcal{H}$ be a user-specified subspace for predictors

with r dimension. Only $r \leq p - \dim(\mathcal{S}_{\mathbf{Y}|\mathbf{X}})$ is considered, otherwise $\mathbf{Y} \perp\!\!\!\perp P_{\mathcal{H}}\mathbf{X}|Q_{\mathcal{H}}\mathbf{X}$ is not true, where $P_{\mathcal{H}}$ is the orthogonal projection onto $\mathcal{H}$ in the usual inner product and, $Q_{\mathcal{H}}$ is the orthogonal projection onto the orthogonal complement of $\mathcal{H}$ and $Q_{\mathcal{H}} = I - P_{\mathcal{H}}$. The null hypothesis $\mathbf{Y} \perp\!\!\!\perp P_{\mathcal{H}}\mathbf{X}|Q_{\mathcal{H}}\mathbf{X}$ is equivalent to test $P_{\mathcal{H}}\mathcal{S}_{\mathbf{Y}|\mathbf{X}} = \mathcal{O}_p$, where $\mathcal{O}_p$ is the origin in $\mathbb{R}^p$. Under the linearity and coverage conditions, $\mathcal{S}_{\mathbf{Y}|\mathbf{X}} = \mathcal{S}_\xi$. Hence, $\mathbf{Y} \perp\!\!\!\perp P_{\mathcal{H}}\mathbf{X}|Q_{\mathcal{H}}\mathbf{X} \Leftrightarrow P_{\mathcal{H}}\mathcal{S}_\xi = \mathcal{O}_p$. Then the following three hypothesis tests are considered in this session:

- Marginal predictor hypotheses: $P_{\mathcal{H}}\mathcal{S}_{\mathbf{Y}|\mathbf{X}} = \mathcal{O}_p$ versus $P_{\mathcal{H}}\mathcal{S}_{\mathbf{Y}|\mathbf{X}} \neq \mathcal{O}_p$.
- Joint hypotheses: $P_{\mathcal{H}}\mathcal{S}_{\mathbf{Y}|\mathbf{X}} = \mathcal{O}_p$ and $d = t$ versus $P_{\mathcal{H}}\mathcal{S}_{\mathbf{Y}|\mathbf{X}} \neq \mathcal{O}_p$ or $d > t$.
- Conditional predictor hypotheses: Given d, $P_{\mathcal{H}}\mathcal{S}_{\mathbf{Y}|\mathbf{X}} = \mathcal{O}_p$ versus $P_{\mathcal{H}}\mathcal{S}_{\mathbf{Y}|\mathbf{X}} \neq \mathcal{O}_p$.

These three hypotheses are discussed for FT-IRE, FT-RIRE, and FT-PIRE, respectively. FT-IRE and FT-RIRE mostly follow the idea of Cook and Ni (2005) and Ni and Cook (2007). The hypothesis tests for FT-PIRE have not been considered before. In the following, we will introduce their test statistics and asymptotic distributions.

FT-IRE Hypothesis Tests

Marginal Predictor Hypothesis for FT-IRE:(Cook and Ni, 2005)

The Marginal predictor hypothesis $P_{\mathcal{H}}\mathcal{S}_\xi = \mathcal{O}_p$ is equivalent to $H^T\xi = 0$, where $H \in \mathbb{R}^{p\times r}$ is a basis for $\mathcal{H}$. The following Wald test statistic of FT-IRE is to test $H^T\xi = 0$.

$$T(\mathcal{H}) = n\text{vec}(H^T\hat{\xi})^T[(I_{2m} \otimes H^T)\hat{\Gamma}(I_{2m} \otimes H)]^{-1}\text{vec}(H^T\hat{\xi})$$

asymptotically follows a chi-square random variable with $2rm$ degrees of freedom. In fact, because $\sqrt{n}[\text{vec}(\hat{\xi}) - \text{vec}(\xi)] \to N(0, \Gamma)$ (Theorem 1), we can have $\sqrt{n}[\text{vec}(H^T\hat{\xi}) - \text{vec}(H^T\xi)] \to N[0, (I_{2m} \otimes H^T)\hat{\Gamma}(I_{2m} \otimes H)]$ using Slutsky's theorem. Let $A = (I_{2m} \otimes H^T)\hat{\Gamma}(I_{2m} \otimes H)$, then the rank of $A^{1/2}A^{-1}A^{1/2}$ is $2rm$.

Joint Dimension Predictor Hypothesis for FT-IRE:(Cook and Ni, 2005)

The predictor part $H^T\xi = 0$ of a joint hypothesis is equivalent to the statement $\xi = Q_{\mathcal{H}}\xi$. The dimension part $d = t$ is considered using $\xi = Q_{\mathcal{H}}\beta v = H_0\beta_{H_0}v$, where $\beta \in \mathbb{R}^{p\times t}, v \in \mathbb{R}^{t\times 2m}$, and the coordinates $\beta_{H_0} \in \mathbb{R}^{(p-r)\times t}$ of β in terms of the basis H_0 for Span$(Q_{\mathcal{H}})$. By minimizing the constrained optimal discrepancy function for FT-IRE

$$F_{t,H}(B,C) = [\text{vec}(\hat{\xi}) - \text{vec}(H_0BC)]^T\hat{\Gamma}^{-1}[\text{vec}(\hat{\xi}) - \text{vec}(H_0BC)]$$

and the test statistics $n\hat{F}_{t,H}(B,C)$ is asymptotically distributed as a chi-square random variable with $(p-t)(2m-t)+tr$ degrees of freedom. Following Theorem 2, the Jacobian matrix for the constrained function is $\Delta_{\xi,H} = (I_{2m}\otimes H_0)(v^T\otimes I_{p-r}, I_{2m}\otimes\beta H_0) \in \mathbb{R}^{2pm\times t(p-r+2m)}$, so the degrees of freedom is $2pm - \text{r}ank(\Delta_{\xi,H}) = 2pm - t(p-r-t+2m) = (p-t)(2m-t)+tr$.

Conditional Predictor Hypothesis for FT-IRE:(Cook and Ni, 2005)

For the conditional predictor hypothesis, $P_{\mathcal{H}}\mathcal{S}_\xi = \mathcal{O}_p$ given d, the difference in minimum discrepancies for FT-IRE is employed, see following:

$$T(\mathcal{H}|d) = nF_{d,H}(B,C) - nF_d(B,C;\hat{\Gamma}^{-1}).$$

And the test statistics $\hat{T}(\mathcal{H}|d)$ is asymptotically distributed as a chi-square random variable with rd degrees of freedom under the null hypothesis. In fact, $T(\mathcal{H}|d)$ is asymptotically equivalent to $U^T(P_\xi - P_{\xi,H})U$, where $U \in \mathbb{R}^{2pm}$ is a standard normal random vector and P_ξ and $P_{\xi,H}$ are the projections with respect to the usual inner product onto Span$(\Gamma^{-1/2}\Delta)$ and Span$(\Gamma^{-1/2}\Delta_{\xi,H})$. It can be shown that Span$(\Delta_{\xi,H}) \subseteq$ Span(Δ), and thus Span$(\Gamma^{-1/2}\Delta_{\xi,H}) \subseteq$ Span$(\Gamma^{-1/2}\Delta)$. Then $(P_\xi - P_{\xi,H})$ is a projection with rank$(\Delta)-$ rank$(\Delta_{\xi,H}) = d(p-d+2m) - d(p-r-d+2m) = rd$.

FT-RIRE Hypothesis Tests

Marginal Predictor Hypothesis for FT-RIRE: (Ni and Cook, 2007)

If we use FT-RIRE and Theorem 5, the Wald-type statistic to test $H^T\xi = 0$ is

$$T_r(\mathcal{H}) = n\text{vec}(H^T\hat{\xi})^T[(I_{2m}\otimes H^T)\widetilde{G}(I_{2m}\otimes H)]^{-1}\text{vec}(H^T\hat{\xi}),$$

which is asymptotically distributed as a linear combination of $2pm$ independent chi-square random variables with one degree of freedom. Let H_0 be an orthogonal basis of Span($Q_{\mathcal{H}}$). The coefficients for the linear combination of chi-square random variables are the eigenvalues of $Q_m G^{-1/2}\Gamma G^{-1/2}Q_m$, where Q_m is the projection onto the complement of Span$[G^{-1/2}(I \otimes H_0)]$.

Joint Dimension Predictor Hypothesis for FT-RIRE:(Ni and Cook, 2007)

For a joint dimension predictor hypothesis, the constrained optimal discrepancy function for FT-RIRE is

$$F^r_{t,H}(B,C) = [\text{vec}(\hat{\xi}) - \text{vec}(H_0 BC)]^T \widetilde{G}^{-1}[\text{vec}(\hat{\xi}) - \text{vec}(H_0 BC)].$$

Under the joint null hypothesis, the test statistic $n\hat{F}^r_{t,H}(B,C)$ is asymptotically distributed as a linear combination of $2pm$ independent chi-square random variables with one degree of freedom. The coefficients are the eigenvalues of $Q_j G^{-1/2}\Gamma G^{-1/2}Q_j$, where Q_j is the projection onto the complement of Span$[G^{-1/2}(v^T \otimes H_0, I \otimes H_0\beta_{H_0})]$.

Conditional Predictor Hypothesis for FT-RIRE:(Ni and Cook, 2007)

For the conditional predictor hypothesis, the difference in minimum discrepancies for FT-RIRE is

$$T_r(\mathcal{H}|d) = nF^r_{d,H}(B,C) - nF_d(B,C;\widetilde{G}^{-1}),$$

and then $\hat{T}_r(\mathcal{H}|d)$ is asymptotically distributed as a linear combination of $2pm$ independent chi-square random variables with one degree of freedom, where the coefficients are the eigenvalues of $Q_c G^{-1/2}\Gamma G^{-1/2}Q_c$ and $Q_c = Q_j - Q_{G^{-1/2}\Delta}$.

FT-PIRE Hypothesis Tests

Marginal Predictor Hypothesis for FT-PIRE:

When the predictors have a categorical variable, we divide the data into K parts (K is the number of class for this categorical variable). Each class is independent of the other classes. Hence, we can apply FT-IRE hypothesis tests for each class (Cook and Ni, 2005). We use notation $\hat{\xi}_l$ and $\hat{\Gamma}_l^{-1}$ in Section 3.2 for each class. The Wald

test statistic to test hypothesis $H^T\xi = 0$, where $\xi = (\xi_1, \cdots, \xi_l)$, is

$$T^K(\mathcal{H}) = \sum_{l=1}^{K} T(\mathcal{H}) = \sum_{l=1}^{K} n_l \text{vec}(H^T\hat{\xi}_l)^T[(I_{2m} \otimes H^T)\hat{\Gamma}(I_{2m} \otimes H)]^{-1}\text{vec}(H^T\hat{\xi}_l),$$

which asymptotically follows a chi-square random variable with $2rmK$ degrees of freedom.

Joint Dimension Predictor Hypothesis for FT-PIRE:

The joint dimension predictor null hypothesis is $H^T\xi = 0$ and $d = t$. H_0 is the basis for $Q_{\mathcal{H}}$. The joint hypothesis by minimizing the constrained optimal discrepancy function for FT-PIRE is,

$$T_t^K(\mathcal{H}) = \sum_{l=1}^{K}[\text{vec}(\hat{\xi}_l) - \text{vec}(H_0BC_l)]^T\hat{\Gamma}_l^{-1}[\text{vec}(\hat{\xi}_l) - \text{vec}(H_0BC_l)]$$

and $\hat{T}_t^K(\mathcal{H})$ is distributed asymptotically as a chi-square random variable with $K(p-t)(2m-t) + Ktr$ degrees of freedom (Cook and Ni, 2005).

Conditional Predictor Hypothesis for FT-PIRE:

A conditional predictor hypothesis is $P_{\mathcal{H}}\mathcal{S}_\xi = \mathcal{O}_p$ given d, and then the difference in minimum discrepancies for FT-IRE is

$$T^K(\mathcal{H}|d) = T_d^K(\mathcal{H}) - \sum_{l=1}^{K} n_l F_d(B, C; \hat{\Gamma}_l^{-1}),$$

$\hat{T}^K(\mathcal{H}|d)$ is distributed asymptotically as a chi-square random variable with rdK degrees of freedom under the null hypothesis (Cook and Ni, 2005).

3.4 Sufficient Variable Selection

In the previous sections, the estimate of the central subspace was discussed in the situation with sample size larger than the number of predictors. In this section, we continue to develop our methods to deal with sparse sufficient variable selection especially for large p and small n data. Sufficient variable selection (SVS)(Yin and Hilafu, 2015), different from variable selection, is to find a subset of relevant variables but without losing any regression information, this is also why it is called *sufficient* variable selection. In the field of sufficient dimension reduction, SVS is devoted to

seeking a few sparse linear combinations to perform dimension reduction. There are five different estimators have been discussed, but we only explore two coordinate-independent sparse estimates for FT-IRE and FT-SIRE. Because other methods can easily follow FT-IRE.

Coordinate-independent Sparse Fourier Transform Inverse Regression Estimator (CIS-FTIRE)

Coordinate-independent sparse sufficient dimension reduction was introduced by Chen et al. (2010), using the weighted group lasso as penalty term which shrinks the variables for all dimension at the same time. We adopt the penalty term of Chen et al. (2010) and optimization algorithm from Qian et al. (2018), that is to simultaneously estimate β and ν with shrinkage using coordinate descent algorithm and Stiefel manifold optimization.

Let $B = (\boldsymbol{B}_1, ..., \boldsymbol{B}_p)^T$ and $\boldsymbol{B}_i \in \mathbb{R}^d$ be the i^{th} row of B. Simultaneous variable selection is to find a set $\mathcal{A}_0 = \{1 \leq j \leq p : \boldsymbol{e}_j^T \beta\beta^T \boldsymbol{e}_j > 0\}$. The number of $\mathcal{A}_0$, $|\mathcal{A}_0|$, is denoted as u, which indicates the number of important predictors. The coordinate-independent penalty is $p_{\boldsymbol{w}}(B) = \sum_{j=1}^{p} w_j ||\boldsymbol{B}_j||_2$, and $\boldsymbol{w} = (w_1, ..., w_p)$ are the penalty weights. We define $\Upsilon_n = \Sigma\hat{\xi}$. The discrepancy function can be written as:

$$\begin{aligned} F(B,C) &= [\text{vec}(\hat{\xi}) - \text{vec}(BC)]^T \Gamma^{-1} [\text{vec}(\hat{\xi}) - \text{vec}(BC)] \\ &= [\text{vec}(\Upsilon_n) - \text{vec}(\Sigma BC)]^T (I_{2m} \otimes \Sigma^{-1}) \Gamma^{-1} (I_{2m} \otimes \Sigma^{-1}) [\text{vec}(\Upsilon_n) - \text{vec}(\Sigma BC)] \\ &= [\text{vec}(\Upsilon_n) - \text{vec}(\Sigma BC)]^T \Lambda [\text{vec}(\Upsilon_n) - \text{vec}(\Sigma BC)], \end{aligned} \tag{3.5}$$

where $\Lambda = \text{Cov}\{\text{vec}[(\mathbf{X} - \boldsymbol{\mu})\boldsymbol{\epsilon}^T]\}^{-1}$.

We combine this coordinate-independent penalty with the quadratic discrepancy function (3.5) with tuning parameter λ, the object function is the following:

$$\begin{aligned} & L_n(B,C) \\ = \; & \tfrac{1}{2}[\text{vec}(\Upsilon_n) - \text{vec}(\Sigma BC)]^T \Lambda [\text{vec}(\Upsilon_n) - \text{vec}(\Sigma BC)] \\ & + \lambda p_{\boldsymbol{w}}(B), \text{ subject to } CC^T = I_d. \end{aligned} \tag{3.6}$$

Let $(\hat{B}, \hat{C}) = \arg\min_{B,C} L_n(B,C)$. The estimate of $\mathcal{S}_{\mathbf{Y}|\mathbf{X}}$ is $\text{Span}(\hat{B})$ and the estimate of $\mathcal{A}_0$ is $\hat{\mathcal{A}}_0 = \{1 \leq j \leq p : \boldsymbol{e}_j^T \hat{B}\hat{B}^T \boldsymbol{e}_j > 0\}$. Before stating the algorithm for estimate,

let us look at the Λ,

$$\begin{aligned}
\Lambda^{-1} &= \mathrm{Cov}\{\mathrm{vec}[(\mathbf{X}-\boldsymbol{\mu})\boldsymbol{\epsilon}^T]\} &&= \mathrm{Cov}[\mathrm{vec}(\Sigma^{1/2}\mathbf{Z}\boldsymbol{\epsilon}^T)] \\
&= \mathrm{Cov}[(I_{2m}\otimes\Sigma^{1/2})\mathrm{vec}(\mathbf{Z}\boldsymbol{\epsilon}^T)] &&= (I_{2m}\otimes\Sigma^{1/2})\mathrm{Cov}[\mathrm{vec}(\mathbf{Z}\boldsymbol{\epsilon}^T)](I_{2m}\otimes\Sigma^{1/2}) \\
&= (I_{2m}\otimes\Sigma^{1/2})\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\otimes\mathbf{Z}\mathbf{Z}^T)(I_{2m}\otimes\Sigma^{1/2})
\end{aligned}$$

Each component of $\boldsymbol{\epsilon}$ can be regarded as the real and imaginary parts of the population residual of $e^{i\boldsymbol{\omega}^T\mathbf{Y}}$ regression on $\mathbf{Z}$. It is easy to prove that $\boldsymbol{\epsilon}$ is uncorrelated with $\mathbf{Z}$ and $\mathrm{E}(\boldsymbol{\epsilon})=0$. If we assume that $\boldsymbol{\epsilon}$ is independent of $\mathbf{Z}$, then $\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\otimes\mathbf{Z}\mathbf{Z}^T) = \mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)\otimes I_p$. Hence, $\Lambda = (I_{2m}\otimes\Sigma^{-1/2})[\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)^{-1}\otimes I_p](I_{2m}\otimes\Sigma^{-1/2}) = \mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)^{-1}\otimes\Sigma^{-1}$.

Lemma 2. *If A and B are both in $\mathbb{R}^{p\times d}$, then* $\mathrm{vec}(A)^T\mathrm{vec}(B) = \mathrm{vec}(B^T)^T\mathrm{vec}(A^T)$.

Lemma (2) is easy to be proved by using definition of Vectorization. Define

$$\begin{aligned}
& U(B,C) \\
=\ & \tfrac{\partial F(B,C)}{2\partial\mathrm{vec}(B^T)} \\
=\ & \tfrac{\partial}{2\partial\mathrm{vec}(B^T)}[-2\mathrm{vec}(\Sigma BC)^T\Lambda\mathrm{vec}(\Upsilon_n)+\mathrm{vec}(\Sigma BC)^T\Lambda\mathrm{vec}(\Sigma BC)] \\
=\ & \tfrac{\partial}{2\partial\mathrm{vec}(B^T)}[-2\mathrm{vec}(B)^T(C\otimes\Sigma)\Lambda\mathrm{vec}(\Upsilon_n)+\mathrm{vec}(B)^T(C\otimes\Sigma)\Lambda(C^T\otimes\Sigma)\mathrm{vec}(B)] \\
=\ & \tfrac{\partial}{2\partial\mathrm{vec}(B^T)}\{-2\mathrm{vec}(B)^T[C\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)^{-1}\otimes I_p]\mathrm{vec}(\Upsilon_n)+\mathrm{vec}(B)^T[C\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)^{-1}C^T\otimes\Sigma]\mathrm{vec}(B)\} \\
=\ & \tfrac{\partial}{2\partial\mathrm{vec}(B^T)}\{-2\mathrm{vec}(B)^T\mathrm{vec}(\Upsilon_n\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)^{-1}C^T)+\mathrm{vec}(B)^T\mathrm{vec}(\Sigma BC\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)^{-1}C^T)\} \\
=\ & \tfrac{\partial}{2\partial\mathrm{vec}(B^T)}\{-2\mathrm{vec}[C\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)^{-1}\Upsilon_n^T]^T\mathrm{vec}(B^T)+\mathrm{vec}[C\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)^{-1}C^TB^T\Sigma]^T\mathrm{vec}(B^T)\} \\
=\ & -\mathrm{vec}[C\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)^{-1}\Upsilon_n^T]+\mathrm{vec}[C\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)^{-1}C^TB^T\Sigma] \\
=\ & -[I_p\otimes C\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)^{-1}\Upsilon_n^T]\mathrm{vec}(I_p)+[\Sigma\otimes C\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)^{-1}C^T]\mathrm{vec}(B^T),
\end{aligned}$$

and

$$H \;=\; \tfrac{\partial^2 F(B,C)}{2\partial^2\mathrm{vec}(B^T)} \;=\; \Sigma\otimes C\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)^{-1}C^T.$$

We notice that neither $U(B,C)$ nor H depend on the inverse of Σ, but it needs to know the inverse of $\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)$. And, $\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)$ is a covariance matrix, so $\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)\geq 0$.

However, $\cos(\boldsymbol{\omega}_i^T\mathbf{Y})$ and $\cos(\boldsymbol{\omega}_j^T\mathbf{Y})$ or $\sin(\boldsymbol{\omega}_i^T\mathbf{Y})$ and $\sin(\boldsymbol{\omega}_j^T\mathbf{Y})$ are highly correlated for the different values $\boldsymbol{\omega}_i$ and $\boldsymbol{\omega}_j$. Hence, we further assume that there exists a constant $c_0>0$ such that $\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) > c_0 I_{2m}$. More generally, we assume $\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\otimes\mathbf{Z}\mathbf{Z}^T) > c_0 I_{2mp}$, without assuming the independence of $\boldsymbol{\epsilon}$ and $\mathbf{Z}$. Then,

$\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)^{-1} < \frac{1}{c_0} I_{2m}$ or $\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T \otimes \mathbf{Z}\mathbf{Z}^T)^{-1} < \frac{1}{c_0} I_{2mp}$. $\frac{1}{c_0}$ is estimated using the maximum of the diagonal elements in $\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)^{-1}$.

Let $\tilde{h} = \frac{\lambda_{\max}(\Sigma)}{\hat{c}_0}$ and $\tilde{H} = \tilde{h} I_p \otimes I_d$. At the $(t+1)$-th iteration, $(B_{(t)}, C_{(t)})$ is the estimate of (B, C) after the t-th iteration, then we update B using the quadratic approximation of L_n in (3.6) by

$$L_n^{(t)}(B) = U_t^T(\mathrm{vec}(B^T) - \mathrm{vec}(B_{(t)}^T)) + \frac{\tilde{h}}{2}(\mathrm{vec}(B^T) - \mathrm{vec}(B_{(t)}^T))^T(\mathrm{vec}(B^T) - \mathrm{vec}(B_{(t)}^T)) + \lambda p_{\boldsymbol{w}}(B),$$

where $U_t = U(B_{(t)}, C_{(t)})$. So $B_{(t+1)} = \mathrm{argmin}_B L_n^{(t)}(B)$. By the Karush-Kuhn-Tucker condition, the l-th row of $B_{(t+1)}$ has the form

$$\boldsymbol{B}_l^{(t+1)} = \frac{1}{\tilde{h}}(1 - \frac{\lambda w_l}{||\tilde{h}\boldsymbol{B}_l^{(t)} - U_l^{(t)}||_2})_+(\tilde{h}\boldsymbol{B}_l^{(t)} - U_l^{(t)}),$$

where $z_+ = \max(z, 0)$, and $U_l^{(t)}$ is the l-th row of U_t after reforming as $p \times d$ matrix. Next, we fix $B = B_{(t+1)}$, and update C using the Reduced Rank Procrustes Rotation (Zou et al., 2006) to solve:

$$C_{(t+1)} \;=\; \arg\min_C - \mathrm{trace}[\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)^{-1}\Upsilon_n^T B_{(t+1)} C].$$

Then, $C_{(t+1)} = W_2 W_1^T$, where $W_1 D W_2^T$ is the singular value decomposition of $\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)^{-1}\Upsilon_n^T B_{(t+1)}$.

Coordinate-independent Sparse Fourier Transform Special Inverse Regression Estimator (CIS-FTSIRE)

Previously, we have developed the coordinate-independent sparse for the optimal inner product matrix Λ. If we assume that $\boldsymbol{\epsilon}$ is independent of $\mathbf{Z}$ and each component of $\boldsymbol{\epsilon}$ has variance 1 and is independent of each other, then $\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T \otimes \mathbf{Z}\mathbf{Z}^T) = I_{2mp}$ and $\Lambda = I_{2m} \otimes \Sigma^{-1}$, in which (3.5) reduce to the discrepancy function of FT-SIRE with the diagonal block inner product matrix. We state algorithm using the quadratic discrepancy function of FT-SIRE, rewrite the quadratic function,

$$\begin{aligned} F_d(B, C; \mathrm{diag}\{\Sigma\}) &= [\mathrm{vec}(\hat{\xi}) - \mathrm{vec}(BC)]^T \mathrm{diag}\{\Sigma\}[\mathrm{vec}(\hat{\xi}) - \mathrm{vec}(BC)] \\ &= [\mathrm{vec}(\hat{\xi}) - \mathrm{vec}(BC)]^T (I_{2m} \otimes \Sigma)[\mathrm{vec}(\hat{\xi}) - \mathrm{vec}(BC)] \\ &= \mathrm{trace}[(\hat{\xi} - BC)^T \Sigma (\hat{\xi} - BC)] \\ &= \mathrm{trace}[(\Sigma\hat{\xi} - \Sigma BC)^T \Sigma^{-1} (\Sigma\hat{\xi} - \Sigma BC)], \end{aligned}$$

and combine this coordinate-independent penalty with the quadratic discrepancy function (3.3) with tuning parameter λ, under the constraint $CC^T = I_d$,

$$L_n(B,C) = \text{trace}[(\Upsilon_n - \Sigma BC)^T \Sigma^{-1}(\Upsilon_n - \Sigma BC)] + \lambda \sum_{j=1}^{p} w_j ||\boldsymbol{B}_j||_2, \text{ subject to } CC^T = I_d. \tag{3.7}$$

The derivative of $L_n(B,C)$ respect to $\text{vec}(B^T)$ is $U(B,C) = \frac{\partial L_n(B,C)}{2\partial \text{vec}(B^T)} = -[I_p \otimes (C\Upsilon_n^T)]\text{vec}(I_p) + (\Sigma \otimes I_d)\text{vec}(B^T)$. The second derivative is $H = \Sigma \otimes I_d$. Let $\tilde{h} = \lambda_{\max}(\Sigma)$, and $\hat{H} = \tilde{h} I_p \otimes I_d$. We update B using the following:

$$\begin{aligned} L_n^{(t)}(B) &= U_t^T(\text{vec}(B^T) - \text{vec}(B_{(t)}^T)) \\ &\quad + \tfrac{\tilde{h}}{2}(\text{vec}(B^T) - \text{vec}(B_{(t)}^T))^T(\text{vec}(B^T) - \text{vec}(B_{(t)}^T)) \\ &\quad + \lambda p_{\boldsymbol{w}}(B), \text{ subject to } CC^T = I_d, \end{aligned}$$

where $U_t = U(B_{(t)}, C_{(t)})$.

$$\boldsymbol{B}_l^{(t+1)} = \tfrac{1}{\tilde{h}}\left(1 - \frac{\lambda w_l}{||\tilde{h}\boldsymbol{B}_l^{(t)} + C_{(t)}\Upsilon_n^T \boldsymbol{e}_l - \sum_{k=1}^p \sigma_{lk}\boldsymbol{B}_k^{(t)}||_2}\right)_+ (\tilde{h}\boldsymbol{B}_l^{(t)} + C_{(t)}\Upsilon_n^T \boldsymbol{e}_l - \textstyle\sum_{k=1}^p \sigma_{lk}\boldsymbol{B}_k^{(t)}),$$

where $\sigma_{lk} = (\Sigma)_{lk}$. With fixed $B = B_{(t+1)}$, we update C using $C_{(t+1)} = W_2 W_1^T$, where $W_1 D W_2^T$ is the singular value decomposition of $\Upsilon_n^T B_{(t+1)}$.

Tuning Parameters and Algorithm

We follow Qian et al. (2018)'s algorithm and their choices of tuning parameters. Assume d is known, and we only use the equal weight instead of updating it as in Qian et al. (2018), for computational efficiency by avoiding a second cross validation.

The proposed method involves two tuning parameters: the weight $\boldsymbol{w}$ and the penalization λ. We use the candidate set of λ as a sequence with length N_λ: $\{\exp\left(\frac{N_\lambda - j}{N_\lambda - 1}\log\lambda_{min} + \frac{j-1}{N_\lambda - 1}\log\lambda_{max}\right)\}_{j=1}^{N_\lambda}$, where the minimum and maximum of candidates are λ_{min} and λ_{max}. By doing this, the sequence of λ has more values around λ_{min} than λ_{max}. The larger the λ, the more penalty putting on the B.

Here we use cross validation to choose the best λ. Divide data into K folds. Let $\mathbf{y}^{(-k)}$ and $\mathbf{x}^{(-k)}$ to be the response vector and predictor matrix after excluding the k-th fold, where $k = 1, ..., K$. In our simulation, we choose $K = 5$. If we don't

know d, then we can use sequential hypothesis tests to estimate d before applying the following algorithm.

1. Given d and a candidate penalization $\tilde{\lambda}$, which is from the potential set, calculate the B with data $(\mathbf{y}^{(-k)}, \mathbf{x}^{(-k)})$ and equal weights $\boldsymbol{w} = \mathbf{1}_p$. Denote the solutions by $\tilde{B}$, and construct the predictor variables with reduced dimensions by $\mathbf{z}_{\tilde{\lambda},k} = \mathbf{x}^{(-k)}\tilde{B}$.

2. Calculate the distance correlation (Li et al., 2012) bewteen $\mathbf{y}^{(-k)}$ and $\mathbf{z}_{\tilde{\lambda},k}$: $\text{dcor}(\mathbf{y}^{(-k)}, \mathbf{z}_{\tilde{\lambda},k})$, and compute the sum of $1 - \text{dcor}(\mathbf{y}^{(-k)}, \mathbf{z}_{\tilde{\lambda},k})$ over K folds. We choose $\hat{\lambda}$ that minimize the $L(\tilde{\lambda}) = \sum_{k=1}^{K} 1 - \text{dcor}(\mathbf{y}^{(-k)}, \mathbf{z}_{\tilde{\lambda},k})$, and calculate $\hat{B}$ using $\hat{\lambda}$ and data $(\mathbf{y}, \mathbf{x})$.

In order to estimate B and C, initial values for B_0 and C_0, are needed in the algorithm. The initial values not only affect the accuracy but also the computation efficiency.

- For B_0: given d, we employ the LassoSIR (Lin et al., 2016) on $(\mathbf{X}, \mathbf{Y})$ to obtain the first direction β_1^0. Then to use the LassoSIR again on $(Q_{\beta_1^0}\mathbf{X}, \mathbf{Y})$ to achieve the second direction β_2^0, keep doing this until β_d^0 is obtained.

- C_0: given B_0, update C_0.

Consistency and Oracle

We further develop the consistency property and oracle property of the sparse estimation, similar to the Qian et al. (2018). Let $\Upsilon_j = \text{E}(e^{i\boldsymbol{\omega}_j^T\mathbf{Y}}\mathbf{X}) - \text{E}(e^{i\boldsymbol{\omega}_j^T\mathbf{Y}})\text{E}(\mathbf{X})$ for $j = 1, ..., m$ and $\Upsilon = (\Upsilon_1^R, \Upsilon_1^I, ..., \Upsilon_m^R, \Upsilon_m^I)$. Note that, we have $\Upsilon = \Sigma\xi$. Because $\xi = \beta\nu$, so $\Upsilon = \Sigma\beta\nu$, then $\hat{\Upsilon}$, the sample version of Υ, can be obtained by replacing the expectation with sample means. As stated above algorithm, $\hat{\Upsilon} = \hat{\Sigma}\hat{\xi} = (\hat{\Upsilon}_1^R, \hat{\Upsilon}_1^I, ..., \hat{\Upsilon}_m^R, \hat{\Upsilon}_m^I)$, where $\hat{\Upsilon}_j = \frac{1}{n}\sum_{k=1}^{n} e^{i\boldsymbol{\omega}_j^T\mathbf{y}_k}\mathbf{x}_k - \frac{1}{n}\sum_{k=1}^{n} e^{i\boldsymbol{\omega}_j^T\mathbf{y}_k}\bar{\mathbf{x}}$. Using CIS-FTSIRE, we don't need to know the inverse of Σ, but Σ is required in estimating the first and second derivative of $F(B, C)$. One choice is to use sample covariance, alternately the threshold covariance is preferable in the high-dimension

setting. See more discussion in Qian et al. (2018). The number of Fourier transform m, the structural dimension d, and the number of important predictors u are allowed to diverge with sample size n. Let $\mathbf{X} = (X_1, ..., X_p)^T$, following conditions are assumed:

C1: For all $\epsilon > 0$ and $1 \leq j \leq p$, there is a constant $C > 0$ such that $P(|X_j - \mu_j| > \epsilon) \leq 2\exp(-\epsilon^2/2C)$.

C2: There are constants $\sigma_l, \sigma_u, \sigma_* > 0$ such that $\sigma_{ij} < \sigma_u$ for every $1 \leq i, j \leq p$, $\lambda_{min}(\mathrm{E}^{-1}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) \otimes I_p) > \sigma_*$, and $\lambda_{min}(\Sigma) > \sigma_l$, where λ_{min} is the minimum eigenvalue of the matrix.

C3: Assume $m^2 u \log p_n = O(n^{1-2\eta})$ and $du^2 \log p_n = O(n^{1-2\eta})$ for some constant $0 < \eta < 1/2$.

C4: Assume $\min_{j \in \mathcal{A}_0} \boldsymbol{e}_j^T \beta\beta^T \boldsymbol{e}_j > C_\phi n^{-\phi}$ for some $0 \leq \phi < 2\eta$ and constant C_ϕ.

C5: The nonzero singular values of β are bounded away from 0.

Condition C1 requires $\mathbf{X}$ to follow uniformly Sub-Gaussian which makes sure that the probability of extreme values $\mathbf{X}$ is small. C2 makes sure that Σ and $\mathrm{E}^{-1}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)$ have uniformly upper bound. Although m, d, and u can diverge with n, condition C3 requires that they do not diverge too fast. Condition C4 means the norm of each important variable is not less than $C_\phi n^{-\phi}$, indicating their signal is large enough to be detected. Condition C5 is needed in Theorem 9 below. We estimate Σ using the sample covariance matrix $\hat{\Sigma} = (\hat{\sigma}_{ij})_{1 \leq i,j \leq p}$ and define $p_n = \max(p, n)$, and $||\cdot||_F$ denotes the Frobenius norm.

Theorem 9. *Under conditions C1-C5, the minimizer $\hat{B}, \hat{C}$ of (3.6) with the sample covariance satisfies*

1. $||P_{\mathcal{S}_{\hat{B}}} - P_{\mathcal{S}_{\mathbf{Y}|\mathbf{X}}}||_F = O_p((m + (du)^{1/2})\sqrt{u \log p_n / n})$;
2. $P(\hat{\mathcal{A}}_0 = \mathcal{A}_0) \to 1$ *as* $n \to \infty$.

Here, tuning parameter $\hat{\lambda}$ and λ are given in the Appendix.

When dimension p is much larger than sample size n, it is difficult to estimate the covariance matrix. Thresholding method (Bickel and Levina, 2008) is computationally fast. Rothman et al. (2009) discussed the generalized thresholding methods for sparse covariance matrix. There are various choices for the thresholding rules such as lasso (Tibshirani, 1996), adaptive lasso (Zou, 2006), SCAD (Fan and Li, 2001), and hard-thresholding (Rothman et al., 2009). We also use the lasso soft-thresholding rule as in Qian et al. (2018): $s_\tau(z) = \text{sign}(z)(|z| - \tau)_+$. The thresholding estimate of $\widetilde{\Sigma} = (\tilde{\sigma}_{ij})_{1\le i,j\le p}$, where $\tilde{\sigma}_{ij} = \hat{\sigma}_{ij}$, if $i = j$, otherwise $\tilde{\sigma}_{ij} = s_\tau(\hat{\sigma}_{ij})$. There are two conditions for estimating sparse covariance matrix:

C6: The covariance matrix Σ satisfies that $\max_{1\le i\le p}\sum_{j=1}^{p}|\sigma_{ij}|^\kappa$ is upper bounded with $0 \le \kappa < 1$.

C7: Assume $m^2 u \log p_n = O(n^{1-2\eta})$ and $u^2 \log p_n = O(n^{1-2\eta})$ for some constant $0 < \eta < 1/2$. Also assume either $du^2 \log p_n = O(n^{1-2\eta})$ or $u(\log p_n)^{1-\kappa} = O(n^{1-\kappa-2\eta})$ holds.

Condition C6 assumes the sparsity of Σ. Similarly, condition C7 restricts the rate of divergence for m and u to ensure the estimate consistent. The following is the same result for thresholding covariance.

Theorem 10. *Under conditions C1-C2 and C4-C7, the minimizer $\hat{B}, \hat{C}$ of (3.6) with the thresholding covariance matrix satisfies*

1. $||P_{\mathcal{S}_{\hat{B}}} - P_{\mathcal{S}_{\mathbf{\Upsilon}|\mathbf{X}}}||_F = O_p((m + \sqrt{(du) \wedge l})\sqrt{u \log p_n / n})$, *where* $l = (n/\log p_n)^\kappa$;
2. $P(\hat{\mathcal{A}}_0 = \mathcal{A}_0) \to 1$ *as* $n \to \infty$.

Here, tuning parameter $\hat{\lambda}$ and λ are given in the Appendix.

3.5 Numerical Study

Suppose that both B and $\hat{B}$ are $p \times d$ orthogonal matrices. We use trace correlation (r_2) (Ye and Weiss, 2003) for measuring accuracy, which is defined as $r_2 = \sqrt{\sum_{i=1}^{d}\rho_i^2/d}$, where ρ_i are the eigenvalues of matrix $\hat{B}^T B B^T \hat{B}$ for $i = 1, ..., d$. Note

that r_2 measures the similarity of two matrices. If $\hat{B} = (\hat{B}_1, ..., \hat{B}_d)$ is the estimate of matrix B, the larger the r_2, the more accurate the estimator.

We use two criteria to compare SVS methods: 1) The true positive (TP) is the number of correctly identified active predictors, and 2) the false positive (FP) is the number of falsely identified active predictors. The bigger the TP and smaller the FP, the better the estimates.

In the following simulations, univariate and multivariate response models (Model 7 and Model 8) are used to demonstrate the estimate performance for FT-IRE, FT-DIRE, FT-SIRE, FT-RIRE, and FT-DRIRE. Model 9 compares FT-IRE, FT-DIRE, and FT-SIRE with FIRE, DIRE (Cook and Zhang, 2014), and IRE (Cook and Ni, 2005). Model 10 compares SVS strategies: TC-SIR (Qian et al., 2018), CIS-FTIRE and CIS-FTSIRE. Model 11 employs tests of dimensions for FT-IRE, FT-DIRE, FT-RIRE, and FT-DRIRE. Model 12 conducts marginal and conditional predictor hypotheses. Finally, a PDB dataset is analyzed using FT-SIRE.

Simulations

Model 7. $Y = X_1 + 0.5X_2^2$, with $p = 5$, $n = 800$ and $d = 2$. Predictors $X_1, X_3, X_5 \overset{iid}{\sim} N(0, 1)$, and $X_2 = X_1 + Z$ where $Z \sim N(0, 1)$ and $X_4 = (1 + X_2)Z$. Let $\{e_i\}$ be $p \times 1$ vectors whose i^{th} entry is 1 and other entries are 0. Then $B = (e_1, e_2)$.

The goal of this simulation is to demonstrate the performance of FT-IRE, FT-DIRE, FT-SIRE, FT-RIRE, and FT-DRIRE for different numbers of $\boldsymbol{\omega}$ in the univariate response. Figure 3.1 plots mean values of r_2 over the 100 simulated data vs. different sizes of $\boldsymbol{\omega}$: $\{5, 10, 15, ..., 100\}$. All five methods have high mean trace correlation: above 0.96. It is hard to distinguish which method performs the best, but most of them fluctuate around 0.98. However, FT-SIRE is slightly lower than the other methods. The larger the number of $\boldsymbol{\omega}$ is, the more stable the mean trace correlation is. Estimates of the five methods keep the same magnitude for different numbers of $\boldsymbol{\omega}$. The estimates are accurate and stable as long as the number of $\boldsymbol{\omega}$ is sufficiently large (Here, say, bigger than 25).

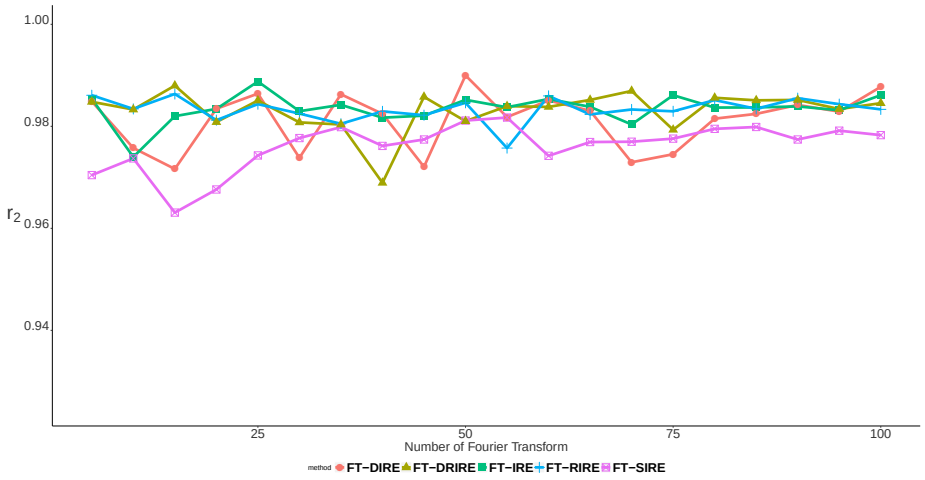

Figure 3.1: Mean values of r_2 over the 100 simulated data vs. different sizes of $\boldsymbol{\omega}$: $\{5, 10, 15, ..., 100\}$ in Model 7.

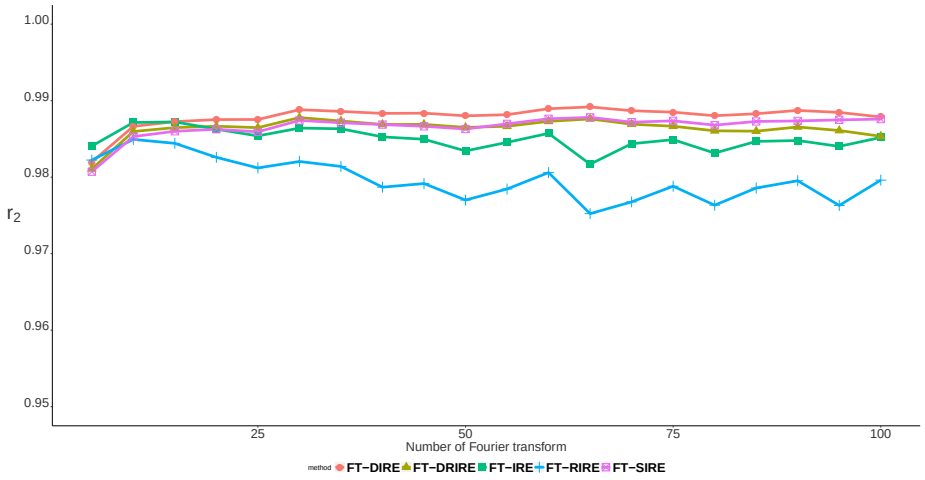

Figure 3.2: Mean values of r_2 over the 100 simulated data vs. different sizes of $\boldsymbol{\omega}$: $\{5, 10, 15, ..., 100\}$ in Model 8.

Model 8. This is Example 3 of Zhu et al. (2010c). $Y_1 = 1 + \beta_1^T\mathbf{X} + \sin(\beta_2^T\mathbf{X}) + \epsilon_1$, $Y_2 = \frac{\beta_2^T\mathbf{X}}{0.5+(\beta_1^T\mathbf{X}+1)^2} + \epsilon_2$, $Y_3 = |\beta_1^T\mathbf{X}|\epsilon_3$, $Y_4 = \epsilon_4, ..., Y_q = \epsilon_q$, with $p = 20$, $q = 5$, $d = 2$, $\beta_1 = e_1$, and $\beta_2 = e_2 + e_3$. Predictor $\mathbf{X}_i \sim N(0, I)$, $n = 2000$ and $\boldsymbol{\epsilon}_i = (\epsilon_1, \epsilon_2, ..., \epsilon_q)^T \sim N_q(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix}$, $A = \begin{pmatrix} 1 & -1/2 \\ -1/2 & 1/2 \end{pmatrix}$, and $D = \text{diag}(1/2, 1/3, ..., 1/q)$.

This is a multivariate responses example exploring the performance of FT-IRE, FT-DIRE, FT-SIRE, FT-RIRE, and FT-DRIRE vs. different sizes of $\boldsymbol{\omega}$. Figure

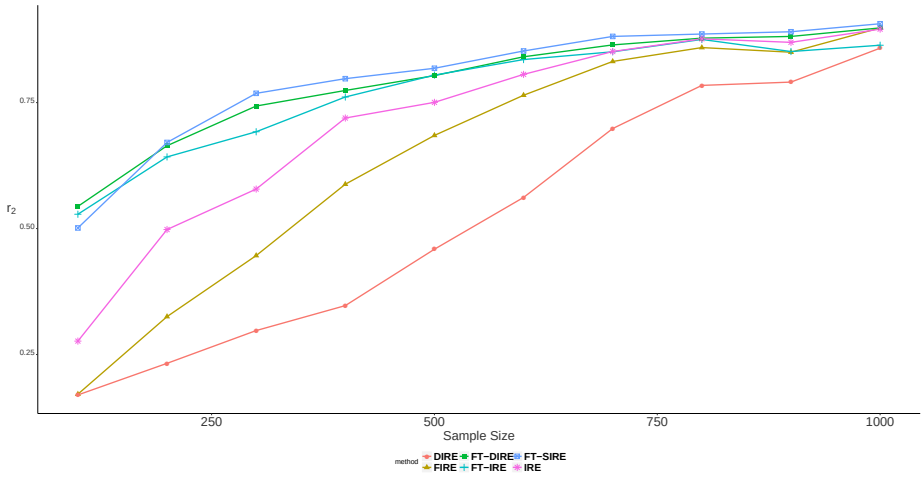

Figure 3.3: Mean values of r_2 over the 100 simulated data vs. various sample sizes from 100 to 1000 at increments of 100 in Model 9.

3.2 plots mean values of r_2 over the 100 simulated data vs. different sizes of $\boldsymbol{\omega}$: $\{5, 10, 15, ..., 100\}$. The mean trace correlations are about 0.98 for all five methods, thus indicating the estimates are accurate and stable as the number of $\boldsymbol{\omega}$ changes. Compared to Model 7, the fluctuation of r_2 for the multivariate model is smaller because multivariate responses provide more information than univariate response. The mean r_2 values of FT-RIRE are lower than other methods. In addition, the degenerated methods (FT-DIRE and FT-DRIRE) have better performance than the general cases (FT-IRE and FT-RIRE). FT-SIRE performs as well as FT-DRIRE.

Model 9. $Y = |\mathbf{X}^T\beta| + 0.2\epsilon$, with $p = 10$, $d = 1$, and $\beta = (1, 1, 1, 1, 0, \cdots, 0)^T \in \mathbb{R}^p$. Predictor $\mathbf{X}$ is generated by $N(\boldsymbol{\mu}_j, \Sigma), j = 1, \cdots, p$ with probability $\frac{1}{p}$ each, where $\boldsymbol{\mu}_j \in \mathbb{R}^p$ is a p-dimensional predictor with j^{th} element two and the other elements zeros, and Σ is a positive definite matrix with $(j_1, j_2)th$ entry $0.5^{|j_1 - j_2|}$. Data with various sample sizes from 100 to 1000 at increments of 100 are simulated. The number of $\boldsymbol{\omega}$ is 50 for all simulated data because models 7 and 8 show that estimates are stable at this size.

Model 9 is compared with FIRE and DIRE (Cook and Zhang, 2014), and IRE (Cook and Ni, 2005), as well as their corresponding robust versions. For FIRE and DIRE, the fused slices is $H = \{3, 4, ..., 15\}$. For IRE, the slice value is $h = 5$.

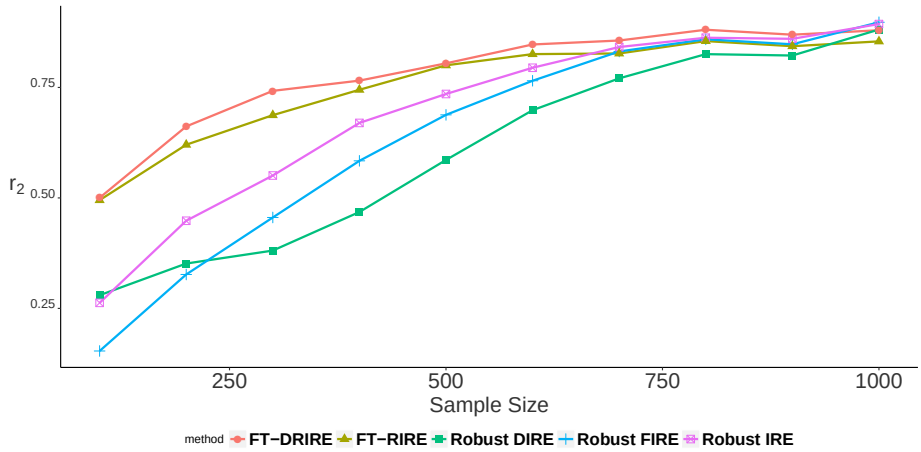

Figure 3.4: Mean values of r_2 over the 100 simulated data vs. various sample sizes from 100 to 1000 at increments of 100 using Robust version in Model 9.

Figure 3.3 compares FT-IRE, FT-DIRE, and FT-SIRE with FIRE, DIRE, and IRE. Overall, Fourier transform approaches have a higher r_2 than FIRE, DIRE, and IRE. The larger the sample size is, the better the estimate for all methods is, indicating the asymptotic efficiency of their estimates. Figure 3.4 also shows similar comparison results between FT-RIRE, FT-DRIRE with robust FIRE, robust DIRE (Cook and Zhang, 2014), and robust IRE (Ni and Cook, 2007). It is interesting to note that the degenerated Fourier transform approaches (FT-DIRE and FT-DRIRE) have larger r_2 compared to the general approaches (FT-IRE and FT-RIRE). However, FIRE and robust FIRE perform better than DIRE and robust DIRE, respectively.

Model 10. $Y = \sin(\mathbf{X}^T\beta)^2 + \mathbf{X}^T\beta + \epsilon$, with $d = 1$ and $\beta = (1, 1, 1, 1, 1, 0, \cdots, 0)^T \in \mathbb{R}^p$. Predictor $\mathbf{X}$ is generated from multivariate normal distribution with mean zeros and covariance matrix Σ, where Σ is: 1) a positive definite matrix with $(j_1, j_2)th$ entry $0.5^{|j_1 - j_2|}$ or 2) a block-diagonal structure with ten predictors into a block, where $\sigma_{ij} = \sigma_{ji} = 0.5$ if i and j are in the same block and 0s otherwise. The sample size is 200 and the dimension of $\mathbf{X}$ is 1000. Again, the number of $\boldsymbol{\omega}$ is 50. Mean values of trace correlation, Frobenius norm of difference between true β and estimate $\hat{\beta}$, the mean number of TP and FP over the 100 simulated data are calculated to compare methods. We compare our two methods CIS-FTIRE and CIS-FTSIRE with the thresholding covariance SIR (TC-SIR) (Qian et al., 2018).

Table 3.1: Compare CIS approaches for SIR, FT-IRE and FT-SIRE in Model 10.

Σ	Methods	R_2	Norm	TP	FP
1)	TC-SIR	0.9643	0.2533	4.99	3.68
	CIS-FTIRE	0.9815	0.1823	5	2.17
	CIS-FTSIRE	0.9637	0.2546	5	3.27
2)	TC-SIR	0.9606	0.2684	5	4.13
	CIS-FTIRE	0.9798	0.1889	5	2.53
	CIS-FTSIRE	0.9566	0.2757	5	3.2

From Table 3.1, CIS-FTIRE is outperforming the other two methods. Not only the estimate is accurate, but also TP numbers are exactly five and FP numbers are the smallest. Even though CIS-FTSIRE provide a slightly less accurate estimate, both of TP and FP numbers are better than TC-SIR. Overall, CIS-FTIRE and CIS-FTSIRE are better than TC-SIR in this model.

Model 11. This is the same model as Model 8 with $q = 10$ by adding extra five independent responses, $Y_q \sim N(0, 1/q)$ for $q = 6 \cdots, 10$ to test dimensions. Because it is a multivariate model, SIR, IRE, FIRE and DIRE cannot be directly employed. Hence, we only investigate the performance of FT-IRE, FT-DIRE, FT-SIRE, FT-RIRE, and FT-DRIRE. Again, m is 50.

Table 3.2 reports the percentages of correctly detecting dimensions ($d = 2$) among 100 simulated data for different sample size: $\{100, 150, 200, 400, 700\}$. In Table 3.2, all these methods have higher percentages of correctly detecting dimension in the larger sample size. First, FT-SIRE converges much quickly compared to the other four methods. However, it attains its stable points at 94% and 96% when the sample size is higher than 200. Because FT-SIRE has the simplest structure of the inner product matrix V, it is computationally cheaper. But it sacrifices the accuracy of testing dimensions. On the other hand, FT-RIRE performs worst with the smaller convergence rate when n from 100 to 200. But the percentage of correct dimensions reach 100% when the sample size is 400. If the sample size is too small, the robust estimate is not accurate, which affects the testing results. When the sample size is large enough, the percentages jump to higher values than other methods. For the

other three methods: FT-IRE, FT-DIRE, and FT-DRIRE, they all perform well when the sample size reaches to 200. Their convergence rates are in the middle of the five methods. In general, all these five methods provide competitive results for sample size larger than 200. If the sample size is too small, we recommend using FT-SIRE.

Table 3.2: Percentage of Correct Dimensions in Model 11.

n	100	150	200	400	700
FT-IRE	0.52	0.85	0.95	0.99	1.00
FT-DIRE	0.53	0.80	0.91	0.94	0.97
FT-SIRE	0.74	0.91	0.91	0.96	0.94
FT-RIRE	0.17	0.54	0.80	1.00	1.00
FT-DRIRE	0.60	0.88	0.93	0.95	0.97

Model 12. $Y = \mathbf{X}^T\beta + 0.2\epsilon$, with $p = 5$, $d = 1$, and $\beta = (1, 0, \cdots, 0)^T \in \mathbb{R}^p$. Predictor $\mathbf{X}$ is generated from multivariate normal distribution with mean zeros and identity covariance matrix. Two user-specified subspaces are $\mathcal{H}_1 = (1, 0, \cdots, 0)^T$ and $\mathcal{H}_2 = (0, 1, \cdots, 1)^T$. The percentages of rejecting the null hypothesis H_0: $\mathcal{H}^T\beta = 0$ ($\mathcal{H}$ represents $\mathcal{H}_1$ or $\mathcal{H}_2$) are presented among 100 simulated data, given the significant level 0.05.

Table 3.3: Percentages of rejecting using Marginal(M) or Conditional(C) predictors hypothesis tests with $n = 200$ in Model 12.

Hypothesis	m=2	m=3	m=4	m=5	m=10	m=20	m=50	m=100
M(FT-IRE_$\mathcal{H}_1$)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
M(FT-IRE_$\mathcal{H}_2$)	0.0500	0.0100	0.0000	0.0200	0.0000	0.0000	0.0000	0.0000
C(FT-IRE_$\mathcal{H}_1$)	1.0000	1.0000	0.9900	1.0000	0.9900	1.0000	1.0000	1.0000
C(FT-IRE_$\mathcal{H}_2$)	0.2000	0.3400	0.5400	0.5400	0.7400	0.8500	0.9400	0.9700
M(FT-RIRE_$\mathcal{H}_1$)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
M(FT-RIRE_$\mathcal{H}_2$)	0.2500	0.4400	0.5500	0.5600	0.5400	0.5900	0.6000	0.5800
C(FT-RIRE_$\mathcal{H}_1$)	1.0000	1.0000	0.9800	0.9900	0.9800	0.9900	1.0000	0.9900
C(FT-RIRE_$\mathcal{H}_2$)	0.2000	0.2900	0.4300	0.3000	0.5800	0.7700	0.7600	0.9400

Model 12 demonstrates the performance of marginal (M) and conditional (C) predictor hypothesis tests ($d = 1$) using FT-IRE and FT-RIRE. The rejection rates for $\mathcal{H}_1$ indicate the power, while the rejection rates for $\mathcal{H}_2$ mean type I error. Regardless of the sample size and the number of $\boldsymbol{\omega}$, the power of marginal or conditional predictor

Table 3.4: Percentages of rejecting using Marginal(M) or Conditional(C) predictors hypothesis tests with $m = 2$ in Model 12.

Hypothesis	n=50	n=100	n=150	n=300	n=600	n=800	n=1000
M(FT-IRE_$\mathcal{H}_1$)	0.9900	0.9900	1.0000	1.0000	1.0000	1.0000	1.0000
M(FT-IRE_$\mathcal{H}_2$)	0.1400	0.0700	0.0200	0.0500	0.0600	0.0500	0.0200
C(FT-IRE_$\mathcal{H}_1$)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
C(FT-IRE_$\mathcal{H}_2$)	0.2800	0.2400	0.1800	0.2400	0.1100	0.1800	0.1500
M(FT-RIRE_$\mathcal{H}_1$)	0.9900	0.9900	1.0000	1.0000	1.0000	1.0000	1.0000
M(FT-RIRE_$\mathcal{H}_2$)	0.2900	0.3000	0.2600	0.2200	0.2300	0.3200	0.2200
C(FT-RIRE_$\mathcal{H}_1$)	0.8500	0.9900	1.0000	1.0000	1.0000	1.0000	1.0000
C(FT-RIRE_$\mathcal{H}_2$)	0.1900	0.2700	0.2900	0.2000	0.1800	0.2600	0.2200

testing is close to one. When the sample size is 200, FT-IRE has not only the higher power but also the lower type I error regardless of the number of m for marginal predictor hypotheses (Table 3.3). But for conditional predictor testing, type I errors of FT-IRE increase as m increases. For FT-RIRE, type I errors also increase as m increases for both marginal and conditional predictor testing, thus indicating that the number of $\boldsymbol{\omega}$ influences the tests' results. The larger the m is, the poorer the performance is. In order to increase accuracy, we use a small value of m when doing tests.

In Table 3.4, we fix $m = 2$, the larger sample, the lower type I error for FT-IRE, but for FT-RIRE, the sample size has no significant effect on type I errors. Because predictor hypothesis tests have low type I error especially when the sample size is small, we do not use conditional predictor test to conduct variable selection in real data analysis.

Real Data analysis

The data set comes from the 2010 Census and 2009-2013 American Community Survey (http://goo.gl/LlcwY7). The 2015 Planning Database (PDB) assembles information of housing, demographic, socioeconomic, and Census operational data. And the data are accumulated at the block-group level, which is the smallest geographic unit used by the Census Bureau. A block-group comprises multiple blocks, usually containing between 600 and 3,000 people.

The PDB comprises approximate 220,000 block groups. The response variable is the number of people with two or more types of health insurance coverage (Y). A total of 10 variables are identified as relevant candidate predictor variables: the number of people (X_1) ages 25 years and over at the time of interview with a college degree or higher in the ACS population; the number of people (X_2) classified as below the poverty level given their total family income within the last year, family size, and family composition in the ACS population; the number of ACS households (X_3) in which the householder and his or her spouse are listed as members of the same household (not including same-sex married couples); the number of ACS households (X_4) where a householder lives alone or with non-relatives only (including same-sex couples where no relatives of the householder are present); the number of ACS households (X_5) where a householder lives alone; the number of ACS families (X_6) with related children under 6 years; the median ACS household income (X_7) for the block group; the median ACS household income (X_8) for the tract; the number of 2010 Census occupied housing units (X_9) that are not owner occupied, whether they are rented or occupied without payment of rent; the number of ACS housing units (X_{10}) where owner or co-owner lives in it. Because most of the variables are count numbers with a large range of values, we treat all of them as continuous variables. We focus on the block groups in Kentucky. We first excluded observations with missing values. There are 4097 blocks left. We then use Box-Cox transformation for the predictors to ensure that the linearity condition is approximately satisfied.

$$\widetilde{X}_1 = (X_1 + 0.5)^{0.33}, \quad \widetilde{X}_2 = (X_2 + 0.5)^{0.35}, \quad \widetilde{X}_3 = (X_3 + 0.5)^{0.53}, \quad \widetilde{X}_4 = (X_4 + 0.5)^{0.33}$$
$$\widetilde{X}_5 = (X_5 + 0.5)^{0.4}, \quad \widetilde{X}_6 = (X_6 + 0.5)^{0.45}, \quad \widetilde{X}_7 = (X_7 + 0.5)^{0.16}, \quad \widetilde{X}_8 = (X_8 + 0.5)^{0.1}$$
$$\widetilde{X}_9 = (X_9 + 0.5)^{0.33}, \quad \widetilde{X}_{10} = (X_{10} + 0.5)^{0.52}.$$

We use FT-IRE, FT-DIRE, FT-SIRE, FT-RIRE, FT-DIRE to conduct dimension testings, comparing to SIR and IRE. All these methods provide two dimensions. We plot the scatter plots (left and middle panel of Figure 3.5) of the response vs. the first two reduced predictors. There is a quadratic relationship between the response and the two reduced predictors, respectively. Hence, we fit ordinal linear regression with two degrees for each variable: $\mathrm{E}(Y) = \beta_0 + \beta_1 X_1^* + \beta_2 X_1^{*2} + \beta_3 X_2^* + \beta_3 X_2^{*2}$, where

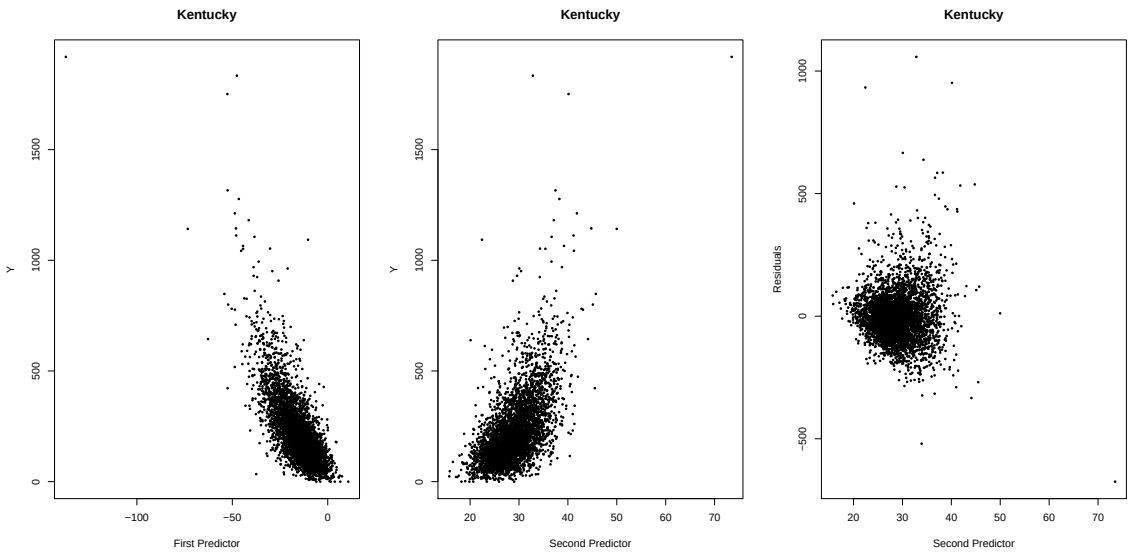

Figure 3.5: Three scatter plots for $d = 2$: (left panel) Y vs. $\hat{\boldsymbol{\beta}}_1^T\mathbf{X}$; (middle panel) Y vs. $\hat{\boldsymbol{\beta}}_2^T\mathbf{X}$; (right panel) residual of ordinal linear regression vs. $\hat{\boldsymbol{\beta}}_2^T\mathbf{X}$.

$*$ means the reduced variables. In the right panel of Figure 3.5, the residual of linear regression vs. the second predictor indicates that there is not strong pattern.

We further apply the SVS approach with FT-SIRE. The second and third columns of Table 3.5 show the estimate using FT-SIRE and the last two columns show the estimate after SVS, which indicates that X_2, X_5, X_7 and X_{10} significantly contribute in the CS.

3.6 Discussion

In this chapter, we developed the optimal quadratic function approach using Fourier transform. Not only the general approach FT-IRE but also special cases FT-DIRE, FT-SIRE, FT-RIRE, and FT-DRIRE are discussed for computational efficiency and robustness. Partial sufficient dimension reduction with Fourier transform is also investigated. Furthermore, SVS and predictor hypotheses are used for sparse situations. For the dimensionality test, smaller m is preferred. While for estimation accuracy of CS, m has less effect. For sufficient variable selection, the simulation results show that the penalized approach outperforms the testing approaches.

Table 3.5: Results from the response with ten predictors for FT-SIRE with two directions.

	FT-SIRE		*SVS FT-SIRE*	
X	First Dir.	Second Dir.	First Dir.	Second Dir.
X_1	0.0405	-0.2291	0.0000	0.0000
X_2	-0.1879	0.2018	-0.0575	-0.0018
X_3	-0.3917	0.0149	0.0000	0.0000
X_4	0.0773	0.2199	0.0000	0.0000
X_5	-0.6808	0.2772	0.0000	-0.0691
X_6	0.2860	0.1900	0.0000	0.0000
X_7	0.3310	0.4501	-0.0146	0.0000
X_8	0.1718	0.4153	0.0000	0.0000
X_9	0.0514	-0.6059	0.0000	0.0000
X_{10}	-0.3417	0.0549	0.0000	-0.0228

Chapter 4 Wavelet transform inverse regression for sufficient dimension reduction

4.1 Introduction

Wavelet analysis is a popular tool and has received much attention in compressed sensing and signal processing. It has been successfully applied in many applications such as image analysis, information system, and other engineering applications.

In the field of statistics, the review articles from Antoniadis (1997) and Antoniadis et al. (2007) summarized the applications of wavelet transform in statistics. To be specific, with the development the popularity of nonparametric function estimation in the past three decades, theoretical and applied research on the field of wavelets has had a noticeable influence on nonparametric regression, partial linear regression models and functional index models, density estimation, and many other related topics.

Although wavelet has been widely used in the estimation of nonparametric function, it hasn't been applied in the field of sufficient dimension reduction. To the best of our knowledge, this is the first attempt to incorporate wavelet transform to sufficient dimension reduction. In this chapter, we adopt wavelet transform to perform sufficient dimension reduction. Different from Chapter 2 and Chapter 3, which perform Fourier transform on the conditional mean $\mathrm{E}(\mathbf{X}|Y)$, we apply wavelet transform. The advantage of wavelet transform over Fourier transform is that wavelet transform is able to decompose complex information including low- and high-frequency into elementary patterns, while Fourier transform can't handle it (See more detail in Chui (2016)). Our wavelet-based approach provides a novel estimator and serves as a complement to the existing SDR method.

Section 4.2 reviews the basic definition of Fourier transform and wavelet transform. Section 4.3 discusses the generalize eigenvalue decomposition approach to estimate the central subspace and applied consistent order-determination procedure by Luo

and Li (2016) to determine the structural dimension. Section 4.4 proposes to obtain the optimal estimator via minimizing the discrepancy function. Section 4.5 give coordinate-independent sparse estimator and perform sufficient dimension reduction by minimizing the discrepancy function with a constructed penalty term. In Section 4.6, simulation studies are conducted to illustrate the superior performance of wavelet transform. Section 4.7 provides some promising research directions relating to wavelets in sufficient dimension reduction.

4.2 Review of Fourier and wavelet transform

The significant difference between Fourier and wavelet transforms is that Fourier transform gains information with equal time width, and wavelet transform adapts the time-widths to various frequency. Based on it, we introduce wavelet transform in SDR to capture different information for both low- and high- frequency.

The well-known Fourier transform representing the frequency information about a function $f \in L^1(\mathbb{R})$ is defined as:

$$(\mathcal{F}f)(\omega) = \frac{1}{\sqrt{2\pi}} \int e^{-i\omega t} f(t) dt,$$

where i is the imaginary unit. However the information about time-localization of high frequency cannot be detected easily from $\hat{f}(\omega)$. Hence, windowed Fourier transform (Daubechies, 1992) is discussed to achieve time-localization, which is to cut off a localized slice of f and is defined as:

$$(T^{win} f)(\omega, t) = \int f(s) g(s-t) e^{-i\omega s} ds, \tag{4.1}$$

where g is a window function. The common choice for g is the Gaussian. The discrete version of windowed Fourier transform, which is more common in the signal analysts, assign t and ω to be spaced values: $t = nt_0$, $\omega = m\omega_0$, where m, n range over $\mathbb{Z}$, and ω_0, $t_0 > 0$ are fixed:

$$T^{win}_{m,n}(f) = \int f(s) g(s - nt_0) e^{-im\omega_0 s} ds. \tag{4.2}$$

Wavelet transform (Daubechies, 1992) describes similar time-frequency as window Fourier transform. Wavelet formulas are similar to (4.1) and (4.2):

$$(T^{wav}f)(a,b) = |a|^{-1/2}\int f(t)\psi(\frac{t-b}{a})dt \tag{4.3}$$

and

$$T^{wav}_{m,n}(f) = a_0^{-m/2}\int f(t)\psi(a_0^{-m}t - nb_0)dt,$$

where $a > 0, b > 0$ and a_0, b_0 are fixed values with $m, n \in \mathbb{Z}$. Both wavelet and windowed Fourier transforms take the inner product of f with a family of functions with two indexes, $g^{\omega,t}(s) = e^{-i\omega s}g(s-t)$ in (4.1), and $\psi^{a,b}(s) = |a|^{-1/2}\psi(\frac{s-b}{a})$ in (4.3). The different frequencies can be covered by changing the value of the scaling parameter a. The larger the a, the smaller frequency it represents. While the parameter b describes the time location of $\psi^{a,b}(s)$. Both (4.1) and (4.3) indicate a time-frequency of the function f. But the difference between wavelet and windowed Fourier transform lies in the shapes of the analyzing functions $g^{\omega,t}$ and $\psi^{a,b}$. In which, $\psi^{a,b}$ adjusts the time-widths correspond to frequency. The higher frequency is, the narrower of the $\psi^{a,b}$. But $g^{\omega,t}$ has the same width but relocate the center and oscillate with different frequency.

Given wavelet transform, the function can be reconstructed by using the resolution of identity formula:

$$f = \frac{1}{C}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{a^2}(T^{wav}f)(a,b)\psi^{a,b}dadb, \tag{4.4}$$

where $C = 2\pi\int_{-\infty}^{\infty}|\hat{\psi}(\omega)|^2|\omega|^{-1}d\omega$. If $\psi \in L^2(\mathbb{R})$, the admissibility condition is satisfied, that is $C < \infty$. If $\psi \in L^1(\mathbb{R})$, then $\hat{\psi}$ is continuous and $C < \infty$ only if $\int\psi(x)dx = 0$. In the following discussion, we only focus on $\psi \in L^2(\mathbb{R})$. It is interesting that f can be constructed by its wavelet transform $T^{wav}f$ or be considered as a superposition of wavelets $\psi^{a,b}$. For the discrete wavelet transform, the orthonormal bases of wavelets (Daubechies, 1992) is discussed. Define

$$\psi_{m,n}(x) := a_0^{-m/2}\psi(a_0^{-m}x - nb_0),\ m, n \in \mathbb{Z}.$$

A function ψ is called an orthogonal wavelet, if the family $\{\psi_{m,n}\}$ is an orthonormal basis of $L^2(\mathbb{R})$; that is $< \psi_{j,k}, \psi_{l,m} >= \delta_{j,l}\cdot\delta_{k,m},\ j,k,l,m \in \mathbb{Z}$, and every $f \in L^2(\mathbb{R})$

can be written as $f(x) = \sum_{j,k=-\infty}^{\infty} c_{j,k}\psi_{j,k}(x)$. Some examples of the orthonormal basis for $L^2(\mathbb{R})$ are the following:

1. Haar basis: $\psi_h(x) = \begin{cases} 1, \ 0 \leq x \leq 1, \\ -1, \ -1 \leq x < 0, \\ 0, \ \text{Otherwise.} \end{cases}$
2. Littlewood-Paley basis: $\psi_l(x) = (\pi x)^{-1}(\sin 2\pi x - \sin \pi x)$.
3. Constant basis: $\psi_c(x) = \begin{cases} 1, \ -1 \leq x \leq 1 \\ 0, \ \text{Otherwise.} \end{cases}$
4. Fourier basis: $\psi_f(x) = e^{-ix}$.

Our goal is to employ wavelet transform into SDR: 1) Developing an estimator from applying generalize eigenvalue decomposition to a proposed kernel matrix, which is constructed from wavelet transform; 2) Constructing a minimum discrepancy approach from wavelet transform; 3) Investigating asymptotically efficient of estimators for CS and test statistics for dimensional structure; 4) Discussing the coordinate-independence sparse estimation (CISE, Chen et al. (2010)) for wavelet transform.

4.3 Generalize Eigenvalue Decomposition

Estimation Procedure

Let $\mathbf{Z} = \Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$, where $\boldsymbol{\mu}$ and Σ are the mean and covariance matrix of $\mathbf{X}$. Under the linearity condition, $m(y) = \Sigma^{-1/2}\mathrm{E}(\mathbf{Z}|Y = y) \in \mathcal{S}_{Y|\mathbf{X}}$. Let $h(y) = m(y)f(y)$, where $f(y)$ is the marginal density distribution of Y. Then

$$\begin{aligned} (T^{wav}h)(a,b) &= |a|^{-1/2} \int \Sigma^{-1/2}\mathrm{E}(\mathbf{Z}|Y = y)f(y)\psi(\tfrac{y-b}{a})dy \\ &= |a|^{-1/2}\mathrm{E}[\Sigma^{-1/2}\mathrm{E}(\mathbf{Z}|y)\psi(\tfrac{y-b}{a})] \\ &= |a|^{-1/2}\mathrm{E}[\Sigma^{-1/2}\mathbf{Z}\psi(\tfrac{y-b}{a})] \\ &\propto \mathrm{E}[\Sigma^{-1/2}\mathbf{Z}\psi^{a,b}(Y)] \end{aligned} \tag{4.5}$$

$\mathcal{S}_{\mathrm{E}(\mathbf{X}|Y)}$ is spanned by $m(y)$, that is $\mathcal{S}_{\mathrm{E}(\mathbf{X}|Y)} = \mathrm{Span}\{m(y), y \in \mathrm{supp}(f)\} = \mathrm{Span}\{m(y)f(y), y \in \mathrm{supp}(f)\} \subseteq \mathcal{S}_{Y|\mathbf{X}}$. From the resolution of identity formula (4.4) and under the linearity condition, then $(T^{wav}h)(a,b) \in \mathcal{S}_{\mathrm{E}(\mathbf{X}|Y)}$. Hence,

$$\mathrm{Span}\{(T^{wav}h)(a,b), a, b \in \mathbb{R}\} = \mathrm{Span}\{m(y), y \in \mathbb{R}\}.$$

In addition, we assume the coverage condition, $\mathcal{S}_{Y|\mathbf{X}} = \mathrm{Span}\{(T^{wav}h)(a,b), a, b \in \mathbb{R}\}$. As $\mathcal{S}_{Y|\mathbf{Z}} = \Sigma^{1/2}\mathcal{S}_{Y|\mathbf{X}}$, so $\mathrm{E}[\mathbf{Z}\psi^{a,b}(Y)] \in \mathcal{S}_{Y|\mathbf{Z}}$, we can work on either space.

Generalize Eigenvalue Decomposition Algorithm

In order to estimate CS, the main goal is to estimate the correlation between the predictor and wavelet transform, that is, $\mathrm{E}[\mathbf{X}\psi^{a,b}(Y)]$ or $\mathrm{E}[\mathbf{Z}\psi^{a,b}(Y)]$. We focus on $\mathbf{X}$ scale, and we use generalize eigenvalue decomposition. We employ the continuous wavelet transform by choosing (a,b) from real values. Here, a is the scale parametric, relating to the shape of $\psi^{a,b}$. The larger the a is, the wider $\psi^{a,b}$ is. While, the smaller the a, the narrower it is. By choosing different values of a, $\psi^{a,b}$ could detect high- or low-frequency information about Y. On the other hand, b is the translation parameter of the $\psi^{a,b}(y)$, which represents the location. We consider $\psi \in L^2(\mathbb{R})$. So $\lim_{|y|\to\infty} \psi(y) = 0$. We say $\psi(y)$ is on the ϵ-compact support $[-S, S]$, that is for any ϵ, there is $S > 0$ such that $|\psi(Y)| < \epsilon$ for $|y| > S$. For example, both Haar and Littlewood-Paley, the ϵ-compact support is $[-1, 1]$. The following are two ways to choose h pairs (a,b) and $\{y_i\}_{i=1}^N$ are random sample:

H1: For fixed h, use the same scale of $\psi^{a,b}$: $a = \frac{y_{\max} - y_{\min}}{2hS}$ and $b = \{y_{min} + a(2i-1)\}_{i=1}^h$.

H2: For fixed h, use the equal number observations within the support of $\psi^{a,b}$: $a = \{\frac{y_{(\frac{i}{h})} - y_{(\frac{i-1}{h})}}{2S}\}_{i=1}^h$ and $b = \{\frac{y_{(\frac{i}{h})} + y_{(\frac{i-1}{h})}}{2}\}_{i=1}^h$, where $y_{(\frac{i}{h})}$ represent the $\frac{i}{h}100\%$ percentile.

Here, we choose a sequence of h's. Because both large and small values of a should be included to capture high- or low-frequency information. In our limited simulation, we choose h from one to ten, that is 55 different (a,b) pairs. Those pairs are enough to achieve accurate results.

1. The sample size is N, $\{\mathbf{x}_j, y_j\}_{j=1}^N$ are random sample, and $\bar{\mathbf{x}}, \hat{\Sigma}$ are the sample mean and covariance of the $\mathbf{X}$.

2. Choose pairs $\{(a_i, b_i)\}_{i=1}^K$. For each pair (a_i, b_i), let $\hat{M}_i = \sum_{j=1}^N \mathbf{x}_j \psi^{a_i,b_i}(y_j)$. Define $\hat{M} = \{\hat{M}_i\}_{i=1}^K \in \mathbb{R}^{K\times p}$ and $\hat{W} = \hat{M} D_m \hat{M}^T$, where $\hat{M}^T$ is the transpose of $\hat{M}$, and $D_m = \text{diag}\{\hat{p}_i\}_{i=1}^K$ with $\hat{p}_i = \frac{1}{N}\sum_{j=1}^N I(\psi^{a_i,b_i}(y_j) \neq 0)$.

3. Conduct the generalize eigenvalue decomposition of $\hat{W}$, and the eigenvectors $\{\hat{\beta}_l\}_{l=1}^d$ corresponding to the first d largest eigenvalues $\{\hat{\lambda}_l\}_{l=1}^d$ of $\hat{W}$, satisfying $\hat{W}\hat{\beta}_l = \hat{\lambda}_l \Sigma \hat{\beta}_l$.

If using $\psi_c(y)$, fixing the value h and employing the equal number of observations (H1), then wavelet transform approach is the same as SIR with slice h. If using $\psi_f(y)$, $a = \frac{1}{\omega}$, and $b = 0$, wavelet transform is equivalent to Fourier transform, but different strategy to choose ω. From this respect, wavelet transform is more flexible than SIR and FT. In the simulations, we illustrate the necessary of introducing wavelet transform, which get more profound and accurate results.

Testing Structural Dimension

This section is to estimate the dimension of CS. From the previous section, we have constructed the kernel matrix W. To find the dimension is the same as to find the rank of W. We employ Luo and Li (2016)'s order determination method, combining eigenvalues and variation of eigenvectors of W. Let $\{\mathbf{X}_j^*, Y_j^*\}_{i=1}^N$ is a bootstrap sample and W^* is the kernel matrix using the bootstrap sample. Assume that the eigenvalues of $\hat{W}$ and W^* are $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$ and $\lambda_1^* \geq \lambda_2^* \geq \cdots \geq \lambda_p^*$, and $(\hat{\beta}_1, \cdots, \hat{\beta}_p)$ and $(\beta_1^*, \cdots, \beta_p^*)$ are the corresponding eigenvectors. Let $\hat{\boldsymbol{\beta}}_k = (\hat{\beta}_1, \cdots, \hat{\beta}_k)$ and $\boldsymbol{\beta}_k^* = (\beta_1^*, \cdots, \beta_k^*)$, where $k < p$, and using $\boldsymbol{\beta}_{k,i}^*$ with subscript i denotes bootstrap sample ith, where $i \in (1, \cdots, B)$. Define a function to represent the discrepancy of $\hat{\boldsymbol{\beta}}_k$ and $\boldsymbol{\beta}_k^*$,

$$f_B^0(k) = \begin{cases} 0, & k = 0, \\ B^{-1}\sum_{i=1}^B \{1 - |\det(\hat{\beta}_k^T \boldsymbol{\beta}_{k,i}^*)|\}, & k = 1, \cdots, p-1. \end{cases}$$

Then to normalize the eigenvalues and f_B^0, that is $\phi(k) = \hat{\lambda}_{k+1}/(1 + \sum_{i=0}^{p-1} \hat{\lambda}_{i+1})$ and $f_B(k) = f_B^0(k)/\{1 + \sum_{i=1}^{p-1} f_n^0(i)\}$. Define $g(k) = \psi(k) + f_B(k)$ and determine the dimension using

$$\hat{d} = \arg\min_k g(k).$$

4.4 Minimum Discrepancy Approach

We will discuss an optimal estimate by constructing discrepancy function between wavelet transform and the population values, which multiply the true basis of the central subspace with the corresponding coefficients. The way we construct the kernel matrix, using one observation multiple times, leads to a high correlation between wavelet transform. It is natural to adjust the discrepancy function by constructing the asymptotic covariance matrix of wavelet function, similar to the weighted least square. We demonstrate the asymptotic distribution to find the covariance matrix and develop algorithms to find the estimate and test statistics.

Notations

We focus on continuous wavelet transform as in Section 4.3, and (a, b) are chosen as in Section 4.3, $\{a_i, b_i\}_{i=1}^K$. Here, we denote $\xi_i := \Sigma T_{a_i,b_i}^{wav}(f) - \mathrm{E}(\mathbf{X})\mathrm{E}[\psi_i(Y)]$ and $\hat{\xi}_i = \frac{1}{N}\sum_{j=1}^N \mathbf{x}_j\psi_i(y_j) - \bar{\mathbf{x}}\bar{\psi_i}(y)$, where $\psi_i(y_j) = \psi^{a_i,b_i}(y_j)$. Let $\xi = \{\xi_i\}_{i=1}^T$ and $\hat{\xi} = \{\hat{\xi}_i\}_{i=1}^K$. From section 4.3, $\Sigma^{-1}\mathrm{Span}\{\xi_i\} \in \mathcal{S}_{Y|\mathbf{X}}$. If $\boldsymbol{\beta}$ is the basis of $\mathcal{S}_{Y|\mathbf{X}}$, there exits C_i such that $\Sigma^{-1}\xi_i = \boldsymbol{\beta}C_i$ for $i \in \{1, \cdots, K\}$. If we can find V such that $\sqrt{n}[\mathrm{vec}(\hat{\Sigma}^{-1}\hat{\xi}) - \mathrm{vec}(\boldsymbol{\beta}C)] \to N(0, V)$, then define the discrepancy function as

$$F_d(\boldsymbol{\beta}, C; V) = N[\mathrm{vec}(\hat{\Sigma}^{-1}\hat{\xi}) - \mathrm{vec}(\boldsymbol{\beta}C)]^T V^{-1}[\mathrm{vec}(\hat{\Sigma}^{-1}\hat{\xi}) - \mathrm{vec}(\boldsymbol{\beta}C)], \tag{4.6}$$

which will follow the chi-square distribution. The following sections are meant to find the V, the algorithm for $\boldsymbol{\beta}$ and C and asymptotic properties.

Asymptotic Properties

Let $\epsilon_i(y) = \psi_i(y) - \mathrm{E}[\psi_i(Y)] - \mathbf{Z}^T\mathrm{E}[\mathbf{Z}\psi_i(Y)]$ and $\mathbf{X}^c = \mathbf{X} - \boldsymbol{\mu}$ denote the center of ψ_i and $\mathbf{X}$. We have the following results:

Theorem 11. *Assume that* $\{y_j, \mathbf{x}_j\}$,$j = 1, ..., N$ *are random samples of* $(Y, \mathbf{X})$ *with finite fourth moments. Then*

$$\sqrt{n}[\text{vec}(\hat{\Sigma}^{-1}\hat{\xi}) - \text{vec}(\boldsymbol{\beta}C)] \xrightarrow{D} N(0, V),$$

where $V = \text{Cov}\{\text{vec}[\Sigma^{-1}\mathbf{X}^c\epsilon^T]\} \in \mathbb{R}^{pK\times pK}$.

Theorem 12. *Assume that* $\{y_j, \mathbf{x}_j\}$,$j = 1, ..., N$ *are random samples of* $(Y, \mathbf{X})$ *with finite fourth moments. Let* $(\hat{\boldsymbol{\beta}}, \hat{C}) = \arg\min_{\boldsymbol{\beta},C} F_d(\boldsymbol{\beta}, C; \hat{V}^{-1})$. *Then the following results hold:*

1. $\text{vec}(\hat{\boldsymbol{\beta}}\hat{C})$ *is asymptotically efficient, and* $\sqrt{n}[\text{vec}(\hat{\boldsymbol{\beta}}\hat{C}) - \text{vec}(\boldsymbol{\beta}C)]$ *is asymptotically normal with mean 0 and some covariance matrix* $\Delta(\Delta^T V^{-1}\Delta)^-\Delta^T$, *where* $\Delta = (C^T \otimes I_p, I_{2m} \otimes \boldsymbol{\beta})$ *with* $Kp \times d(p + K)$ *dimensions.*

2. $\hat{F}_d$ *has an asymptotic chi-square distribution with degrees of freedom* $(p-d)(K-d)$.

3. $\text{Span}(\hat{\boldsymbol{\beta}})$ *is a consistent estimator of* $\mathcal{S}_{Y|\mathbf{X}}$.

The second statement can be used to determine the dimension with sequential tests.

Algorithm

1. Choose an initial value for $\boldsymbol{\beta}_0 \in \mathbb{R}^{p\times d}$. An initial choice will affect the speed of convergence. We use the general eigenvalue decomposition estimation.

2. Fixed $\boldsymbol{\beta}$, update C by minimizing $F_d(\boldsymbol{\beta}, C; V)$. Here, $\text{vec}(C)$ can be constructed by fitting linear regression $V^{1/2}\text{vec}(\hat{\Sigma}^{-1}\hat{\xi})$ on $V^{1/2}(I_K \otimes \boldsymbol{\beta})$, that is $\text{vec}(C) = [(I_K \otimes \boldsymbol{\beta}^T)V(I_K \otimes \boldsymbol{\beta})]^{-1}(I_K \otimes \boldsymbol{\beta}^T)V\text{vec}(\hat{\Sigma}^{-1}\hat{\xi})$. Assign err to be $F_d(\boldsymbol{\beta}, C; V)$.

3. Fixed C, minimize $F_n(\boldsymbol{\beta}, C; V)$ with respect to one column of $\boldsymbol{\beta}$, subject to unit norm and orthogonal to other columns (keeping them constants). For this partial minimization problem, the quadratic discrepancy function is $F(\beta_k) = (\boldsymbol{\alpha}_k - (\boldsymbol{c}_k^T \otimes I_p)Q_{\boldsymbol{\beta}_{(-k)}}\beta_k)^T V(\boldsymbol{\alpha}_k - (\boldsymbol{c}_k^T \otimes I_p)Q_{\boldsymbol{\beta}_{(-k)}}\beta_k)$, where $\boldsymbol{\alpha}_k = \text{vec}(\hat{\Sigma}^{-1}\hat{\xi} -$

$\boldsymbol{\beta}_{(-k)}C_{(-k)})$, $\boldsymbol{c}_k$ is kth column of C, $C_{(-k)}$ (or $\boldsymbol{\beta}_{(-k)}$) are deleting k^{th} column from C (or $\boldsymbol{\beta}$) and $Q_{\boldsymbol{\beta}_{(-k)}}$ is orthogonal complement of Span($\boldsymbol{\beta}_{(-k)}$).

a) For $k = 1, ..., d$:

i. Denote $\boldsymbol{\beta} = (\beta_1, ..., \beta_{k-1}, \beta_k, \beta_{k+1}, ..., \beta_d)$ and update $\hat{\beta}_k = Q_{\boldsymbol{\beta}_{(-k)}}[Q_{\boldsymbol{\beta}_{(-k)}}(\boldsymbol{c}_k^T \otimes I_p)V(\boldsymbol{c}_k^T \otimes I_p)Q_{\boldsymbol{\beta}_{(-k)}}]^{-}Q_{\boldsymbol{\beta}_{(-k)}}(\boldsymbol{c}_k^T \otimes I_p)V\boldsymbol{\alpha}_k$, then normalize $\hat{\beta}_k$ using $\hat{\beta}_k/||\hat{\beta}_k||$.

ii. Update $\boldsymbol{\beta}$ by replace β_k with $\hat{\beta}_k$ and update C like step 2.

b) Update err with $F_d(\boldsymbol{\beta}, C; V)$.

4. Return to step 3 until err less than 10^{-6}, say.

5. The resulting estimates are: $\hat{\boldsymbol{\beta}} = (\beta_1, \cdots, \beta_d)$.

4.5 Sufficient Variable Selection

We now discuss SVS, following the ideas of Chen et al. (2010) and Qian et al. (2018). Let $\boldsymbol{\beta} = (\alpha_1, ..., \alpha_p)^T$ and $\alpha_i \in \mathbb{R}^d$ be the i^{th} row of $\boldsymbol{\beta}$. Simultaneous variable selection is to find a set $\mathcal{A}_0 = \{1 \leq j \leq p : \boldsymbol{e}_j^T\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{e}_j > 0\}$. The number of $\mathcal{A}_0$, $|\mathcal{A}_0|$, is denote as u, which indicates the number of important predictors. The coordinate-independent penalty is $p_{\boldsymbol{w}}(\boldsymbol{\beta}) = \sum_{j=1}^{p} w_j||\alpha_j||_2$ and $\boldsymbol{w} = (w_1, ..., w_p)$ are the penalty weights. The object function can be written as:

$$L_n(\boldsymbol{\beta}, C) = [\text{vec}(\hat{\xi}) - \text{vec}(\Sigma\boldsymbol{\beta}C)]^T\Lambda[\text{vec}(\hat{\xi}) - \text{vec}(\Sigma\boldsymbol{\beta}C)] + \lambda p_{\boldsymbol{w}}(\boldsymbol{\beta}), \text{ s.t. } CC^T = I_d, \tag{4.7}$$

where $\Lambda = \text{Cov}\{\text{vec}[(\mathbf{X}^c)\epsilon^T]\}^{-1}$.

$$\begin{aligned}
\Lambda^{-1} &= \text{Cov}\{\text{vec}[(\mathbf{X} - \boldsymbol{\mu})\epsilon^T]\} &&= \text{Cov}[\text{vec}(\Sigma^{1/2}\mathbf{Z}\epsilon^T)] \\
&= \text{Cov}[(I_{2m} \otimes \Sigma^{1/2})\text{vec}(\mathbf{Z}\epsilon^T)] &&= (I_{2m} \otimes \Sigma^{1/2})\text{Cov}[\text{vec}(\mathbf{Z}\epsilon^T)](I_{2m} \otimes \Sigma^{1/2}) \\
&= (I_{2m} \otimes \Sigma^{1/2})\text{E}[\epsilon\epsilon^T \otimes \mathbf{Z}\mathbf{Z}^T](I_{2m} \otimes \Sigma^{1/2})
\end{aligned}$$

It can be proved that ϵ is uncorrelated with $\mathbf{Z}$ and $\text{E}(\epsilon) = 0$. If we assume that ϵ is independent of $\mathbf{Z}$, then $\text{E}(\epsilon\epsilon^T \otimes \mathbf{Z}\mathbf{Z}^T) = \text{E}(\epsilon\epsilon^T) \otimes I_p$. Hence, $\Lambda = (I_{2m} \otimes$

$\Sigma^{-1/2})[\mathrm{E}(\epsilon\epsilon^T)^{-1} \otimes I_p](I_{2m} \otimes \Sigma^{-1/2}) = \mathrm{E}(\epsilon\epsilon^T)^{-1} \otimes \Sigma^{-1}$. Define the first derivative of $L_n(\boldsymbol{\beta}, C)$ as the following:

$$\begin{aligned}
U(\boldsymbol{\beta}, C) &= \frac{\partial L_n(\boldsymbol{\beta},C)}{2\partial \mathrm{vec}(\boldsymbol{\beta}^T)} \\
&= \frac{\partial}{2\partial \mathrm{vec}(\boldsymbol{\beta}^T)}[-2\mathrm{vec}(\Sigma\boldsymbol{\beta}C)^T\mathrm{vec}(\hat{\xi}) + \mathrm{vec}(\Sigma\boldsymbol{\beta}C)^T\Lambda\mathrm{vec}(\Sigma\boldsymbol{\beta}C)] \\
&= \frac{\partial}{2\partial \mathrm{vec}(\boldsymbol{\beta}^T)}[-2\mathrm{vec}(\boldsymbol{\beta})^T(C \otimes \Sigma)\Lambda\mathrm{vec}(\hat{\xi}) + \mathrm{vec}(\boldsymbol{\beta})^T(C \otimes \Sigma)\Lambda(C^T \otimes \Sigma)\mathrm{vec}(\boldsymbol{\beta})] \\
&= \frac{\partial}{2\partial \mathrm{vec}(\boldsymbol{\beta}^T)}\{-2\mathrm{vec}(\boldsymbol{\beta})^T[C\mathrm{E}(\epsilon\epsilon^T)^{-1} \otimes I_p]\mathrm{vec}(\hat{\xi}) + \mathrm{vec}(\boldsymbol{\beta})^T[C\mathrm{E}(\epsilon\epsilon^T)^{-1}C^T \otimes \Sigma]\mathrm{vec}(\boldsymbol{\beta})\} \\
&= \frac{\partial}{2\partial \mathrm{vec}(\boldsymbol{\beta}^T)}\{-2\mathrm{vec}(\boldsymbol{\beta})^T\mathrm{vec}(\hat{\xi}\mathrm{E}(\epsilon\epsilon^T)^{-1}C^T) + \mathrm{vec}(\boldsymbol{\beta})^T\mathrm{vec}(\Sigma\boldsymbol{\beta}C\mathrm{E}(\epsilon\epsilon^T)^{-1}C^T)\} \\
&= \frac{\partial}{2\partial \mathrm{vec}(\boldsymbol{\beta}^T)}\{-2\mathrm{vec}[C\mathrm{E}(\epsilon\epsilon^T)^{-1}\hat{\xi}^T]^T\mathrm{vec}(\boldsymbol{\beta}^T) + \mathrm{vec}[C\mathrm{E}(\epsilon\epsilon^T)^{-1}C^T\boldsymbol{\beta}^T\Sigma]^T\mathrm{vec}(\boldsymbol{\beta}^T)\} \\
&= -\mathrm{vec}[C\mathrm{E}(\epsilon\epsilon^T)^{-1}\hat{\xi}^T] + \mathrm{vec}[C\mathrm{E}(\epsilon\epsilon^T)^{-1}C^T\boldsymbol{\beta}^T\Sigma] \\
&= -[I_p \otimes C\mathrm{E}(\epsilon\epsilon^T)^{-1}\hat{\xi}^T]\mathrm{vec}(I_p) + [\Sigma \otimes C\mathrm{E}(\epsilon\epsilon^T)^{-1}C^T]\mathrm{vec}(\boldsymbol{\beta}^T),
\end{aligned}$$

and

$$H = \frac{\partial^2 L_n(\boldsymbol{\beta},C)}{2\partial^2 \mathrm{vec}(\boldsymbol{\beta}^T)} = \Sigma \otimes C\mathrm{E}(\epsilon\epsilon^T)^{-1}C^T.$$

We notice that neither $U(\boldsymbol{\beta}, C)$ nor H depend on the inverse of Σ, but it needs to know the inverse of $\mathrm{E}(\epsilon\epsilon^T)$. $\mathrm{E}(\epsilon\epsilon^T)$ is a covariance matrix, so $\mathrm{E}(\epsilon\epsilon^T) \geq 0$.

Let $\tilde{h} = \lambda_{\max}(\Sigma \otimes C\mathrm{E}(\epsilon\epsilon^T)^{-1}C^T)$ and $\tilde{H} = \tilde{h}I_p \otimes I_d$. At the $(t+1)^{th}$ iteration, $(\boldsymbol{\beta}_{(t)}, C_{(t)})$ is the estimate of $(\boldsymbol{\beta}, C)$ after the t^{th} iteration, and then we update $\boldsymbol{\beta}$ using the quadratic approximation of L_n by

$$L_n^{(t)}(\boldsymbol{\beta}) = U_t^T(\mathrm{vec}(\boldsymbol{\beta}^T) - \mathrm{vec}(\boldsymbol{\beta}_{(t)}^T)) + \frac{\tilde{h}}{2}(\mathrm{vec}(\boldsymbol{\beta}^T) - \mathrm{vec}(\boldsymbol{\beta}_{(t)}^T))^T(\mathrm{vec}(\boldsymbol{\beta}^T) - \mathrm{vec}(\boldsymbol{\beta}_{(t)}^T)) + \lambda p_{\boldsymbol{w}}(\boldsymbol{\beta}),$$

where $U_t = U(\boldsymbol{\beta}_{(t)}, C_{(t)})$. So $\boldsymbol{\beta}_{(t+1)} = \mathrm{argmin}_{\boldsymbol{\beta}} L_n^{(t)}(\boldsymbol{\beta})$. By the Karush-Kuhn-Tucker condition, the l^{th} row of $\boldsymbol{\beta}_{(t+1)}$ has the form

$$\beta_l^{(t+1)} = \frac{1}{\tilde{h}}(1 - \frac{\lambda w_l}{||\tilde{h}\beta_l^{(t)} - U_l^{(t)}||_2})_+(\tilde{h}\beta_l^{(t)} - U_l^{(t)}),$$

where $z_+ = \max(z, 0)$, and $U_l^{(t)}$ is the l^{th} row of U_t after reforming as $p \times d$ matrix. Next, we fix $\boldsymbol{\beta} = \boldsymbol{\beta}_{(t+1)}$, and update C using the Reduced Rank Procrustes Rotation to solve:

$$C_{(t+1)} = \mathrm{argmin}_C - \mathrm{trace}[\mathrm{E}(\epsilon\epsilon^T)^{-1}\hat{\xi}^T\boldsymbol{\beta}_{(t+1)}C]$$

Thus, $C_{(t+1)} = W_2W_1^T$, where $W_1DW_2^T$ is the singular value decomposition of $\mathrm{E}(\epsilon\epsilon^T)^{-1}\hat{\xi}^T\boldsymbol{\beta}_{(t+1)}$.

4.6 Numerical Study

This section we use two simulation examples to illustrate the advantages of wavelet transform. The benefit of wavelet transform is to deal with more than one 'wave'. We use the distance measure of two matrices: $D(\beta_1, \beta_2) = ||P_1 - P_2||_f$, where P_i represent the projection on Span(β_i) for $i = 1, 2$. This distance measures the difference between two spaces spanned by $\beta_i, i = 1, 2$.

Model 1: $Y = 3/(|\beta_r^T \mathbf{X}| + 1)(\sin(1.5\beta_r^T \mathbf{X}) + \sin(2\beta_r^T \mathbf{X})) + 0.1\epsilon$,

Model 2: $Y = (\beta_r^T \mathbf{X})^2 + 0.2\epsilon$,

where β_r has 1s at random $|\sqrt{p}|_-$ positions. $|\cdot|_-$ is the largest integer smaller than the given values. Let $\mathbf{X}$ be $N_p(0, \Sigma)$, where $\Sigma = (\sigma_{ij}) = (0.5^{|i-j|})$. And ϵ is a standard normal random variable. We compare three methods: FT, wavelets with Littlewood-Paley and Haar. Here, we have two settings are 1) fix $p = 15$, vary sample size from $N = (100, 200, 400, 800, 1500, 2000)$, and 2) fix $N = 1000$, vary the dimension $p = (10, 20, \cdots, 100)$. For each setting, simulations run for 100 times and the mean distance are reported. Also, we use H2 to choose (a, b). The smaller the mean distance, the more accurate the estimate. In Model 1, Table 4.1 and 4.2 report that wavelet transform with Haar basis has the lowest distance. In the model, there are two frequency $\sin(1.5\beta_r^T \mathbf{X})$ and $\sin(2\beta_r^T \mathbf{X})$ in Model 1 because wavelet transform performs better to capture different frequencies at a time. Littlewood-Paley basis fails to get better result comparing to FT. Because it self involves lots of 'wave', which brings in noises instead of information. But Littlewood-Paley could detect symmetric relationship between Y and $\beta^T \mathbf{X}$ better than FT in Model 2, because of containing small and big waves that could not cancel out symmetric information.

4.7 Discussion

In the future study, we will prove the theoretical properties in wavelet method, and optimize wavelet algorithm for large p small n case when generalize eigenvalue algorithm fails. The alternating direction method of multipliers (ADMM) is an algorithm that

Table 4.1: Mean of distance D over 100 simulations for $p = 15$ in Models 1 and 2.

	Model 1			Model 2		
Sample Size	FT	W-littlewood	W-haar	FT	W-littlewood	W-haar
N=100	1.3707	1.4556	1.2572	1.4546	1.0644	0.9699
N=200	1.0944	1.3569	0.9001	1.388	0.8606	0.8216
N=400	0.7642	1.3083	0.6293	1.3788	0.7295	0.7754
N=800	0.5293	1.1158	0.4147	1.3591	0.6562	0.733
N=1500	0.3665	0.7329	0.2952	1.3443	0.5865	0.6701
N=2000	0.3273	0.5993	0.2687	1.3551	0.5677	0.67

Table 4.2: Mean of distance D over 100 simulations for $N = 1000$ in Models 1 and 2.

	Model 1			Model 2		
Dimension	FT	W-littlewood	W-haar	FT	W-littlewood	W-haar
p=10	0.356	0.8114	0.2998	1.291	0.5006	0.5208
p=20	0.5785	1.1652	0.4461	1.3842	0.7144	0.7668
p=30	0.8142	1.3189	0.5403	1.4036	0.8672	0.9344
p=40	1.0684	1.3636	0.6464	1.4193	0.9646	1.0471
p=50	1.2372	1.4121	0.7431	1.4324	1.0529	1.1337
p=60	1.2751	1.4304	0.8104	1.4442	1.1339	1.1688
p=70	1.3498	1.4376	0.9	1.4503	1.1925	1.2565
p=80	1.3753	1.4501	0.9466	1.4619	1.2391	1.2704
p=90	1.4237	1.463	1.0262	1.4725	1.2697	1.2842
p=100	1.4443	1.4752	1.096	1.4822	1.321	1.321

solves convex optimization problems, which might help us to solve the non-invertible of the covariance matrix. For SVS, we will code the algorithms to investigate the performance of coordinate-independence sparse estimate. Simulations will be conducted to compare different wavelet basis: Haar, Littlewood, constant and Fourier.

Appendices

Appendix A: Proof for Chapter 2

Proof of Equivalent of FT and SIR when response variable is categorical. Assume Y is univariate, and Y has K levels $\{0, 1, \cdots, K-1\}$ with probability $P_y = P(Y = y) > 0$, $y \in \{0, 1, \cdots, K-1\}$. Let $\mathcal{S}_{ft} = \text{Span}\{\psi(\omega), \omega \in \mathbb{R}\}$ and $\mathcal{S}_{sir} = \text{Span}\{\text{E}(\mathbf{Z}|Y = y), y \in \{0, 1, \cdots, K-1\}\}$.

$$\begin{aligned}
\psi(\omega) &= \text{E}[\text{E}(e^{i\omega y}\mathbf{Z}|Y = y)] \\
&= \text{E}(\mathbf{Z}|Y = 0)P(Y = 0) + \text{E}(e^{i\omega}\mathbf{Z}|Y = 1)P(Y = 1) + \cdots \\
&\quad + \text{E}(e^{i\omega(K-1)}\mathbf{Z}|Y = K-1)P(Y = K-1) \\
&= P_0\text{E}(\mathbf{Z}|Y = 0) + P_1 e^{i\omega}\text{E}(\mathbf{Z}|Y = 1) + \cdots + P_{K-1}e^{i\omega(K-1)}\text{E}(\mathbf{Z}|Y = K-1).
\end{aligned}$$

Because $\text{E}(\mathbf{Z}|Y) \in \mathcal{S}_{sir}$, then $\mathcal{S}_{ft} \subseteq \mathcal{S}_{sir}$.

Now, choose $\omega_1, \cdots, \omega_{K-1}$ such that they are all different numbers.

$$\begin{aligned}
\psi(0) &= P_0\text{E}(\mathbf{Z}|Y = 0) + P_1\text{E}(\mathbf{Z}|Y = 1) + \cdots + P_{K-1}\text{E}(\mathbf{Z}|Y = K-1), \\
\psi(\omega_1) &= P_0\text{E}(\mathbf{Z}|Y = 0) + P_1 e^{i\omega_1}\text{E}(\mathbf{Z}|Y = 1) + \cdots + P_{K-1}e^{i\omega_1(K-1)}\text{E}(\mathbf{Z}|Y = K-1), \\
&\vdots \\
\psi(\omega_{K-1}) &= P_0\text{E}(\mathbf{Z}|Y = 0) + P_1 e^{i\omega_{K-1}}\text{E}(\mathbf{Z}|Y = 1) + \cdots + P_{K-1}e^{i\omega_{K-1}(K-1)}\text{E}(\mathbf{Z}|Y = K-1).
\end{aligned}$$

And the following matrix is nonsingular:

$$A = \begin{pmatrix}
1 & 1 & 1 & \cdots & 1 \\
1 & e^{i\omega_1} & e^{i2\omega_1} & \cdots & e^{i(K-1)\omega_1} \\
1 & e^{i\omega_2} & e^{i2\omega_2} & \cdots & e^{i(K-1)\omega_2} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
1 & e^{i\omega_{K-1}} & e^{i2\omega_{K-1}} & \cdots & e^{i(K-1)\omega_{K-1}}
\end{pmatrix} \text{diag}(P_y).$$

Because $|A| = \prod_{y=1}^{K-1} P_y \prod_{y=2}^{K-1}(e^{i\omega_y} - e^{i\omega_1}) \prod_{y=3}^{K-1}(e^{i\omega_y} - e^{i\omega_2}) \cdots (e^{i\omega_{K-1}} - e^{i\omega_{K-2}}) \neq 0$, then we have $(\text{E}(\mathbf{Z}|Y = 0), \cdots, \text{E}(\mathbf{Z}|Y = K-1))^T = A^{-1}(\psi(0), \psi(\omega_1), \cdots \psi(\omega_{K-1}))^T$. Because $\psi(0), \psi(\omega_1), \cdots, \psi(\omega_{K-1}) \in \mathcal{S}_{ft}$, then $\text{E}(\mathbf{Z}|Y = y) \in \mathcal{S}_{ft}$ for $y \in \{0, 1, \cdots, K-1\}$. That is, $\mathcal{S}_{sir} \subseteq \mathcal{S}_{ft}$. Hence, $\mathcal{S}_{ft} = \mathcal{S}_{sir}$. □

Proof of Proposition 2. To obtain the asymptotic distribution of $\hat{\Lambda}_d$, fix t and choose $\{\boldsymbol{\omega}_j\}_{j=1}^t$. For $j = 1, \cdots, t$, define:

$$\hat{\psi}_{j1} = \frac{1}{n}\sum_{k=1}^n \hat{\mathbf{z}}_k \cos(\boldsymbol{\omega}_j^T \mathbf{y}_k), \qquad \psi_{j1} = \mathrm{E}[\mathbf{Z}\cos(\boldsymbol{\omega}_j^T\mathbf{Y})],$$
$$\hat{\psi}_{j2} = \frac{1}{n}\sum_{k=1}^n \hat{\mathbf{z}}_k \sin(\boldsymbol{\omega}_j^T \mathbf{y}_k), \qquad \psi_{j2} = \mathrm{E}[\mathbf{Z}\sin(\boldsymbol{\omega}_j^T\mathbf{Y})].$$

Let $\hat{\Psi} = (\hat{\psi}_{11}, \hat{\psi}_{12}, \cdots, \hat{\psi}_{t1}, \hat{\psi}_{t2})$ and $\Psi = (\psi_{11}, \psi_{12}, \cdots, \psi_{t1}, \psi_{t2})$. Following (Cook, 1998, Page: 207), by Singular-Value Decomposition, $\Psi = \Gamma^T \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \Phi$, where Γ and Φ are respective $p \times p$ and $2t \times 2t$ orthogonal matrices, and D is a $d \times d$ diagonal matrix of positive values. Let $\Gamma^T = (\Gamma_1, \Gamma_0)$ and $\Phi^T = (\Phi_1, \Phi_0)$, where Γ_0 is $p \times (p-d)$ and Φ_0 is $2t \times (2t-d)$. In X-scale and $j = 1, \cdots, t$, define:

$$\hat{\theta}_{j1} = \frac{1}{n}\sum_{k=1}^n \mathbf{x}_k \cos(\boldsymbol{\omega}_j^T \mathbf{y}_k), \qquad \theta_{j1} = \mathrm{E}[\mathbf{X}\cos(\boldsymbol{\omega}_j^T\mathbf{Y})],$$
$$\hat{\theta}_{j2} = \frac{1}{n}\sum_{k=1}^n \mathbf{x}_k \sin(\boldsymbol{\omega}_j^T \mathbf{y}_k), \qquad \theta_{j2} = \mathrm{E}[\mathbf{X}\sin(\boldsymbol{\omega}_j^T Y)].$$
$$\hat{\Theta} = \left(\hat{\theta}_{11}, \hat{\theta}_{12}, \cdots, \hat{\theta}_{t1}, \hat{\theta}_{t2}\right), \qquad \Theta = (\theta_{11}, \theta_{12}, \cdots, \theta_{t1}, \theta_{t2}).$$
$$V = \Psi\Psi^T, \qquad \hat{V} = \hat{\Psi}\hat{\Psi}^T.$$

Set $\hat{Q} = \left(\frac{1}{n}\sum \cos(\boldsymbol{\omega}_1^T\mathbf{y}_k), \frac{1}{n}\sum \sin(\boldsymbol{\omega}_1^T\mathbf{y}_k), \cdots, \frac{1}{n}\sum \cos(\boldsymbol{\omega}_t^T\mathbf{y}_k), \frac{1}{n}\sum \sin(\boldsymbol{\omega}_t^T\mathbf{y}_k)\right)^T$, and $Q = \left(\mathrm{E}(\cos\boldsymbol{\omega}_1^T\mathbf{Y}), \mathrm{E}(\sin\boldsymbol{\omega}_1^T\mathbf{Y}), \cdots, \mathrm{E}(\cos\boldsymbol{\omega}_t^T\mathbf{Y}), \mathrm{E}(\sin\boldsymbol{\omega}_t^T\mathbf{Y})\right)^T$.

Look at the one column of Ψ as an example, say, $\mathrm{E}[\mathbf{Z}\cos(\boldsymbol{\omega}^T\mathbf{Y})]$. Then

$$\begin{aligned}
\mathrm{E}[\mathbf{Z}\cos(\boldsymbol{\omega}^T\mathbf{Y})] &= \mathrm{E}[\Sigma^{-1/2}(\mathbf{X}-\boldsymbol{\mu})\cos(\boldsymbol{\omega}^T\mathbf{Y})] \\
&= \Sigma^{-1/2}\mathrm{E}[(\mathbf{X}-\boldsymbol{\mu})\cos(\boldsymbol{\omega}^T\mathbf{Y})] \\
&= \Sigma^{-1/2}\{\mathrm{E}[\mathbf{X}\cos(\boldsymbol{\omega}^T\mathbf{Y})] - \boldsymbol{\mu}\mathrm{E}[\cos(\boldsymbol{\omega}^T\mathbf{Y})]\}.
\end{aligned}$$

So $\Psi = \Sigma^{-1/2}\Theta - \Sigma^{-1/2}\boldsymbol{\mu}Q^T$. And $\Gamma_0^T\Psi\Phi_0 = 0$, that is, $\Gamma_0^T\Sigma^{-1/2}(\Theta - \boldsymbol{\mu}Q^T)\Phi_0 = 0$. Define $\hat{A} = \hat{\Sigma}^{-1/2}\Sigma^{1/2}$, then

$$\begin{aligned}
\sqrt{n}\Gamma_0^T\hat{\Psi}\Phi_0 &= \sqrt{n}\Gamma_0^T\hat{\Sigma}^{-1/2}(\hat{\Theta} - \bar{\boldsymbol{x}}\hat{Q}^T)\Phi_0 = \sqrt{n}\Gamma_0^T\hat{A}\Sigma^{-1/2}(\hat{\Theta} - \bar{\boldsymbol{x}}\hat{Q}^T)\Phi_0 \\
&= \sqrt{n}\Gamma_0^T(\hat{A} - I + I)\Sigma^{-1/2}[\hat{\Theta} - \Theta + \Theta - \boldsymbol{\mu}Q^T + \boldsymbol{\mu}(Q^T - \hat{Q}^T) + (\boldsymbol{\mu} - \bar{\boldsymbol{x}})\hat{Q}^T]\Phi_0.
\end{aligned}$$

Here $(\hat{A} - I)\Sigma^{-1/2}(\hat{\Theta} - \Theta) = O_p(\frac{1}{n})$, $(\hat{A} - I)\Sigma^{-1/2}\boldsymbol{\mu}(Q^T - \hat{Q}^T) = O_p(\frac{1}{n})$, $(\hat{A} - I)\Sigma^{-1/2}(\boldsymbol{\mu} - \bar{\boldsymbol{x}})\hat{Q}^T = O_p(\frac{1}{n})$ and $\Gamma_0^T(\hat{A} - I)\Sigma^{-1/2}(\Theta - \boldsymbol{\mu}Q^T)\Phi_0 = 0$. Hence,

$$\begin{aligned}
\sqrt{n}\Gamma_0^T\hat{\Psi}\Phi_0 &= \sqrt{n}\Gamma_0^T\Sigma^{-1/2}[\hat{\Theta} - \Theta + \boldsymbol{\mu}(Q^T - \hat{Q}^T) + (\boldsymbol{\mu} - \bar{\boldsymbol{x}})Q^T]\Phi_0 + O_p(\tfrac{1}{n}) \\
&= \sqrt{n}\Gamma_0^T\Sigma^{-1/2}[\hat{\Theta} - \Theta + \boldsymbol{\mu}(Q - \hat{Q})^T + (\boldsymbol{\mu} - \bar{\boldsymbol{x}})Q^T]\Phi_0 + O_p(\tfrac{1}{n}).
\end{aligned}$$

By central limit theorem, we have

$$\sqrt{n}\left((\mathrm{vec}(\hat{\Theta} - \Theta))^T, (\hat{Q} - Q)^T, (\bar{\boldsymbol{x}} - \boldsymbol{\mu})^T\right)^T \to N_{2pt+2t+p}\left(\mathbf{0}, \tau = \begin{pmatrix} \Delta_{xy} & \Delta_{xy,y} & \Delta_{xy,x} \\ \Delta_{xy,y}^T & \Delta_y & \Delta_{y,x} \\ \Delta_{xy,x}^T & \Delta_{y,x}^T & \Sigma \end{pmatrix}\right),$$

where the τ will be defined as follows: $\mathrm{Cov}(\mathbf{X}\cos(\boldsymbol{\omega}_j^T\mathbf{Y}), \mathbf{X}\cos(\boldsymbol{\omega}_k^T\mathbf{Y})) = \Delta_{xy}^{j1,k1}$,
$\mathrm{Cov}(\mathbf{X}\cos(\boldsymbol{\omega}_j^T\mathbf{Y}), \mathbf{X}\sin(\boldsymbol{\omega}_k^T\mathbf{Y})) = \Delta_{xy}^{j1,k2}$, $\mathrm{Cov}(\mathbf{X}\sin(\boldsymbol{\omega}_j^T\mathbf{Y}), \mathbf{X}\sin(\boldsymbol{\omega}_k^T\mathbf{Y})) = \Delta_{xy}^{j2,k2}$,
$\mathrm{Cov}(\mathbf{X}\cos(\boldsymbol{\omega}_j^T\mathbf{Y}), \cos(\boldsymbol{\omega}_k^T\mathbf{Y})) = \Delta_{xy,y}^{j1,k1}$, $\mathrm{Cov}(\mathbf{X}\cos(\boldsymbol{\omega}_j^T\mathbf{Y}), \sin(\boldsymbol{\omega}_k^T\mathbf{Y})) = \Delta_{xy,y}^{j1,k2}$,
$\mathrm{Cov}(\mathbf{X}\sin(\boldsymbol{\omega}_j^T\mathbf{Y}), \sin(\boldsymbol{\omega}_k^T\mathbf{Y})) = \Delta_{xy,y}^{j2,k2}$, $\mathrm{Cov}(\mathbf{X}\cos(\boldsymbol{\omega}_j^T\mathbf{Y}), \mathbf{X}) = \Delta_{xy,x}^{j1}$,
$\mathrm{Cov}(\mathbf{X}\sin(\boldsymbol{\omega}_j\mathbf{Y}), \mathbf{X}) = \Delta_{xy,x}^{j2}$, $\mathrm{Cov}(\cos(\boldsymbol{\omega}_j\mathbf{Y}), \cos(\boldsymbol{\omega}_k^T\mathbf{Y})) = \Delta_y^{j1,k1}$,
$\mathrm{Cov}(\cos(\boldsymbol{\omega}_j^T\mathbf{Y}), \sin(\boldsymbol{\omega}_k^T\mathbf{Y})) = \Delta_y^{j1,k2}$, $\mathrm{Cov}(\sin(\boldsymbol{\omega}_j^T\mathbf{Y}), \sin(\boldsymbol{\omega}_k^T\mathbf{Y})) = \Delta_y^{j2,k2}$,
$\mathrm{Cov}(\cos(\boldsymbol{\omega}_j^T\mathbf{Y}), \mathbf{X}) = \Delta_{y,x}^{j1}$, $\mathrm{Cov}(\sin(\boldsymbol{\omega}_j^T\mathbf{Y}), \mathbf{X}) = \Delta_{y,x}^{k2}$,

$$\Delta_{xy} = \begin{array}{c} \\ p \\ p \\ \vdots \\ p \\ p \\ \\ 1 \end{array} \begin{array}{c} \begin{array}{ccccc} p & p & \dots & p & p \end{array} \\ \begin{pmatrix} \Delta_{xy}^{11,11} & \Delta_{xy}^{11,12} & \dots & \Delta_{xy}^{11,t1} & \Delta_{xy}^{11,t2} \\ \Delta_{xy}^{12,11} & \Delta_{xy}^{12,12} & \dots & \Delta_{xy}^{12,t1} & \Delta_{xy}^{12,t2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \Delta_{xy}^{t1,11} & \Delta_{xy}^{t1,12} & \dots & \Delta_{xy}^{t1,t1} & \Delta_{xy}^{t1,t2} \\ \Delta_{xy}^{t2,11} & \Delta_{xy}^{t2,12} & \dots & \Delta_{xy}^{t2,t1} & \Delta_{xy}^{t2,t2} \end{pmatrix} \\ \begin{array}{ccccc} 1 & 1 & \dots & 1 & 1 \end{array} \end{array},$$

$$\Delta_{xy,y} = \begin{array}{c} \\ p \\ p \\ \vdots \\ p \\ p \\ \\ 1 \end{array} \begin{array}{c} \begin{array}{ccccc} p & p & \dots & p & p \end{array} \\ \begin{pmatrix} \Delta_{xy,y}^{11,11} & \Delta_{xy,y}^{11,12} & \dots & \Delta_{xy,y}^{11,t1} & \Delta_{xy,y}^{11,t2} \\ \Delta_{xy,y}^{12,11} & \Delta_{xy,y}^{12,12} & \dots & \Delta_{xy,y}^{12,t1} & \Delta_{xy,y}^{12,t2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \Delta_{xy,y}^{t1,11} & \Delta_{xy,y}^{t1,12} & \dots & \Delta_{xy,y}^{t1,t1} & \Delta_{xy,y}^{t1,t2} \\ \Delta_{xy,y}^{t2,11} & \Delta_{xy,y}^{t2,12} & \dots & \Delta_{xy,y}^{t2,t1} & \Delta_{xy,y}^{t2,t2} \end{pmatrix} \\ \begin{array}{ccccc} 1 & 1 & \dots & 1 & 1 \end{array} \end{array},$$

$$\Delta_y = \begin{array}{c} \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 1 & \dots & 1 & 1 \end{array} \\ \begin{pmatrix} \Delta_y^{11,11} & \Delta_y^{11,12} & \dots & \Delta_y^{11,t1} & \Delta_y^{11,t2} \\ \Delta_y^{12,11} & \Delta_y^{12,12} & \dots & \Delta_y^{12,t1} & \Delta_y^{12,t2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \Delta_y^{t1,11} & \Delta_y^{t1,12} & \dots & \Delta_y^{t1,t1} & \Delta_y^{t1,t2} \\ \Delta_y^{t2,11} & \Delta_y^{t2,12} & \dots & \Delta_y^{t2,t1} & \Delta_y^{t2,t2} \end{pmatrix} \end{array},$$

$$\Delta_{xy,x} = \begin{array}{c} \\ p \\ p \\ \vdots \\ p \\ p \end{array} \begin{array}{c} p \\ \begin{pmatrix} \Delta_{xy,x}^{11} \\ \Delta_{xy,x}^{12} \\ \vdots \\ \Delta_{xy,x}^{t1} \\ \Delta_{xy,x}^{t2} \end{pmatrix} \end{array}, \Delta_{y,x} = \begin{array}{c} \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{array} \begin{array}{c} p \\ \begin{pmatrix} \Delta_{y,x}^{11} \\ \Delta_{y,x}^{12} \\ \vdots \\ \Delta_{y,x}^{t1} \\ \Delta_{y,x}^{t2} \end{pmatrix} \end{array}.$$

$$\text{Let } A = \begin{array}{c} \\ p \\ p \\ \vdots \\ p \\ p \end{array} \begin{array}{c} \begin{array}{cccccccccc} p & p & \dots & p & p & 1 & 1 & \dots & 1 & 1 \end{array} \quad p \\ \begin{pmatrix} I_p & 0 & \dots & 0 & 0 & \boldsymbol{\mu} & 0 & \dots & 0 & 0 & \mathrm{E}\cos(\omega_1 Y) I_p \\ 0 & I_p & \dots & 0 & 0 & 0 & \boldsymbol{\mu} & \dots & 0 & 0 & \mathrm{E}\sin(\omega_1 Y) I_p \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & I_p & 0 & 0 & 0 & \dots & \boldsymbol{\mu} & 0 & \mathrm{E}\cos(\omega_t Y) I_p \\ 0 & 0 & \dots & 0 & I_p & 0 & 0 & \dots & 0 & \boldsymbol{\mu} & \mathrm{E}\sin(\omega_t Y) I_p \end{pmatrix} \end{array}.$$

Then, $\sqrt{n}\mathrm{vec}[\hat{\Theta} - \Theta + \boldsymbol{\mu}(Q - \hat{Q})^T + (\boldsymbol{\mu} - \bar{\boldsymbol{x}})Q^T] \sim N_{2pt}(\mathbf{0}, A\tau A^T)$.

Hence, $\sqrt{n}\mathrm{vec}\{\Gamma_0^T \Sigma^{-1/2}[\hat{\Theta} - \Theta + \boldsymbol{\mu}(Q - \hat{Q})^T + (\boldsymbol{\mu} - \bar{\boldsymbol{x}})Q^T]\Phi_0\} = (\Phi_0^T \otimes \Gamma_0^T \Sigma^{-1/2})\sqrt{n}\mathrm{vec}[\hat{\Theta} - \Theta + \boldsymbol{\mu}(Q - \hat{Q})^T + (\boldsymbol{\mu} - \bar{\boldsymbol{x}})Q^T]$, which has normal distribution $N_{(2t-d)\times(p-d)}(\mathbf{0}, \Omega = (\Phi_0^T \otimes \Gamma_0^T \Sigma^{-1/2}) A\tau A^T (\Phi_0 \otimes \Sigma^{-1/2}\Gamma_0))$. Let $\Psi_0 = \Gamma_0^T \hat{\Psi} \Phi_0$, then $\hat{\Lambda}_d = n\mathrm{trace}(\Psi_0 \Psi_0^T) = n\mathrm{vec}(\Psi_0)^T \mathrm{vec}(\Psi_0)$. Because Ω is a positive definite matrix, there exist an orthogonal matrix P and diagonal matrix D such that $\Omega = P^T D P$. Because $\sqrt{n}\mathrm{vec}(\Psi_0) \sim N(\mathbf{0}, \Omega)$, then $\sqrt{n} P \mathrm{vec}(\Psi_0) \sim N(\mathbf{0}, D)$. So $n\mathrm{vec}(\Psi_0)^T \mathrm{vec}(\Psi_0) \sim \sum_{k=1}^{(p-d)(2t-d)} \lambda_k C_k$, where the C_ks are independent chi-square random variables with one degree of freedom and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{(p-d)(2t-d)}$ are eigenvalues of the covariance matrix Ω.

Remark: When $\mathbf{X}$ follows multivariate normal distribution, Proposition 2 will not result in a chi-square distribution. In fact, $\hat{\psi}_{j1} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{Z}_k \cos(\boldsymbol{\omega}_j^T \mathbf{y}_k)$ is not a linear transformation of normal distribution. Hence, it is not a normal distribution. However, because the sample mean in different slices Li (1991) is independent normal under assumption, we expect that the test statistic follows chi-square distribution. □

Proof of Proposition 3. The proof is similar to the proof of Proportion 4.2 of Chiaromonte et al. (2002). First, fix the number of transformations $\{t_k\}_{k=1}^K$ and choose $\{\boldsymbol{\omega}_j^{(k)}\}_{j=1}^{t_k}$ for each level. For each level k, $j = 1, \cdots, t_k$, define:

$\hat{\psi}_{j1}^{(k)} = \frac{1}{n_k}\sum_{l=1}^{n_k} \hat{\mathbf{z}}_l^{(k)} \cos(\boldsymbol{\omega}_j^{(k)T}\mathbf{y}_l^{(k)}), \qquad \psi_{j1}^{(k)} = \mathrm{E}[\mathbf{Z}\cos(\boldsymbol{\omega}_j^{(k)T}\mathbf{Y})]$,

$\hat{\psi}_{j2}^{(k)} = \frac{1}{n_k}\sum_{l=1}^{n_k} \hat{\mathbf{z}}_l^{(k)} \sin(\boldsymbol{\omega}_j^{(k)T}\mathbf{y}_l^{(k)}), \qquad \psi_{j2}^{(k)} = \mathrm{E}[\mathbf{Z}\sin(\boldsymbol{\omega}_j^{(k)T}\mathbf{Y})]$.

Let $\hat{\Psi}_k = (\hat{\psi}_{11}^{(k)}, \hat{\psi}_{12}^{(k)}, \cdots, \hat{\psi}_{t1}^{(k)}, \hat{\psi}_{t2}^{(k)})$ and $\Psi_k = (\psi_{11}^{(k)}, \psi_{12}^{(k)}, \cdots, \psi_{t1}^{(k)}, \psi_{t2}^{(k)})$. Let $\hat{f}_k = \sqrt{\frac{n_k}{n}}$, $\Psi^W = (f_1\Psi_1, \cdots, f_K\Psi_K)$ and $\hat{\Psi}^W = (\hat{f}_1\hat{\Psi}_1, \cdots, \hat{f}_K\hat{\Psi}_K)$. For each level k, $j = 1, \cdots, t_k$, define:

$\hat{\theta}_{j1}^{(k)} = \frac{1}{n_k}\sum_{l=1}^{n_k} \mathbf{x}_l^{(k)} \cos(\boldsymbol{\omega}_j^{(k)T}\mathbf{y}_l^{(k)}), \qquad \theta_{j1}^{(k)} = \mathrm{E}[\mathbf{X}\cos(\boldsymbol{\omega}_j^{(k)T}\mathbf{Y})]$,

$\hat{\theta}_{j2}^{(k)} = \frac{1}{n_k}\sum_{l=1}^{n_k} \mathbf{x}_l^{(k)} \sin(\boldsymbol{\omega}_j^{(k)T}\mathbf{y}_l^{(k)}), \qquad \theta_{j2}^{(k)} = \mathrm{E}[\mathbf{X}\sin(\boldsymbol{\omega}_j^{(k)T}\mathbf{Y})]$.

Let $\hat{\Theta}_k = \left(\hat{\theta}_{11}^{(k)}, \hat{\theta}_{12}^{(k)}, \cdots, \hat{\theta}_{t1}^{(k)}, \hat{\theta}_{t2}^{(k)}\right)$ and $\Theta_k = \left(\theta_{11}^{(k)}, \theta_{12}^{(k)}, \cdots, \theta_{t1}^{(k)}, \theta_{t2}^{(k)}\right)$. Set $\hat{Q}_k = \left(\frac{1}{n_k}\sum_{l=1}^{n_k}\cos(\boldsymbol{\omega}_1^{(k)T}\mathbf{y}_l^{(k)}), \frac{1}{n_k}\sum_{l=1}^{n_k}\sin(\boldsymbol{\omega}_1^{(k)T}\mathbf{y}_l^{(k)}), \cdots, \frac{1}{n_k}\sum_{l=1}^{n_k}\cos(\boldsymbol{\omega}_t^{(k)T}\mathbf{y}_l^{(k)}), \frac{1}{n_k}\sum_{l=1}^{n_k}\sin(\boldsymbol{\omega}_t^{(k)T}\mathbf{y}_l^{(k)})\right.$ and $Q_k = \left(\mathrm{E}(\cos\boldsymbol{\omega}_1^{(k)T}\mathbf{Y}), \mathrm{E}(\sin\boldsymbol{\omega}_1^{(k)T}\mathbf{Y}), \cdots, \mathrm{E}(\cos\boldsymbol{\omega}_t^{(k)T}\mathbf{Y}), \mathrm{E}(\sin\boldsymbol{\omega}_t^{(k)T}\mathbf{Y})\right)^T$. Then by SVD, $\Psi^W = \boldsymbol{\Gamma}^T \begin{bmatrix} \boldsymbol{D} & 0 \\ 0 & 0 \end{bmatrix} \boldsymbol{\Phi}$, where $\boldsymbol{\Gamma}$ and $\boldsymbol{\Phi}$ are respective $p\times p$ and $2\sum t_k \times 2\sum t_k$ orthogonal matrices and $\boldsymbol{D}$ is a $d \times d$ diagonal matrix of positive values. Let $\boldsymbol{\Gamma}^T = (\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_0)$ and $\boldsymbol{\Phi}^T = (\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_0)$, where $\boldsymbol{\Gamma}_0$ is $p \times (p-d)$ and $\boldsymbol{\Phi}_0$ is $2\sum t_k \times (2\sum t_k - d)$. Thus $\hat{\Lambda}_d^W = n \times \mathrm{trace}[(\boldsymbol{\Gamma}_0^T\hat{\Psi}^W\boldsymbol{\Phi}_0)(\boldsymbol{\Gamma}_0^T\hat{\Psi}^W\boldsymbol{\Phi}_0)^T] = n\ \mathrm{vec}(\boldsymbol{\Gamma}_0^T\hat{\Psi}^W\boldsymbol{\Phi}_0)^T\ \mathrm{vec}(\boldsymbol{\Gamma}_0^T\hat{\Psi}^W\boldsymbol{\Phi}_0)$. Partition $\boldsymbol{\Phi}_0 = (\boldsymbol{\Phi}_{01}^T, \cdots, \boldsymbol{\Phi}_{0K}^T)^T$, where $\boldsymbol{\Phi}_{0k}$ has dimension $2t_k \times (2\sum t_k - d)$. Then $\sqrt{n}\boldsymbol{\Gamma}_0^T\hat{\Psi}^W\boldsymbol{\Phi}_0 = \sqrt{n}\boldsymbol{\Gamma}_0^T(\sum_{k=1}^K \hat{f}_k\hat{\Psi}_k\boldsymbol{\Phi}_{0k}) = \sum_{k=1}^K \sqrt{n_k}\boldsymbol{\Gamma}_0^T\hat{\Psi}_k\boldsymbol{\Phi}_{0k}$. As $\boldsymbol{\Gamma}_0^T\Psi^W\boldsymbol{\Phi}_0 = 0$, that is $\boldsymbol{\Gamma}_0^T\Psi^W\boldsymbol{\Phi}_0 = \boldsymbol{\Gamma}_0^T\sum_{k=1}^K f_k\Psi_k\boldsymbol{\Phi}_{0k} = \boldsymbol{\Gamma}_0^T\sum_{k=1}^K f_k\Sigma^{-1/2}(\Theta_k - \boldsymbol{\mu}Q_k)\boldsymbol{\Phi}_{0k} = 0$.

Define $\hat{A} = \hat{\Sigma}^{-1/2}\Sigma^{1/2}$, then

$$\begin{aligned}
\sqrt{n}\Gamma_0^T\hat{\Psi}^W\boldsymbol{\Phi}_0 &= \sqrt{n}\Gamma_0^T\hat{\Sigma}^{-1/2}\textstyle\sum_{k=1}^K \hat{f}_k(\hat{\Theta}_k - \bar{\boldsymbol{x}}^{(k)}\hat{Q}_k^T)\boldsymbol{\Phi}_{0k} \\
&= \sqrt{n}\Gamma_0^T\hat{A}\Sigma^{-1/2}\textstyle\sum_{k=1}^K \hat{f}_k(\hat{\Theta}_k - \bar{\boldsymbol{x}}^{(k)}\hat{Q}_k^T)\boldsymbol{\Phi}_{0k} \\
&= \sqrt{n}\Gamma_0^T\hat{A}\Sigma^{-1/2}\textstyle\sum_{k=1}^K \frac{\hat{f}_k}{f_k} f_k(\hat{\Theta}_k - \bar{\boldsymbol{x}}^{(k)}\hat{Q}_k^T)\boldsymbol{\Phi}_{0k} \\
&= \sqrt{n}\Gamma_0^T(\hat{A} - I + I)\Sigma^{-1/2}\textstyle\sum_{k=1}^K (\frac{\hat{f}_k}{f_k} - 1 + 1) f_k(\hat{\Theta}_k - \bar{\boldsymbol{x}}^{(k)}\hat{Q}_k^T)\boldsymbol{\Phi}_{0k} \\
&= \sqrt{n}\Gamma_0^T(\hat{A} - I)\Sigma^{-1/2}\textstyle\sum_{k=1}^K f_k(\hat{\Theta}_k - \bar{\boldsymbol{x}}^{(k)}\hat{Q}_k^T)\boldsymbol{\Phi}_{0k} \\
&\quad + \sqrt{n}\Gamma_0^T\Sigma^{-1/2}\textstyle\sum_{k=1}^K \hat{f}_k(\hat{\Theta}_k - \bar{\boldsymbol{x}}^{(k)}\hat{Q}_k^T)\boldsymbol{\Phi}_{0k} + O_p(\frac{1}{n}).
\end{aligned}$$

So $\hat{\Theta}_k - \bar{\boldsymbol{x}}^{(k)}\hat{Q}_k^T = \hat{\Theta}_k - \Theta_k + \Theta_k - \boldsymbol{\mu}Q_k^T + \boldsymbol{\mu}(Q_k^T - \hat{Q}_k^T) + (\boldsymbol{\mu} - \bar{\boldsymbol{x}}^{(k)})\hat{Q}_k^T$. Then we use $(\hat{A} - I)\Sigma^{-1/2}(\hat{\Theta}_k - \Theta_k) = O_p(\frac{1}{n_k})$, $(\hat{A} - I)\Sigma^{-1/2}\boldsymbol{\mu}(Q_k^T - \hat{Q}_k^T) = O_p(\frac{1}{n_k})$, $(\hat{A} - I)\Sigma^{-1/2}(\boldsymbol{\mu} - \bar{\boldsymbol{x}}_{(k)})\hat{Q}_k^T = O_p(\frac{1}{n})$ and $\sum_{k=1}^K f_k\Sigma^{-1/2}(\Theta_k - \boldsymbol{\mu}Q_k)\boldsymbol{\Phi}_{0k} = 0$.

$$\begin{aligned}
\sqrt{n}\Gamma_0^T\hat{\Psi}^W\boldsymbol{\Phi}_0 &= \sqrt{n}\Gamma_0^T\Sigma^{-1/2}\textstyle\sum_{k=1}^K \hat{f}_k(\hat{\Theta}_k - \Theta_k + \boldsymbol{\mu}(Q_k^T - \hat{Q}_k^T) + (\boldsymbol{\mu} - \bar{\boldsymbol{x}}^{(k)})\hat{Q}_k^T)\boldsymbol{\Phi}_{0k} + O_p(\tfrac{1}{n}). \\
&= \Gamma_0^T\Sigma^{-1/2}\textstyle\sum_{k=1}^K \sqrt{n_k}(\hat{\Theta}_k - \Theta_k + \boldsymbol{\mu}(Q_k^T - \hat{Q}_k^T) + (\boldsymbol{\mu} - \bar{\boldsymbol{x}}^{(k)})\hat{Q}_k^T)\boldsymbol{\Phi}_{0k} + O_p(\tfrac{1}{n}).
\end{aligned}$$

Let $\Omega_k = (\mathbf{\Phi}_{0k}^T \otimes \mathbf{\Gamma}_0^T \Sigma^{-1/2}) A_k \tau_k A_k^T (\mathbf{\Phi}_{0k} \otimes \Sigma^{-1/2} \mathbf{\Gamma}_0)$ and τ_k are defined in Proposition 2. Then $\sqrt{n}\mathbf{\Gamma}_0^T \hat{\Psi}^W \mathbf{\Phi}_0 \sim N(\mathbf{0}, \sum \Omega_k)$. Furthermore, the rank for $\sum \Omega_k$ is $(p-d)(2\sum t_k - Kd)$. So $\hat{\Lambda}_d^W \sim \sum_{i=1}^{(p-d)(2\sum t_k - Kd)} \lambda_i C_i$, where the C_is are independent chi-square random variables, each with one degree of freedom, and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{(p-d)(2\sum t_k - Kd)}$ are eigenvalues of the covariance matrix $\Omega^W = \sum \Omega_k$. □

Appendix B: Proof for Chapter 3

Proof of Theorem 1: Let $\hat{\boldsymbol{\varphi}}_{\boldsymbol{\omega}} = \frac{1}{n}\sum_{k=1}^n e^{i\boldsymbol{\omega}^T \mathbf{y}_k} \mathbf{x}_k$, $\boldsymbol{\varphi}_{\boldsymbol{\omega}} = \mathrm{E}(e^{i\boldsymbol{\omega}^T \mathbf{Y}} \mathbf{X})$, $\mathbf{C}_{\boldsymbol{\omega}} = \mathrm{E}(e^{i\boldsymbol{\omega}^T \mathbf{Y}})$, $\hat{\mathbf{C}}_{\boldsymbol{\omega}} = \frac{1}{n}\sum_{k=1}^n e^{i\boldsymbol{\omega}^T \mathbf{y}_k}$, and $\mathbf{Z}_k = \Sigma^{-1/2}(\mathbf{x}_k - \boldsymbol{\mu})$. We need a lemma which is in the Appendix B: Lemma A.2 of Cook and Ni (2005).

Lemma 3. *Li et al. (2003) Suppose that a random vector* $\mathbf{X}$ *has covariance matrix* $\Sigma > 0$. *Then*

$$\hat{\Sigma}^{-1} - \Sigma^{-1} = -n^{-1}\Sigma^{-1/2} \sum_{k=1}^n (\mathbf{Z}_k \mathbf{Z}_k^T - I)\Sigma^{-1/2} + O_p(n^{-1}).$$

Consider

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\xi}}_{\boldsymbol{\omega}} - \boldsymbol{\xi}_{\boldsymbol{\omega}}) &= \sqrt{n}\hat{\Sigma}^{-1}(\hat{\boldsymbol{\varphi}}_{\boldsymbol{\omega}} - \hat{\mathbf{C}}_{\boldsymbol{\omega}}\bar{\mathbf{x}}) - \sqrt{n}\Sigma^{-1}(\boldsymbol{\varphi}_{\boldsymbol{\omega}} - \mathbf{C}_{\boldsymbol{\omega}}\boldsymbol{\mu}) \\ &= \sqrt{n}(\hat{\Sigma}^{-1} - \Sigma^{-1})(\boldsymbol{\varphi}_{\boldsymbol{\omega}} - \mathbf{C}_{\boldsymbol{\omega}}\boldsymbol{\mu}) \\ &\quad + \sqrt{n}\Sigma^{-1}[(\hat{\boldsymbol{\varphi}}_{\boldsymbol{\omega}} - \hat{\mathbf{C}}_{\boldsymbol{\omega}}\bar{\mathbf{x}}) - (\boldsymbol{\varphi}_{\boldsymbol{\omega}} - \mathbf{C}_{\boldsymbol{\omega}}\boldsymbol{\mu})] + O_p(n^{-1/2}). \end{aligned} \tag{4.8}$$

The first item can be written as:

$$\begin{aligned} \sqrt{n}(\hat{\Sigma}^{-1} - \Sigma^{-1})(\boldsymbol{\varphi}_{\boldsymbol{\omega}} - \mathbf{C}_{\boldsymbol{\omega}}\boldsymbol{\mu}) &= -n^{-1/2}\Sigma^{-1/2}\textstyle\sum_{k=1}^n (\mathbf{Z}_k \mathbf{Z}_k^T - I)\Sigma^{-1/2}(\boldsymbol{\varphi}_{\boldsymbol{\omega}} - \mathbf{C}_{\boldsymbol{\omega}}\boldsymbol{\mu}) + O_p(n^{-1/2}) \\ &= -n^{-1/2}\Sigma^{-1/2}\textstyle\sum_{k=1}^n (\mathbf{Z}_k \mathbf{Z}_k^T - I)\mathrm{E}(e^{i\boldsymbol{\omega}^T \mathbf{Y}}\mathbf{Z}) + O_p(n^{-1/2}). \end{aligned} \tag{4.9}$$

Then

$$\begin{aligned} \hat{\boldsymbol{\varphi}}_{\boldsymbol{\omega}} - \hat{\mathbf{C}}_{\boldsymbol{\omega}}\bar{\mathbf{x}} &= \textstyle\frac{1}{n}\sum_{k=1}^n e^{i\boldsymbol{\omega}^T \mathbf{y}_k}\mathbf{x}_k - \frac{1}{n}\sum_{k=1}^n e^{i\boldsymbol{\omega}^T \mathbf{y}_k}\bar{\mathbf{x}} \\ &= \textstyle\frac{1}{n}\sum_{k=1}^n (e^{i\boldsymbol{\omega}^T \mathbf{y}_k} - \mathrm{E}e^{i\boldsymbol{\omega}^T \mathbf{Y}})(\mathbf{x}_k - \boldsymbol{\mu}) - \frac{1}{n}(\bar{\mathbf{x}} - \boldsymbol{\mu})\sum_{k=1}^n (e^{i\boldsymbol{\omega}^T \mathbf{y}_k} - \mathrm{E}e^{i\boldsymbol{\omega}^T \mathbf{Y}}) \\ &= \textstyle\frac{1}{n}\sum_{k=1}^n (e^{i\boldsymbol{\omega}^T \mathbf{y}_k} - \mathrm{E}e^{i\boldsymbol{\omega}^T \mathbf{Y}})(\mathbf{x}_k - \boldsymbol{\mu}) + O_p(n^{-1}). \end{aligned}$$

Therefore, the second term can be simplified as

$$\begin{aligned}\sqrt{n}\Sigma^{-1}[(\hat{\boldsymbol{\varphi}}_{\boldsymbol{\omega}} - \hat{\mathbf{C}}_{\boldsymbol{\omega}}\bar{\mathbf{x}}) - (\boldsymbol{\varphi}_{\boldsymbol{\omega}} - \mathbf{C}_{\boldsymbol{\omega}}\boldsymbol{\mu})] &= n^{-1/2}\Sigma^{-1/2}\textstyle\sum_{k=1}^{n}[\Sigma^{-1/2}(e^{i\boldsymbol{\omega}^T\mathbf{y}_k} - \mathrm{E}e^{i\boldsymbol{\omega}^T\mathbf{Y}})(\mathbf{x}_k - \boldsymbol{\mu})] \\ &\quad -\sqrt{n}\Sigma^{-1}(\boldsymbol{\varphi}_{\boldsymbol{\omega}} - \mathbf{C}_{\boldsymbol{\omega}}\boldsymbol{\mu}) + O_p(n^{-1/2}) \\ &= n^{-1/2}\Sigma^{-1/2}\textstyle\sum_{k=1}^{n}[\mathbf{Z}_k(e^{i\boldsymbol{\omega}^T\mathbf{y}_k} - \mathrm{E}e^{i\boldsymbol{\omega}^T\mathbf{Y}}) - \mathrm{E}(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\mathbf{Z})] \\ &\quad +O_p(n^{-1/2}).\end{aligned} \tag{4.10}$$

Then we put equation (4.9) and (4.10) into (4.8):

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\xi}}_{\boldsymbol{\omega}} - \boldsymbol{\xi}_{\boldsymbol{\omega}}) &= \sqrt{n}\hat{\Sigma}^{-1}(\hat{\boldsymbol{\varphi}}_{\boldsymbol{\omega}} - \hat{\mathbf{C}}_{\boldsymbol{\omega}}\bar{\mathbf{x}}) - \sqrt{n}\Sigma^{-1}(\boldsymbol{\varphi}_{\boldsymbol{\omega}} - \mathbf{C}_{\boldsymbol{\omega}}\boldsymbol{\mu}) \\ &= n^{-1/2}\Sigma^{-1/2}\textstyle\sum_{k=1}^{n}[\mathbf{Z}_k(e^{i\boldsymbol{\omega}^T\mathbf{y}_k} - \mathrm{E}e^{i\boldsymbol{\omega}^T\mathbf{Y}}) - \mathrm{E}(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\mathbf{Z}) - (\mathbf{Z}_k\mathbf{Z}_k^T - I)\mathrm{E}(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\mathbf{Z})] \\ &\quad +O_p(n^{-1/2}) \\ &= n^{-1/2}\Sigma^{-1/2}\textstyle\sum_{k=1}^{n}\{\mathbf{Z}_k[e^{i\boldsymbol{\omega}^T\mathbf{y}_k} - \mathrm{E}e^{i\boldsymbol{\omega}^T\mathbf{Y}} - \mathbf{Z}_k^T\mathrm{E}(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\mathbf{Z})]\} + O_p(n^{-1/2}) \\ &= n^{-1/2}\Sigma^{-1/2}\textstyle\sum_{k=1}^{n}\mathbf{Z}_k\varepsilon_{\boldsymbol{\omega},k} + O_p(n^{-1/2}),\end{aligned}$$

where $\varepsilon_{\boldsymbol{\omega},k} = e^{i\boldsymbol{\omega}^T\mathbf{y}_k} - \mathrm{E}e^{i\boldsymbol{\omega}^T\mathbf{Y}} - \mathbf{Z}_k^T\mathrm{E}(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\mathbf{Z})$. Let $\boldsymbol{\varepsilon}_k = (\varepsilon^R_{\boldsymbol{\omega}_1,k}, \varepsilon^I_{\boldsymbol{\omega}_1,k}, ..., \varepsilon^R_{\boldsymbol{\omega}_m,k}, \varepsilon^I_{\boldsymbol{\omega}_m,k})^T$, where $k = 1, ..., n$.

Then we have

$$\sqrt{n}[\mathrm{vec}(\hat{\boldsymbol{\xi}}) - \mathrm{vec}(\beta\nu)] = n^{-1/2}\sum_{k=1}^{n}(\Sigma^{-1/2}\mathbf{Z}_k\boldsymbol{\varepsilon}_k^T) + O_p(n^{-1/2}).$$

Thus

$$\sqrt{n}[\mathrm{vec}(\hat{\boldsymbol{\xi}}) - \mathrm{vec}(\beta\nu)] \xrightarrow{D} N(0, \Gamma),$$

where $\Gamma = \mathrm{Cov}[\mathrm{vec}(\Sigma^{-1/2}Z\boldsymbol{\epsilon}^T)] \in \mathbb{R}^{2pm\times 2pm}$. □

Proof of Theorem 2: Because $\hat{\Gamma}$ converges to Γ in probability, the asymptotic distribution of $n\hat{F}_d$ is the same as that of $n\hat{H}_d$ using the Lemma A.3 on the Cook and Ni (2005)'s Appendix C, where

$$H_d(B, C) = [\mathrm{vec}(\hat{\boldsymbol{\xi}}) - \mathrm{vec}(BC)]^T\Gamma^{-1}[\mathrm{vec}(\hat{\boldsymbol{\xi}}) - \mathrm{vec}(BC)].$$

Again, we also need to show the asymptotic distribution of $\mathrm{vec}(\hat{\beta}\hat{\nu})$ of $F_d(B, C)$ is the same as that of $H_d(B, C)$. Similar to Cook and Ni (2005), we can show that there is one parameterization that satisfies the conditions in the statement of Lemma

A.4 of Cook and Ni (2005). Let $\beta = (\beta_1^T, \beta_2^T)^T$, where $\beta_1 \in \mathbb{R}^{d\times d}$, $\beta_2 \in \mathbb{R}^{(p-d)\times d}$. We can assume that $\beta_1 = I_d$, which is nonsingular. The new parameterization brings a full-rank Jacobian matrix and an open parameter space in $\mathbb{R}^{d(2m+p-d)}$.

Let

$$\boldsymbol{\theta} = \begin{pmatrix} \text{vec}(B) \\ \text{vec}(C) \end{pmatrix} \in \mathbb{R}^{d(p+2m)},$$

$$g(\boldsymbol{\theta}) = \text{vec}(BC) \in \mathbb{R}^{2pm},$$

$$\boldsymbol{\tau}_n = \text{vec}(\hat{\xi}),$$

and

$$g(\boldsymbol{\theta}_0) = \text{vec}(\beta\nu).$$

Using the Proposition A.1 in Cook and Ni (2005) by checking all these conditions, we can get

$$\sqrt{n}[\text{vec}(\hat{\beta}\hat{\nu}) - \text{vec}(\beta\nu)] \xrightarrow{D} N[0, \Delta(\Delta^T\Gamma^{-1}\Delta)^-\Delta^T],$$

which is the conclusion 3 of the Proposition A.1.

And $n\hat{H}_d \xrightarrow{D} \chi^2_k$, where $k = 2pm - rank(\Delta)$ and $rank(\Delta) = d(p-d)+2md$. Hence $k = (2m-d)(p-d)$. Hence the conclusion 2 is proved. The consistency of Span$(\hat{\beta})$ follows directly from conclusion 1. □

Proof of Theorem 3: Follow the proof in Appendix D in Cook and Ni (2005). Let $V = \Gamma_D^{-1} = \text{diag}\{\Gamma_l^{-1}\}$ and $V_n = \hat{\Gamma}_D^{-1} = \text{diag}\{\hat{\Gamma}_l^{-1}\}$ The discrepancy function $F_d(B, C, \hat{\Gamma}_D^{-1})$ can be written as

$$F_d(B, C; V_n) = [\text{vec}(\hat{\xi}) - \text{vec}(BC)]^T V_n [\text{vec}(\hat{\xi}) - \text{vec}(BC)].$$

Because $\widetilde{V}_n$ converges to V, it follows from Lemma A.3 in Cook and Ni (2005) that the asymptotic distribution of $n\hat{F}_d$ is the same as that of $n\hat{H}_d$, where

$$H_d(B, C; V) = [\text{vec}(\hat{\xi}) - \text{vec}(BC)]^T V [\text{vec}(\hat{\xi}) - \text{vec}(BC)].$$

Theorem 3 can be proven in the same way as theorem 1. From the conclusion 1 from Proposition A.1 in Cook and Ni (2005), the asymptotic distribution of $n\hat{F}_d$ is the same as that of $||Q_\Phi V^{1/2} W||^2$, where W is normal with mean 0 and covariance

matrix Γ and $\Phi = V^{1/2}\Delta$. Consequently, $n\hat{F}_d$ is asymptotically distributed as a linear combination of independent chi-squared random variables each with 1 degree of freedom. The coefficient of the chi-squared variables are the eigenvalues of $Q_\Phi \Omega Q_\Phi$, where $\Omega = V^{1/2}\Gamma V^{1/2}$. What's more, the dimension of $\dim(Q_\Phi \Omega Q_\Phi)$ =$\dim(Q_\Phi)$ = $2pm-\dim(\Delta) = (p-d)(2m-d)$.

Finally, consistency follows from the conclusion 3 of Proposition A.1 in Cook and Ni (2005) in combination with Lemma A.3.

□

Proof of Lemma 1: Firstly, we know that $\text{Span}(\hat{\beta})$ is a consistent estimator of $\sum_{j=1}^{m} \text{Span}\{\xi_j\}$ by Theorem 3 part 1. Under coverage condition $\sum_{j=1}^{m} \text{Span}\{\xi_j\} = \mathcal{S}_{Y|\mathbf{X}}$. Secondly, under the linearity condition, $\text{Span}(\hat{u}_1, \cdots, \hat{u}_d) \subseteq \mathcal{S}_{Y|\mathbf{X}}$. All the $\text{Span}(\hat{\beta})$, $\text{Span}(\hat{u}_1, \cdots, \hat{u}_d)$ and $\mathcal{S}_{Y|\mathbf{X}}$ have dimension d. Hence, $\text{Span}(\hat{\beta}) = \text{Span}(\hat{u}_1, \cdots, \hat{u}_d)$. □

Proof of Theorem 4: The Proof of this theorem is similar to the theorem 3, but just replace $V = \text{diag}\{\Sigma\}$ and $\hat{V} = \text{diag}\{\hat{\Sigma}\}$. □

Proof of Theorem 5: Consider for fix $\boldsymbol{\omega}$,

$$\sqrt{n}(\widetilde{\boldsymbol{\xi}}_{\boldsymbol{\omega}} - \boldsymbol{\xi}_{\boldsymbol{\omega}}) \;\; = \;\; \sqrt{n}\Sigma^{-1}(\hat{\boldsymbol{\varphi}}_{\boldsymbol{\omega}} - \hat{\mathbf{C}}_{\boldsymbol{\omega}}\bar{\mathbf{x}}) - \sqrt{n}\Sigma^{-1}(\boldsymbol{\varphi}_{\boldsymbol{\omega}} - \mathbf{C}_{\boldsymbol{\omega}}\boldsymbol{\mu}).$$

Therefore,

$$\sqrt{n}\Sigma^{-1}[(\hat{\boldsymbol{\varphi}}_{\boldsymbol{\omega}} - \hat{\mathbf{C}}_{\boldsymbol{\omega}}\bar{\mathbf{x}}) - (\boldsymbol{\varphi}_{\boldsymbol{\omega}} - \mathbf{C}_{\boldsymbol{\omega}}\boldsymbol{\mu})] \;\; = \;\; n^{-1/2}\Sigma^{-1/2}\textstyle\sum_{k=1}^{n}[\mathbf{z}_k(e^{i\boldsymbol{\omega}^T\mathbf{y}_k} - \mathrm{E}e^{i\boldsymbol{\omega}^T\mathbf{Y}}) - \mathrm{E}(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\mathbf{Z})].$$

Thus, $\sqrt{n}[\text{vec}(\widetilde{\xi}) - \text{vec}(\beta\nu)] \xrightarrow{D} N(0, \widetilde{\Gamma})$. □

Proof of Theorem 7: The Proof of this theorem will be exactly similar to the theorem 3, but just replace $V = \text{diag}\{\widetilde{\Gamma}_1^{-1}, ..., \widetilde{\Gamma}_K^{-1}\}$ and $\hat{V} = \text{diag}\{\widetilde{G}_1^{-1}, ..., \widetilde{G}_K^{-1}\}$, and replace Γ with $\widetilde{\Gamma}$. □

Proof of Theorem 9: This proof follows the idea of Theorem 6 in the Qian et al. (2018). Let (B_0, C_0) be a minimizer of (4.7) with $B_0 = (B_{01}, \cdots, B_{0p})^T$ and the constraint that $C_0 C_0^T = I_d$. Let $\hat{B} = (\hat{B}_1, \cdots, \hat{B}_p)^T$ and $\hat{C}$ be a minimizer of (3.6). Then $L_n(\hat{B}, \hat{C}) \leq L_n(B_0, C_0)$. Here we assume that $\boldsymbol{\epsilon}$ is independent of $\mathbf{Z}$.

$\hat{\Lambda} = \widehat{\mathrm{E}^{-1}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)} \otimes \hat{\Sigma}^{-1}$. After simplification, we have

$$
\begin{aligned}
& -[\mathrm{vec}(\Upsilon_n) - \mathrm{vec}(\Sigma B_0 C_0)]^T [\widehat{\mathrm{E}^{-1}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)} \otimes I_p][\mathrm{vec}(\hat{B}\hat{C}) - \mathrm{vec}(B_0 C_0)] \\
& + \mathrm{vec}(B_0 C_0)^T (I_{2m} \otimes \hat{\Sigma} - I_{2m} \otimes \Sigma)[\widehat{\mathrm{E}^{-1}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)} \otimes I_p][\mathrm{vec}(\hat{B}\hat{C}) - \mathrm{vec}(B_0 C_0)] \\
& + \tfrac{1}{2}[\mathrm{vec}(\hat{B}\hat{C}) - \mathrm{vec}(B_0 C_0)]^T [\widehat{\mathrm{E}^{-1}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)} \otimes \hat{\Sigma}][\mathrm{vec}(\hat{B}\hat{C}) - \mathrm{vec}(B_0 C_0)] + \lambda \textstyle\sum_{j=1}^p w_j ||\hat{B}_j||_2 \\
\leq \; & \lambda \textstyle\sum_{j=1}^p w_j ||B_{0j}||_2.
\end{aligned}
$$

Then define

$$
\begin{aligned}
D_1 &= -[\mathrm{vec}(\Upsilon_n) - \mathrm{vec}(\Sigma B_0 C_0)]^T [\widehat{\mathrm{E}^{-1}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)} \otimes I_p][\mathrm{vec}(\hat{B}\hat{C}) - \mathrm{vec}(B_0 C_0)] \\
D_2 &= \mathrm{vec}(B_0 C_0)^T (I_{2m} \otimes \hat{\Sigma} - I_{2m} \otimes \Sigma)[\widehat{\mathrm{E}^{-1}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)} \otimes I_p][\mathrm{vec}(\hat{B}\hat{C}) - \mathrm{vec}(B_0 C_0)] \\
D_3 &= [\mathrm{vec}(\hat{B}\hat{C}) - \mathrm{vec}(B_0 C_0)]^T [\widehat{\mathrm{E}^{-1}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)} \otimes \hat{\Sigma}][\mathrm{vec}(\hat{B}\hat{C}) - \mathrm{vec}(B_0 C_0)]
\end{aligned}
$$

We will find upper bound for D_1 and D_2 at first. We assume $\mathrm{E}(\mathbf{X}) = \mathbf{0}$. Because $X_j (1 \leq j \leq p)$ satisfies sub-Gaussian distribution (C1), $e^{i\boldsymbol{\omega}\mathbf{Y}} X_j$ also follow sub-Gaussian, which means that there exist constants $v, C_1 > 0$ such that for every $k \geq 2$,

$$
\mathrm{E}\{|\boldsymbol{e}_j^T [e^{i\boldsymbol{\omega}\mathbf{Y}}\mathbf{X} - \mathrm{E}(e^{i\boldsymbol{\omega}\mathbf{Y}}\mathbf{X})]|^k\} \leq \frac{k! v^2 C_1^{k-2}}{2}.
$$

Then for $\epsilon > 0$ and large enough n using the Bernstein inequality,

$$
P\left(\frac{\sum_{k=1}^n \boldsymbol{e}_j^T [e^{i\boldsymbol{\omega}\mathbf{y}_k}\mathbf{x}_k - \mathrm{E}(e^{i\boldsymbol{\omega}\mathbf{Y}}\mathbf{X})]}{n} > \epsilon \right) \leq \exp\left(-\frac{n\epsilon^2}{2(v^2 + C_1\epsilon)} \right).
$$

Take $\epsilon = C_2 \sqrt{\frac{\log p_n}{n}}$ and $C_2^2 > 12v^2$, then $C_2 \sqrt{\frac{\log p_n}{n}} \leq \frac{v^2}{2C_0}$ for large enough n, we have

$$
P\left(\frac{\sum_{k=1}^n \boldsymbol{e}_j^T [e^{i\boldsymbol{\omega}\mathbf{y}_k}\mathbf{x}_k - \mathrm{E}(e^{i\boldsymbol{\omega}\mathbf{Y}}\mathbf{X})]}{n} > C_2 \sqrt{\frac{\log p_n}{n}} \right) \leq \frac{4}{p_n^4}.
$$

As a result, with probability greater than $1 - \frac{4}{p_n^3}$, for large enough n, and by union bound,

$$
\max_{1 \leq j \leq p} \left| \frac{\sum_{k=1}^n \boldsymbol{e}_j^T e^{i\boldsymbol{\omega}\mathbf{y}_k}\mathbf{x}_k}{n} - \boldsymbol{e}_j^T \mathrm{E}(e^{i\boldsymbol{\omega}\mathbf{Y}}\mathbf{X}) \right| \leq C_2 \sqrt{\frac{\log p_n}{n}}.
$$

With the same argument, we have $P\left(|\boldsymbol{e}_j^T \bar{\mathbf{x}}| \geq C_2 \sqrt{\frac{\log p_n}{n}} \right) \leq \frac{2}{p_n^4}$, and with probability larger than $1 - \frac{2}{p_n^3}$,

$$
\max_{1 \leq j \leq p} |\boldsymbol{e}_j^T \bar{\mathbf{x}}| \leq C_2 \sqrt{\frac{\log p_n}{n}}.
$$

With probability greater than $1-\frac{C_3}{p_n^2}$ for some $C_3 zh > 0$, for large enough n, and every $1 \le j \le p$,

$$\begin{aligned} & |\tfrac{1}{n}\sum_{k=1}^n \boldsymbol{e}_j^T e^{i\boldsymbol{\omega}^T \mathbf{y}_k}\mathbf{x}_k - \tfrac{1}{n}\sum_{k=1}^n \boldsymbol{e}_j^T e^{i\boldsymbol{\omega}^T \mathbf{y}_k}\bar{\mathbf{x}} - \boldsymbol{e}_j^T \mathrm{E}(e^{i\boldsymbol{\omega}^T \mathbf{Y}}\mathbf{X})| \\ \le\ & |\tfrac{1}{n}\sum_{k=1}^n \boldsymbol{e}_j^T e^{i\boldsymbol{\omega}^T \mathbf{y}_k}\mathbf{x}_k - \boldsymbol{e}_j^T \mathrm{E}(e^{i\boldsymbol{\omega}^T \mathbf{Y}}\mathbf{X})| + |\tfrac{1}{n}\sum_{k=1}^n \boldsymbol{e}_j^T e^{i\boldsymbol{\omega}^T \mathbf{y}_k}\bar{\mathbf{x}}| \\ \le\ & 2C_2\sqrt{\tfrac{\log p_n}{n}}, \end{aligned}$$

then

$$\max_{1\le j\le p} ||\boldsymbol{e}_j^T(\Upsilon - \Sigma B_0 C_0)||_2 \le 4mC_2\sqrt{\frac{\log p_n}{n}}.$$

From the assumption $\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T \otimes \mathbf{Z}\mathbf{Z}^T) > c_0 I_{2mp}$, and $\boldsymbol{\epsilon}$ and $\mathbf{Z}$ are independent, we have there exists a constant $C_4 > 0$ such that $\widehat{\mathrm{E}^{-1}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)} \otimes I_p < C_4 I_{2mp}$, then

$$|D_1| \le C_5 m\sqrt{\frac{\log p_n}{n}}\sum_{j=1}^p ||\hat{\boldsymbol{\eta}}_j||_2, \tag{4.11}$$

where $C_5 = 4C_2C_4$ and $\hat{\boldsymbol{\eta}}_j = \hat{C}^T B_j - \hat{C}_0^T B_{0j}$.

Under the condition (C1), there exist constant $C_7, C_8 > 0$ for every $1 \le i, j \le p$,

$$P(|\hat{\sigma}_{ij} - \sigma_{ij}| > \epsilon) \le C_7 \exp(-\frac{8n\epsilon^2}{C_8}),$$

where $\hat{\sigma}_{ij}$ and σ_{ij} are $(\hat{\Sigma})_{i,j}$ and $(\Sigma)_{i,j}$, respectively. Let $\epsilon = C_6\sqrt{\frac{\log p_n}{n}}$ with $C_6 > \sqrt{C_8}$, then $P(|\hat{\sigma}_{ij}-\sigma_{ij}| > C_6\sqrt{\frac{\log p_n}{n}}) \le \frac{C_7}{p_n^8}$. Using the union bound, with probability greater than $1-\frac{C_7}{p_n^6}$,

$$\max_{1\le i,j\le p} |\hat{\sigma}_{ij} - \sigma_{ij}| < C_6\sqrt{\frac{\log p_n}{n}}.$$

Because $B_0C_0 = \Sigma^{-1}\mathrm{E}(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\mathbf{X}) = \Sigma^{-1/2}\mathrm{E}(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\mathbf{Z})$, so

$$B_0C_0C_0^TB_0^T = \Sigma^{-1/2}\mathrm{E}(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\mathbf{Z})\mathrm{E}(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\mathbf{Z}^T)\Sigma^{-1/2} = \Sigma^{-1/2}[I_d - \mathrm{Cov}(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\mathbf{Z})]\Sigma^{-1/2}.$$

$||B_0C_0||_F^2 = ||B_0||_F^2 = \mathrm{trace}\{\Sigma^{-1}[I_d - \mathrm{Cov}(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\mathbf{Z})]\} \le d/\sigma_l$. Then,

$$||B_0^T(\hat{\Sigma}-\Sigma)_{\cdot j}||_2 \le \max_{1\le i,j\le p}|\hat{\sigma}_{ij}-\sigma_{ij}|\sqrt{u}||B_0||_F \le C_6\sqrt{\frac{du\log p_n}{n\sigma_l}},$$

which implies that

$$D_2 \le C_4C_6\sqrt{\frac{du\log p_n}{n\sigma_l}}\sum_{j=1}^p ||\hat{\boldsymbol{\eta}}_j||_2. \tag{4.12}$$

Then using the bound for D_1 and D_2, (4.11) and (4.12) respectively,

$$\begin{aligned} & \tfrac{1}{2}D_3 + \lambda \textstyle\sum_{j=1}^p w_j ||\hat{\boldsymbol{\eta}}_j||_2 \\ \leq \;& \lambda \textstyle\sum_{j=1}^p w_j ||B_{0j}||_2 - \lambda \textstyle\sum_{j=1}^p w_j ||\hat{B}_j||_2 \\ & + \lambda \textstyle\sum_{j=1}^p w_j ||\hat{\boldsymbol{\eta}}_j||_2 + C_9 \sqrt{\tfrac{\log p_n}{n}} \textstyle\sum_{j=1}^p ||\hat{\boldsymbol{\eta}}_j||_2, \end{aligned}$$

where $\tilde{C} = C_9(m + \sqrt{du})$ and $C_9 = C_5 + C_4 C_6 \sigma_l^{-1/2}$. For every $j \in \mathcal{A}_0$, $|||B_{0j}||_2 - ||\hat{B}_j||_2| \leq ||\hat{\boldsymbol{\eta}}||_2$, and for every $j \in \mathcal{A}_0^c$, $||\hat{B}_j||_2 = ||\hat{\boldsymbol{\eta}}_j||_2$, then

$$\begin{aligned} & \tfrac{1}{2}D_3 + \lambda \textstyle\sum_{j=1}^p w_j ||\hat{\boldsymbol{\eta}}_j||_2 \leq 2\lambda \textstyle\sum_{j \in \mathcal{A}_0} w_j ||\hat{\boldsymbol{\eta}}_j||_2 + \tilde{C} \sqrt{\tfrac{\log p_n}{n}} \textstyle\sum_{j=1}^p ||\hat{\boldsymbol{\eta}}_j||_2 \\ \Rightarrow \;& \tfrac{1}{2}D_3 + \textstyle\sum_{j \in \mathcal{A}_0^c} \left(\lambda w_j - \tilde{C} \sqrt{\tfrac{\log p_n}{n}} \right) ||\hat{\boldsymbol{\eta}}_j||_2 \leq \textstyle\sum_{j \in \mathcal{A}_0} \left(\lambda w_j + \tilde{C} \sqrt{\tfrac{\log p_n}{n}} \right) ||\hat{\boldsymbol{\eta}}_j||_2. \end{aligned}$$

By choosing the $\tilde{\lambda} = 2C_9(m + \sqrt{du})\sqrt{\frac{\log p_n}{n}}$, $\lambda = 2^{1-\rho} C_9 C_\phi^{\rho/2} (m + \sqrt{du}) \sqrt{\frac{\log p_n}{n^{1+\rho\phi}}}$, and $2\rho(\eta - \phi/2) > 1 - 2\eta$, with conditions (C3) and (C4), we have

$$\lambda w_j \leq 2\tilde{C} \sqrt{\frac{\log p_n}{n}}, \; \forall j \in \mathcal{A}_0 \text{ and } \lambda w_j \geq 2\tilde{C} \sqrt{\frac{\log p_n}{n}}, \; \forall j \in \mathcal{A}_0^c,$$

then

$$\frac{1}{2} \sum_{j \in \mathcal{A}_0^c} \lambda w_j ||\hat{\boldsymbol{\eta}}_j||_2 \leq \frac{1}{2} D_3 + \frac{1}{2} \sum_{j \in \mathcal{A}_0^c} \lambda w_j ||\hat{\boldsymbol{\eta}}_j||_2 \leq 3 \sum_{j \in \mathcal{A}_0} \tilde{C} \sqrt{\frac{\log p_n}{n}} ||\hat{\boldsymbol{\eta}}_j||_2,$$

and

$$\sum_{j \in \mathcal{A}_0^c} ||\hat{\boldsymbol{\eta}}_j||_2 \leq 3 \sum_{j \in \mathcal{A}_0} ||\hat{\boldsymbol{\eta}}_j||_2.$$

Let $B_3 = [\mathrm{vec}(\hat{B}\hat{C}) - \mathrm{vec}(B_0 C_0)]^T [\widehat{\mathrm{E}^{-1}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)} \otimes \Sigma] [\mathrm{vec}(\hat{B}\hat{C}) - \mathrm{vec}(B_0 C_0)]$, we get

$$\begin{aligned} & |D_3 - B_3| \\ \leq \;& C_4 |\textstyle\sum_{j=1}^p \sum_{k=1}^p (\hat{\sigma}_{jk} - \sigma_{jk}) \hat{\boldsymbol{\eta}}_j^T \hat{\boldsymbol{\eta}}_k| \\ \leq \;& C_4 C_6 \sqrt{\tfrac{\log p_n}{n}} \left[|\textstyle\sum_{j \in \mathcal{A}_0} \sum_{k \in \mathcal{A}_0} \hat{\boldsymbol{\eta}}_j^T \hat{\boldsymbol{\eta}}_k| + |\sum_{j \in \mathcal{A}_0^c} \sum_{k \in \mathcal{A}_0^c} \hat{\boldsymbol{\eta}}_j^T \hat{\boldsymbol{\eta}}_k| + 2|\sum_{j \in \mathcal{A}_0} \sum_{k \in \mathcal{A}_0^c} \hat{\boldsymbol{\eta}}_j^T \hat{\boldsymbol{\eta}}_k| \right] \\ \leq \;& C_4 C_6 \sqrt{\tfrac{\log p_n}{n}} [(\textstyle\sum_{k \in \mathcal{A}_0} ||\hat{\boldsymbol{\eta}}_j||_2)^2 + (3 \textstyle\sum_{k \in \mathcal{A}_0} ||\hat{\boldsymbol{\eta}}_j||_2)^2 + 6(\textstyle\sum_{k \in \mathcal{A}_0} ||\hat{\boldsymbol{\eta}}_j||_2)^2] \\ \leq \;& 16 C_4 C_6 \sqrt{\tfrac{\log p_n}{n}} (\textstyle\sum_{k \in \mathcal{A}_0} ||\hat{\boldsymbol{\eta}}_j||_2)^2 \end{aligned}$$

Also, define $\hat{\mathcal{A}}_{11}$ be the index subset in $\mathcal{A}_0^c$ that corresponds to the u largest $||\hat{\boldsymbol{\eta}}_j||_2$'s for $j \in \mathcal{A}_0^c$. Define $\tilde{\mathcal{A}}_0 = \mathcal{A}_0 \cup \hat{\mathcal{A}}_{11}$,

$$\begin{aligned} B_3 \;\leq\; & |D_3 - B_3| + D_3 \\ \leq \;& 32 C_4 C_6 \sqrt{\tfrac{\log p_n}{n}} (\textstyle\sum_{k \in \mathcal{A}_0} ||\hat{\boldsymbol{\eta}}_j||_2)^2 + 6\tilde{C} \sqrt{\tfrac{\log p_n}{n}} \textstyle\sum_{j \in \mathcal{A}_0} ||\hat{\boldsymbol{\eta}}_j||_2 \\ \leq \;& 64 C_4 C_6 u \sqrt{\tfrac{\log p_n}{n}} (\textstyle\sum_{k \in \tilde{\mathcal{A}}_0} ||\hat{\boldsymbol{\eta}}_j||_2^2) + 6\tilde{C} (\tfrac{2u \log p_n}{n} \textstyle\sum_{j \in \tilde{\mathcal{A}}_0} ||\hat{\boldsymbol{\eta}}_j||_2^2)^{1/2}, \end{aligned}$$

which implies that

$$\begin{aligned}(\textstyle\sum_{j\in\tilde{\mathcal{A}}_0}||\hat{\boldsymbol{\eta}}_j||_2^2)^{1/2} &\leq \frac{6\tilde{C}\sqrt{\frac{2u\log p_n}{n}}}{B_3/\sum_{j\in\tilde{\mathcal{A}}_0}||\hat{\boldsymbol{\eta}}_j||_2^2-64C_4C_6u\sqrt{\frac{\log p_n}{n}}} \\ &\leq \frac{12\tilde{C}}{\sigma_*\sigma_l}\sqrt{\frac{u\log p_n}{n}}\end{aligned}$$

We can also have $\sum_{j\in\mathcal{A}_0^c\setminus\hat{\mathcal{A}}_{11}}||\hat{\boldsymbol{\eta}}_j||_2^2 \leq 9\sum_{j\in\tilde{\mathcal{A}}_0}||\hat{\boldsymbol{\eta}}_j||_2^2$ (Qian et al., 2018), combine these two results above

$$||\hat{B}\hat{C}-B_0C_0||_F \leq \frac{48\tilde{C}}{\sigma_*\sigma_l}\sqrt{\frac{u\log p_n}{n}}.$$

Hence, by Wedin's Theorem,

$$||P_{\mathcal{S}_{\hat{B}}}-P_{\mathcal{S}_{\mathbf{Y}|\mathbf{X}}}||_F = O_p((m+(du)^{1/2})\sqrt{u\log p_n/n}).$$

So far we have proved the first statement of Theorem (9), we can prove the second statement, oracle property, which is similar to the proof of Qian et al. (2018). □

Proof of Theorem 10: We let $\tau = 2C_6\sqrt{\frac{\log p_n}{n}}$, $\tilde{\lambda} = 2C_9^*(m+\sqrt{du\wedge l})\sqrt{\frac{\log p_n}{n}}$, $\lambda = 2^{1-\rho}C_9^*C_\phi^{\rho/2}(m+\sqrt{du\wedge l})\sqrt{\frac{\log p_n}{n^{1+\rho\phi}}}$, and $2\rho(\eta-\phi/2)>1-2\eta$, where C_9^* is defined later in the proof. We want to find the updated upper bound for D_2. From the C6, we denote the upper bound of $\max_{1\leq i\leq p}\sum_{j=1}^p|\sigma_{ij}|^\kappa$ to be s, then we can have $||\widetilde{\Sigma}-\Sigma||_2 \leq 13C_6s\left(\frac{\log p_n}{n}\right)^{\frac{1-\kappa}{2}}$ (Qian et al., 2018), then $||B_0^T(\widetilde{\Sigma}-\Sigma)||_2 \leq \frac{13C_6s}{\sigma_l^{1/2}}\left(\frac{\log p_n}{n}\right)^{\frac{1-\kappa}{2}}$, which implies that

$$D_2 \leq (13s+1)C_4C_6\sigma_l^{-1/2}[du\wedge l]^{1/2}\sqrt{\frac{\log p_n}{n}}\sum_{j=1}^p||\hat{\boldsymbol{\eta}}_j||_2.$$

We replace $\tilde{C}$ wich $\tilde{C}^* = C_9^*(m+\sqrt{du\wedge l})$, where $C_9^* = C_5+(13s+1)C_4C_6\sigma_l^{-1/2}$, then use the same way as the theorem 9. □

Bibliography

Antoniadis, A. (1997). Wavelets in statistics: a review. *Journal of the Italian Statistical Society*, 6(2):97.

Antoniadis, A. et al. (2007). Wavelet methods in statistics: Some recent developments and their applications. *Statistics Surveys*, 1:16–55.

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

Bentler, P. M. and Xie, J. (2000). Corrections to test statistics in principal hessian directions. *Statistics and Probability Letters*, 47(4):381–389.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Annals of Statistics*, 36(6):2577–2604.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384.

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

Chiaromonte, F., Cook, R. D., and Li, B. (2002). Sufficient dimension reduction in regressions with categorical predictors. *Annals of Statistics*, 30(2):475–497.

Chui, C. K. (2016). *An introduction to wavelets*. Elsevier.

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

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. John Wiley & Sons, Inc., New York, NY.

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

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

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

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

Cook, R. D. and Weisberg, S. (1991). Comment on "Sliced inverse regression for dimension reduction" by K.-C. Li. *Journal of the American Statistical Association*, 86(414):328–332.

Cook, R. D. and Zhang, X. (2014). Fused estimators of the central subspace in sufficient dimension reduction. *Journal of the American Statistical Association*, 109(506):815–827.

Daubechies, I. (1992). *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32(2):407–451.

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

Hsing, T. (1999). Nearest neighbor inverse regression. *Annals of Statistics*, 27(2):697–731.

Hsing, T. and Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression. *Annals of Statistics*, 20(2):1040–1061.

Li, B., Cook, R. D., and Chiaromonte, F. (2003). Dimension reduction for the conditional mean in regressions with categorical predictors. *Annals of Statistics*, 31(5):1636–1668.

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

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

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

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

Li, L. and Nachtsheim, C. J. (2006). Sparse sliced inverse regression. *Technometrics*, 48(4):503–510.

Li, R., Zhong, W., and Zhu, L.-P. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139.

Lin, Q., Zhao, Z., and Liu, J. S. (2016). Sparse sliced inverse regression via lasso.

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

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

Ni, L. and Cook, R. D. (2007). A robust inverse regression estimator. *Statistics and Probability Letters*, 77(3):343–349.

Ni, L., Cook, R. D., and Tsai, C. L. (2005). A note on shrinkage sliced inverse regression. *Biometrika*, 92(1):2425–247.

Qian, W., Ding, S., and Cook, R. D. (2018). Sparse minimum discrepancy approach to sufficient dimension reduction with simultaneous variable selection in ultrahigh dimension. *Journal of the American Statistical Association*.

Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186.

Saracco, J. (2005). Asymptotic for pooled marginal slicing estimator based on sir approach. *Journal of Multivariate Analysis*, 96(1):117–135.

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

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

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

Weng, J. and Yin, X. (2018). Fourier transform approach for inverse dimension reduction method. *Submitted to Journal of Nonparametric Statistics*.

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

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

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

Yin, X. and Hilafu, H. (2015). Sequential sufficient dimension reduction for large p, small n problems. *Journal of the Royal Statistical Society: Series B*, 77:879–892.

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

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

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67.

Zhu, L.-P., Yu, Z., and Zhu, L.-X. (2010a). A sparse eigen-decomposition estimation in semiparametric regression. *Computational Statistics & Data Analysis*, 54(4):976–986.

Zhu, L.-P. and Zhu, L.-X. (2009). Dimension reduction for conditional variance in regressions. *Statistica Sinica*, 19(2):869–883.

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

Zhu, L.-P., Zhu, L.-X., and Wen, S.-Q. (2010c). On dimension reduction in regressions with multivariate responses. *Statistica Sinaca*, 20(1):1291–1307.

Zhu, L.-X. and Ng, K. W. (1995). Asymptotic of sliced inverse regression. *Statistica Sinica*, 5(2):727–736.

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

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

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B*, 67(2):301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286.

Vita

Education

- M.S. in Statistics, University of Kentucky.
- M.S. in Mathematics, Fudan University.
- B.S. in Mathematics, Sun Yat-Sen University.

Awards

1. R.L. Anderson Teaching Excellence Award, Department of Statistics, University of Kentucky. 2017-2018
2. David Allen Fellow for Statistical Excellence, Department of Statistics, University of Kentucky. 2017-2018

Publications

1. Weng, J. and Yin, X. (2018), "Fourier Transform Approach for Inverse Dimension Reduction Method", Journal of Nonparametric Statistics. 30(4):1049-1071, 2018.
2. Weng, J. and Young, D. (2017), "Some Dimension Reduction Strategies for the Analysis of Survey Data", Journal of Big Data, 43(4), 2017.