The Pennsylvania State University
The Graduate School

DIMENSION REDUCTION AND SUFFICIENT GRAPHICAL MODELS

A Dissertation in
Statistics
by
Kyongwon Kim

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2020

The dissertation of Kyongwon Kim was reviewed and approved by the following:

Bing Li
Verne M. Willaman Professor of Statistics
Dissertation Advisor, Chair of Committee

David Hunter
Professor of Statistics

Bharath Sriperumbudur
Associate Professor of Statistics

Rongling Wu
Distinguished Professor of Public Health Sciences

Ephraim Hanks
Associate Professor of Statistics
Graduate Program Chair

Abstract

The methods I develop in my thesis are based on linear or nonlinear sufficient dimension reduction. The basic principle of linear sufficient dimension reduction is to extract a small number of linear combinations of predictor variables, which can represent original predictor variables without loss of information on the conditional distribution of response variable given predictor variables. Nonlinear sufficient dimension reduction is a more generalized version of linear sufficient dimension reduction to the nonlinear context.

I am focusing on applying sufficient dimension reduction methods into two areas, regression modeling and graphical models. The first project is about statistical inference in regression context after sufficient dimension reduction. Second, I apply nonlinear sufficient dimension reduction method to the well known statistical graphical models in machine learning. These projects have consistency in a context that discovering areas that sufficient dimension reduction can be applied and establishing statistical theory behind their applications.

My first project is about post sufficient dimension reduction statistical inference. The methodologies of sufficient dimension reduction have undergone extensive developments in the past three decades. However, there has been a lack of systematic and rigorous development of post dimension reduction inference, which has seriously hindered its applications. The current common practice is to treat the estimated sufficient predictors as the true predictors and use them as the starting point of the downstream statistical inference. However, this naive inference approach would grossly overestimate the confidence level of an interval, or the power of a test, leading to the distorted results. In this project, we develop a general and comprehensive framework of post dimension reduction inference, which can accommodate any dimension reduction method and model building method, as long as their corresponding influence functions are available. Within this general framework, we derive the influence functions and present the explicit post reduc-

tion formulas for the combinations of numerous dimension reduction and model building methods. We then develop post reduction inference methods for both confidence interval and hypothesis testing. We investigate the finite-sample performance of our procedures by simulations and a real data analysis.

My second project is about applying nonlinear dimension reduction technique to graphical models. We introduce the Sufficient Graphical Model by applying the recently developed nonlinear sufficient dimension reduction techniques to the evaluation of conditional independence. Graphical model is nonparametric in nature, as it does not make distributional assumptions such as the Gaussian or copula Gaussian assumptions. However, unlike fully nonparametric graphical model, which relies on the high-dimensional kernel to characterize a conditional independence, our graphical model is based on a conditional independence given a set of sufficient predictors with a substantially reduced dimension. In this way, we avoid the curse of dimensionality that comes with a high-dimensional kernel. We develop the population-level properties, convergence rate, and consistency of our estimate. By simulation comparisons and an analysis of the DREAM 4 Challenge data set, we demonstrate that our method outperforms the existing methods when the Gaussian or copula Gaussian assumptions are violated, and its performance remains excellent in the high-dimensional setting.

Table of Contents

List of Figures viii

List of Tables x

Acknowledgments xi

Chapter 1
Introduction 1

Chapter 2
Literature Review 4
2.1 Sufficient Dimension Reduction 4
2.1.1 Basic Formulation 4
2.1.2 Linear SDR methods 6
2.1.3 Nonlinear generalized SDR method 11
2.2 Graphical Models 14

Chapter 3
On Post Dimension Reduction Statistical Inference 17
3.1 Introduction 17
3.2 General framework post reduction inference 19
3.2.1 Reduction, estimation and composite functionals 19
3.2.2 Influence function and asymptotic distribution 22
3.2.3 Asymptotic comparison of naive and objective inference . . . 25
3.2.4 Identifiability of reduction parameter 26
3.3 Influence functions for estimation functionals 27
3.3.1 Differentiable estimating equations 27
3.3.2 Non-differentiable estimating equations 29

3.3.3 Generalized method of moments . . . 33
3.4 Influence functions for reduction functionals . . . 36
3.5 Post dimension reduction inference . . . 40
3.6 Conclusions . . . 43

Chapter 4
Sufficient Graphical Models **44**
4.1 Introduction . . . 44
4.2 Sufficient graphical model . . . 44
4.3 Estimation: population-level development . . . 47
4.3.1 Preliminaries . . . 47
4.3.2 Step 1: nonlinear dimension reduction . . . 48
4.3.3 Step 2: estimation of Sufficient Graphical Model . . . 51
4.4 Estimation: sample-level implementation . . . 53
4.4.1 Coordinating mapping . . . 53
4.4.2 Implementation of step 1 . . . 55
4.4.3 Implementation of step 2 . . . 57
4.5 Tuning . . . 59
4.6 Asymptotic theory . . . 60
4.6.1 Overview . . . 60
4.6.2 Transparent kernel . . . 62
4.6.3 Convergence rates of (i) and (iii) in (4.6.1) . . . 64
4.6.4 Optimal rates of tuning parameters . . . 67
4.7 Discussion . . . 68

Chapter 5
Proofs of Asymptotic Results for Sufficient Graphical Models **71**
5.1 Preliminaries . . . 71
5.1.1 Hilbert-Schmidt norm and operator norm . . . 71
5.1.2 Sample mean of operators . . . 72
5.1.3 Tychonoff regularized inverse . . . 73
5.1.4 Negative square root . . . 75
5.1.5 Notations for order of magnitude . . . 75
5.2 Proof of Theorem 4 . . . 76
5.3 Proof of Theorem 5 . . . 77
5.4 Proof of Theorem 6 . . . 78
5.5 Proof of Theorem 7 . . . 81
5.6 Proof of Theorem 8 . . . 83
5.7 Proof of Theorem 9 . . . 92
5.8 Proof of Theorem 11 . . . 92

Chapter 6
Simulation Studies and Real Data Analysis **94**
6.1 Simulation Studies for Post Dimension Reduction Inference 94
6.1.1 Comparison of confidence interval 95
6.1.2 Comparison of local power 98
6.2 Application for Post Dimension Reduction Inference 100
6.3 Simulation Studies for Sufficient Graphical Models 108
6.4 Application for Sufficient Graphical Models 114

Chapter 7
Future Research **117**
7.1 Functional sufficient graphical models with application to f-MRI and EEG dataset 117
7.2 Extension of Post Dimension Reduction Inference 121
7.3 Learning Causal Networks via Sufficient Faithfulness 121
7.4 Functional Component Selection 122

Appendix A
Computer Codes for Sufficient Dimension Reduction **123**

Appendix B
Computer Codes for Sufficient Graphical Models **135**

Bibliography **139**

List of Figures

6.1 Local power of hypothesis testing in model III with sample size $n = 500$. The five panels, left to right, top to bottom, correspond to the results from five SDR methods, SIR, SAVE, DR, y-PHD, and r-PHD. The red curve denotes the naive inference method, and the blue curve denotes our proposed inference method. 99

6.2 Local power of hypothesis testing in model IV with sample size $n = 500$. The five panels, left to right, top to bottom, correspond to the results from five SDR methods, SIR, SAVE, DR, y-PHD, and r-PHD. The red curve denotes the naive inference method, and the blue curve denotes our proposed inference method. 100

6.3 Response versus the first SIR predictor in BigMac data. 101

6.4 Confidence intervals for θ_2 (upper panel) and θ_4 (lower panel) in the BigMac data analysis. 102

6.5 Local power for θ_2 in the BigMac data analysis. The five panels, left to right, top to bottom, correspond to the results from five SDR methods, SIR, SAVE, DR, y-PHD, and r-PHD. The red curve denotes the naive inference method, and the blue curve denotes our proposed inference method. 103

6.6 Confidence interval for θ_1 in the BigMac data analysis. 104

6.7 Local power of hypothesis testing in model III with sample size $n = 300$. The five panels, left to right, top to bottom, correspond to the results from five SDR methods, SIR, SAVE, DR, y-PHD, and r-PHD. The red curve denotes the naive inference method, and the blue curve denotes our proposed inference method. 104

6.8 Local power of hypothesis testing in model III with sample size $n = 800$. The rest of setup is the same as in Figure 6.7. 105

6.9 Local power of hypothesis testing in model III with sample size $n = 1200$. The rest of setup is the same as in Figure 6.7. 105

6.10 Local power of hypothesis testing in model IV with sample size $n = 300$. The five panels, left to right, top to bottom, correspond to the results from five SDR methods, SIR, SAVE, DR, y-PHD, and r-PHD. The red curve denotes the naive inference method, and the blue curve denotes our proposed inference method. 106
6.11 Local power of hypothesis testing in model IV with sample size $n = 800$. The rest of setup is the same as in Figure 6.10. 106
6.12 Local power of hypothesis testing in model IV with sample size $n = 1200$. The rest of setup is the same as in Figure 6.10. 107
6.13 Averaged ROC curves of Model I. Left panel: $n = 100$; right panel: $n = 1000$ 110
6.14 Averaged ROC curves of Model II. Left panel: $n = 100$; right panel: $n = 1000$ 110
6.15 Averaged ROC curves for Model III. Left panel: $n = 50$; right panel: $n = 100$. 112
6.16 Averaged ROC curves for Model IV. Left panel: $n = 50$; right panel: $n = 100$. 112
6.17 Averaged ROC curves for Model I. Left panel: $n = 100$; right panel: $n = 1000$ when we choose one dimension. 113
6.18 Averaged ROC curves for Model II. Left panel: $n = 100$; right panel: $n = 1000$ when we choose one dimension. 113
6.19 Averaged ROC curves for Model III. Left panel: $n = 50$; right panel: $n = 100$ when we choose one dimension. 114
6.20 Averaged ROC curves for Model IV. Left panel: $n = 50$; right panel: $n = 100$ when we choose one dimension. 114

7.1 Example of human brain network (left) and f-MRI functional data for control group (Upper right) and ADHD group (Lower right). . . 119

List of Tables

6.1 Coverage probability of confidence interval for θ_1 and θ_2 in model I. 96
6.2 Coverage probability of confidence interval for θ_1 in model II. 97
6.3 Standard errors for θ_1 and θ_2 in model I and comparison to the oracle method. 97
6.4 Comparison of AUC for SGM, APCO, and the champion method for the five networks in the DREAM 4 Challenge dataset. 115

Acknowledgments

First of all, I would like to express my sincere appreciation to Professor Bing Li for all his encouragement, guidance, and the time he spent on me in my academic research. It is hard for me to overstate my gratitude toward him. Without his support, I do not think it would be possible to do the research and complete my dissertation. He gave me meaningful advice and inspiration with extreme patience to help me complete my Ph.D. degree. I am very proud of being one of his students.

I would like to thank my committee members, Dr. Hunter, Dr. Sriperumbudur, and Dr. Wu for their suggestions on my research and encouragement for my career. I would like to give my special thanks to Professor Buchanan for her advice on my statistics teaching methods.

I would like to thank all my friends at Penn State. They always support me and motivate me to work hard. I also want to give special thanks to the Penn State Department of Statistics for their support. It is fortunate to be part of the members of this friendly and wonderful department.

Last but not least, I wish to acknowledge the support and great love of family members - my father, Dong Wook Kim, my mother, Keum Joong Choi, my son, Daniel Woojin Kim, and my wife, Ha Eun Cho. They make this possible by giving me an opportunity to pursue a doctoral degree. I truly want to say I love you all.

Chapter 1

Introduction

Sufficient dimension reduction (SDR) embodies a family of methods that, in a regression setup, seek a reduction of dimensionality without loss of regression information. It has proven to be a powerful tool to extract useful information from high dimensional data and has found wide applications in high dimensional data analysis and regression graphics (Cook, 1998b; Li, 2018b,d).

Since the pioneering work of sliced inverse regression (Li, 1991a), the research in SDR has been flourishing, and numerous SDR methods have been proposed, including sliced average variance estimation (Cook and Weisberg, 1991), principal hessian directions (Li, 1992), minimum average variance estimation (Xia et al., 2002), and directional regression (Li and Wang, 2007), among many others. There have also been developments of SDR based variable selection and screening (Bondell and Li, 2009; Zhu et al., 2011), semiparametric SDR (Ma and Zhu, 2012, 2013), and nonlinear SDR (Li et al., 2011a; Li and Song, 2017a). For a comprehensive review, see Li (2018b).

Sufficient dimension reduction has proven to be a powerful tool to extract core information hidden in high-dimensional data, for the purpose of classifying, clustering, and predicting one or several response variables. Despite the extensive development of sufficient dimension reduction over the past three decades, there has been a lack of a rigorous and systematic method for conducting statistical inference after dimension reduction. Currently, the common practice is to feed the sufficient predictors into subsequent modeling as if they were the true predictors, but this leads to overly optimistic confidence intervals and p-values. Thus, in

some sense, sufficient dimension reduction is not complete as a statistical method without an adequate procedure for post dimension reduction inference. The goal of my first project is to fill this gap.

The central issue for post dimension reduction inference is to take into account both the statistical error produced model estimation and statistical error the underlying dimension reduction method. In other words, we need to track how the error caused by dimension reduction propagates into the estimation process. The basic idea of the first project is as follows. First, we develop the general asymptotic distribution for post dimension reduction estimation for a given set of influence functions of a SDR and a model estimation method. Then, we derive the influence functions for both estimating equations and the Generalized method of moments (GMM) case. After that, we derive the influence function for the four SDR methods. Therefore, we provide 5×3 combinations SDR and estimation methods already cover a wide variety of applications. More importantly, they serve as an illustration of how to derive the influence functions for a given SDR method and estimation methods, and how to plug them into our post dimension reduction inference framework to find the desired asymptotic distribution. Based on the asymptotic form under the general case and specific case, we make an asymptotic normal distribution for the estimator of parameter. And we prove asymptotic variance by utilizing estimated central space is bigger than that of made by assuming we know true effective dimension reduction space. By using that distribution, we set up statistical inference like a confidence interval, general, and local alternative hypothesis test for both estimating equation and GMM case.

Throughout my study, I have tried to apply sufficient dimension reduction methods. One of the areas that I am focusing on is graphical models in machine learning. Statistical graphical models are one of the most momentous areas in current statistical research due to the demands from many contemporary applications involving the estimation of networks. A popular graphical model is the Gaussian graphical model (Yuan and Lin, 2007), which has a uniquely simple interaction structure that reduces the model estimation problem into a sparse estimation of the precision matrix. However, the Gaussian assumption can be violated by some commonly used interaction structures, and this can hinder the performance of the Gaussian graphical model. This motivates our development of estimation and in-

ference methods for non-Gaussian graphical models by using nonparametric kernel mapping.

We propose combining principal ideas and techniques from the most recent development in nonlinear sufficient dimension reduction into the graphical models to reduce the dimension of the mapping kernels. By using nonlinear SDR, we are able to not only avoid "curse of dimensionality"(Bellman, 1961) but also make algorithms computationally economic when handling high-dimension networks. Furthermore, my model also has a strength in relaxing Gaussian and Copula Gaussian assumptions. Additionally, because we use reproducing kernel Hilbert space (RKHS), the asymptotic structures of the kernel estimates of linear operators are mathematically tractable, and some fundamental tools have been developed in the recent literature. This gives us a real expectation of developing a reasonably complete asymptotic theory for statistical inference, consistency, and convergence rate for the non-gaussian graphical models.

Chapter 2

Literature Review

2.1 Sufficient Dimension Reduction

2.1.1 Basic Formulation

With the considerable improvement of computational power and storage capacity of the computer, a modern dataset has extremely high-dimension throughout many scientific areas such as Genomics, Sociology, Machine Learning, and many others. To utilize useful information from such high-dimensional dataset, many methods for finding dimension reduction space have been developed during the past decades. Therefore, the methodologies of sufficient dimension reduction have undergone extensive developments in the past three decades.

For a response variable Y and the p-dimensional predictor vector X, Sufficient dimension reduction (SDR) is a method to extract a small number of linear combinations of predictor variables, which can represent original predictor variables without loss of information on the conditional distribution of Y given X. Mathematically, we can write

$$Y \perp\!\!\!\perp X | \eta^{\mathsf{T}} X \tag{2.1.1}$$

where $\perp\!\!\!\perp$ indicates independence and η is $p \times q$ matrix with $q \leq p$. This implies that Y depends on X only through $\eta^{\mathsf{T}} X$. It is straightforward to see that η always exists, as it can trivially take the form of the identity matrix.

Relationship (2.1.1) can also be written as follows. Let $f(\cdot)$ be the distribution

of $Y|X$. Then the basic philosophy of SDR can also be finding a function $\eta^\intercal X$ such that

$$f(Y|X) = f(Y|\eta^\intercal X). \tag{2.1.2}$$

This means we can replace predictor variable X to $\eta^\intercal X$ without loss of information. However, it is not unique, as one can rotate or amend η so that the relationship (2.1.1) still holds. That is, for any invertible matrix $A \in \mathbb{R}^{q\times q}$, $A^\intercal\eta^\intercal X$ also satisfies the conditional independence (2.1.1), $Y \perp\!\!\!\perp X|A^\intercal\eta^\intercal X$. Furthermore, any space which includes column space of η also satisfies the above sufficient condition. Thus, SDR turns to the subspace spanned by the columns of η. It is called a dimension reduction subspace, and under very minor conditions (Yin et al., 2008), the intersection of all such subspaces is itself a dimension reduction subspace. Such an intersection, by definition, is a unique and parsimonious population parameter that captures full regression information of Y given X. It is called the central subspace, is denoted as $\mathcal{S}_{Y|X}$, and is the main object of interest in the SDR inquiry. Because its definition central subspace is the smallest SDR subspaces and according to Yin et al. (2008) existence can be guaranteed when X is in convex support.

Oftentimes, instead of a whole information of $Y|X$, we are interested in partial information such as the mean $E[Y|X], \mathrm{var}(Y|X)$, and $\mathrm{median}(Y|X)$. If we are interested in the relationship between Y and $E[Y|X]$, investigating $Y|X$ to withdraw information of $E[Y|X]$ contains inefficient task because the distribution of $E[Y|X]$ is always a subset of $Y|X$. As SDR subspaces imply, we are finding η such that $\eta^\intercal X$ explains all the relationship between Y and X, mean dimension reduction subspace uses a similar concept. That is to find η satisfies

$$E[Y|X] = E[Y|\eta^\intercal X].$$

Therefore Cook and Li (2002) define central mean subspace as a subspace spanned by the columns of $\eta \in \mathbb{R}^{p\times d}$ such that

$$Y \perp\!\!\!\perp E[Y|X]|\eta^\intercal X.$$

Similar to SDR subspaces, mean dimension reduction subspaces has an identifiabil-

ity issue. Thus Cook and Li (2002) present intersection of all the mean subspaces as a central mean subspace and it can be expressed as $\mathcal{S}_{E[Y|X]}$. According to Cook and Li (2002) and Zhu and Zhu (2009), an existence of $\mathcal{S}_{E[Y|X]}$ can also be guaranteed by the same condition with $\mathcal{S}_{Y|X}$. Finally note that

$$\mathcal{S}_{E[Y|X]} \subseteq \mathcal{S}_{Y|X}.$$

Before move on to discuss the preliminary conditions for the SDR, we want to introduce an invariant property of SDR which is useful when we implement SDR methods. Let $U = W^\top X + a$ where $W \in \mathbb{R}^{p\times p}$ nonsingular matrix, and $a \in \mathbb{R}^p$. Then $\mathcal{S}_{Y|U} = W^{-1}\mathcal{S}_{Y|X}$. This means we are able to use standardization when we use SDR in practice, for example $Z = \Sigma^{-\frac{1}{2}}(X - E[X])$ where Σ is a covariance matrix of X. Therefore, from now on, without loss of generality, we assume X is standardized, that is $E[X] = 0$ and $\mathrm{var}(X) = I_p$.

Here we discuss critical conditions for inverse regression methods in SDR which will be introduced later.

Assumption 1. *(Linearity condition) Let $\eta \in \mathbb{R}^{p\times d}$ be a matrix whose column space is $\mathcal{S}_{Y|X}$. We assume $E[X|\eta^\top X]$ is a linear in $\eta^\top X$.*

This condition is required for most of the first-order methods of SDR and is known to be satisfied when X is an elliptically distributed. In addition to the linearity condition, some SDR methods, especially second-order methods which will be discussed later, need constant variance assumption.

Assumption 2. *(Constant Variance condition) Let $\eta \in \mathbb{R}^{p\times d}$ be the same as Assumption 1.* $\mathrm{var}(X|\eta^\top X)$ *is a nonrandom matrix.*

According to Cook and Li (2002), constant variance condition hold when X is normally distributed. Besides, it is known that the constant variance assumption approximately holds when X is elliptically distributed.

2.1.2 Linear SDR methods

We present a brief introduction of linear SDR methods based on inverse regression such as Sliced Inverse Regression, Sliced Average Variance Estimation, Contour

Regression, and Directional Regression. There are several advantages of inverse regression methods.

First, they do not need smoothing regardless of the relationship between X and Y and the dimension of X. Second, most of these methods can be calculated as a generalized eigen-decomposition problem, which is easy to implement. Third, the methods do not require any assumption on the link function of regression models.

There are two categories of inverse conditional moment methods. First order methods which are based on $E[X|Y]$ include Sliced Inverse Regression (Li, 1991a), Ordinary Least Squares(Li and Duan, 1989), Parametric Inverse Regression (Bura and Cook, 2001), and Kernel Inverse Regression (Zhu and Fang, 1996). Second order methods are based on $E[XX^\intercal]$ and it contains Sliced Average Variance Estimation (Cook and Weisberg, 1991), Contour Regression (Li et al., 2005), Directional Regression (Li and Wang, 2007), and Principal Hessian Direction (Li, 1992).

Inverse moment methods require assumptions on the predictors. First-order methods are hold based on the Assumption 1. Second-order methods require Assumption 2 in addition to linearity assumption.

Ordinary Least Squares

Based on the single index model, when $d = 1$, Duan and Li (1991) presents Ordinary Least Squares (OLS) method can be a consistant estimator for η under the Assumption 1. We can also express this as

$$E[XY] \in \mathcal{S}_{E[Y|X]}$$

Where $E[XY]$ is the OLS estimator. This can be shown as follows. Define $\text{span}(\eta) = \mathcal{S}_{E[Y|X]}$ and $P_\eta = \eta(\eta^\intercal\eta)^{-1}\eta^\intercal$ is a projection matrix to $\text{span}(\eta)$. Because $Y \perp\!\!\!\perp E[Y|X]|\eta^\intercal X$ and the linearity assumption

$$E[XY] = E[XE[Y|X]] = E[XE[Y|\eta^\intercal X]] = P_\eta E[XY].$$

The advantage of the OLS method is that the OLS estimator is $\sqrt{n}$ consistent estimator to $\mathcal{S}_{E[Y|X]}$. However, because OLS is based on the single-index model, it

is not able to estimate $d > 1$ in the central mean subspace.

Sliced Inverse Regression

Sliced Inverse Regression (SIR)(Li, 1991a) generalized the idea called slicing regression (Duan and Li, 1991) under the Assumption 1. The theoretical foundation of SIR is that conditional distribution Y given X can be obtained by an inverse regression and can be described as follows. Because

$$E[X|Y] = E[E[X|Y, \eta^\intercal X]|Y] = E[E[X|\eta^\intercal X]|Y] = E[P_\eta X|Y] = P_\eta E[X|Y].$$

Therefore, $E[X|Y] \in \mathcal{S}_{Y|X}$ and it is the fundamental idea of SIR. We can easily derive

$$\text{span}(E[X|Y]) = \text{span}(\text{var}(E[X|Y])) \subseteq \mathcal{S}_{Y|X}.$$

Therefore $\text{var}(E[X|Y])$ is a subspace of central subspace. SIR is now one of the most common SDR methods because it is easy to implement, and it has unbiasedness and consistency. Furthermore, by using a sample inverse moment, SIR always has a clear estimate. However, it also has several limitations. First of all, SIR does not work well enough if Y and X have a symmetric structure, and it cannot recover the entire central subspace. Second, when the response has a binary structure, SIR also does not work well because it captures at most one direction of the central subspace.

Sliced Average Variance Estimation

Sliced Average Variance Estimation (Cook and Weisberg, 1991) can overcome the disadvantage of SIR when Y and X have a symmetric relationship as it means the within slice variance instead of averaging the within each slice. Therefore, SAVE can achieve exhaustiveness when estimate $\mathcal{S}_{Y|X}$. SAVE is a second-order method as it is based on conditional variance, $\text{var}(X|Y)$. We can express

$$\begin{aligned}\text{var}(X|Y) =& E[\text{var}(X|Y, \eta^\intercal X)|Y)] + \text{var}(E[X|Y, \eta^\intercal X]|Y) \\ =& E[\text{var}(X|\eta^\intercal X)|Y] + \text{var}(E[X|\eta^\intercal X]|Y).\end{aligned}$$

The second equality is because of $Y \perp\!\!\!\perp X|\eta^{\intercal}X$. Let $Q_{\eta} = I_p - P_{\eta}$. Then Under Assumptions 1 and 2

$$\text{var}(X|Y) = P_{\eta}\text{var}(X|Y)P_{\eta} + Q_{\eta}.$$

Therefore, $I_p - \text{var}(X|Y) = P_{\eta}(I_p - \text{var}(X|Y))P_{\eta}$. Hence, we can use the column space of the matrix $I_p - \text{var}(X|Y)$ to estimate $\mathcal{S}_{Y|X}$. In practice, span($E[I_p - \text{var}(X|Y)]^2$) can be used to estimate $\mathcal{S}_{Y|X}$.

SAVE has an advantage compared to SIR when Y is a categorical variable. If the response variable has a s distinct values, SIR can only estimate at most $q = s - 1$, but SAVE can recover larger subspace. As we can see from the point that SAVE can be applied to the situation when X and Y have a symmetric dependence and because SAVE is exhaustive, SAVE covers a larger proportion than SIR, that is

$$\mathcal{S}_{\text{SIR}} \subseteq \mathcal{S}_{\text{SAVE}} = \mathcal{S}_{Y|X}.$$

Similar to SIR, SAVE is also a model-free method. However, SAVE also has some drawbacks. When the sample size is finite, SAVE does not work well when Y and X has a monotone relationship. Furthermore, SAVE lost its efficiency in terms of computation because it includes a conditional covariance. Finally, SAVE requires constant variance assumption in addition to the linearity condition.

Contour Regression

Slicing methods such as SIR and SAVE only consider the interaction of X within each slice of Y. Therefore, it can easily miss the inter-slice information. This is one of the reasons why SIR is not exhaustive, and SAVE does not work well in practice. Thus, Contour Regression (Li et al., 2005) is introduced by estimating $\mathcal{S}_{Y|X}$ based on the empirical directions, which can be defined as a direction of $X - \tilde{X}$ where $\tilde{X}$ is the set of independent copy of X. Contour Regression (CR) basically considers conditional moment based on these empirical directions, and

this conditional moment spans the orthogonal complement of the central subspace. We can define contour directions as the empirical direction with response variable changes within a very small boundary. Then CR can be seen as regressing X within each contour of Y. There are two versions of CR, which are simple contour regression (SCR) and general contour regression (GCR). The following procedures are about SCR. Let $\tilde{X}$ and $\tilde{Y}$ are an independent copy of X and Y, respectively. Then objectie matrix for SCR is as follows.

$$E[(X - \tilde{X})(X - \tilde{X})^{\intercal} || Y - \tilde{Y}| \leq c].$$

Here $|Y - \tilde{Y}| \leq c$ determines the volume of contour, and we can use an orthogonal complement of the above objective matrix to estimate $\mathcal{S}_{Y|X}$. Based on Li et al. (2005), CR has exhaustiveness under the Assumption 1 in addition to some mild assumptions. This is because CR can consider both within a slice and inter-slice information. Because CR becomes insensitive when X and Y have a symmetric relationship, Li et al. (2005) also introduced GCR, which has a more complicated version of contour directions and GCR can also enjoy exhaustiveness with milder conditions. CR has a computational drawback because it involves calculating empirical directions whose cost is $O(n^2)$ whereas SIR and SAVE enjoy $O(n)$.

Directional Regression

To overcome the higher computational cost of CR, Li and Wang (2007) introduced Directional Regression (DR), which utilizes empirical directions efficiently than CR by regressing the empirical directions, $X - \tilde{X}$ directly on the directions on the paired responses $(Y, \tilde{Y})$. In short, it projects the variations of empirical directions onto the space of response Y. According to Li and Wang (2007), if we consider

$$E[(X - \tilde{X})(X - \tilde{X})^{\intercal} | Y, \tilde{Y}]$$

then it is true that

$$\text{span}(2I_p - E[(X - \tilde{X})(X - \tilde{X})^{\intercal} | Y, \tilde{Y}]) \subseteq \mathcal{S}_{Y|X}.$$

Because it can be reexpressed as

$$\begin{aligned} 2I_p - E[(X-\tilde{X})(X-\tilde{X})^\intercal|Y,\tilde{Y}] =& E[X|Y]E[\tilde{X}^\intercal|\tilde{Y}] - E[X|Y]E[X|Y] \\ &+ E[\tilde{X}|\tilde{Y}]E[X^\intercal|Y] - E[\tilde{X}|\tilde{Y}]E[\tilde{X}^\intercal|\tilde{Y}] \\ &+ I_p - \mathrm{var}(X|Y) + I_p - \mathrm{var}(\tilde{X}|\tilde{Y}). \end{aligned}$$

Thus, we can see that DR can be expressed as a combination of SIR and SAVE. Besides subspaces, which are estimated from DR. includes a span of SIR and SAVE as their subspaces. DR enjoys several advantages. First of all, under the Assumption 1 and other mild assumptions, DR is an exhaustive method. Furthermore, it enjoys computational simplicity, $O(n)$, same order as SIR and SAVE.

2.1.3 Nonlinear generalized SDR method

Li, Artemiou, and Li (2011a) and Lee, Li, and Chiaromonte (2013) generalizes this framework to nonlinear sufficient dimension reduction (nonlinear SDR), where the goal is to recover the smallest sub σ-field $\mathcal{G}$ on the support of X such that $Y \perp\!\!\!\perp X|\mathcal{G}$. The collection of square-integrable functions measurable with respect to $\mathcal{G}$ is called central class, written as $\mathfrak{S}_{Y|X}$. They pointed out that several existing nonlinear dimension reduction methods developed in the statistics and machine learning communities are actually estimators of $\mathfrak{S}_{Y|X}$, and introduced several new estimators for this class. Nonlinear SDR greatly expands the scope, flexibility, and effectiveness of SDR.

Generalized Sliced Inverse Regression (GSIR)

Let $(\Omega_X, \mathcal{F}_X, P)$ be a probability space, Ω_X a subset of $\mathbb{R}^p$, and $X : \Omega \to \Omega_X$ a random vector, where Ω is a sample space. Also, let P_X be the distribution of X, Ω_X be the support of X. We can define a response variable Y in a similar context. Let $L_2(P)$ denote the class of all functions

$$\{(f : \Omega_X \to \mathbb{R}) : \int f dP = 0, \int f^2 dP < \infty\}.$$

Here, instead of L_2 structure, we use Reproducing Kernel Hilbert Space (RKHS) feature to define our new operator and derive related methods. This is because

various asymptotic tools for linear operators have been developed in the RKHS setting. Also, the asymptotic structures of the kernel estimates of linear operators are mathematically tractable, and some rudimentary tools have been developed in the recent literature. For example Fukumizu, Bach, and Gretton (2007) and Bach (2008) give us a real prospect of developing a reasonably complete asymptotic theory for consistency rate and statistical inference for non-gaussian graphical models. Define $k_X : \Omega_X \times \Omega_X \to \mathbb{R}$ is a positive definite kernel. Let $\mathscr{H}_X$ be the RKHS of functions of X based on kernel k; that is, $\mathscr{H}_X$ is the space spanned by $\{k(\cdot, X) : X \in \Omega_X\}$ with its inner product given by $\langle k(\cdot, X), k(\cdot, X)\rangle_{\mathscr{H}_X} = k(X, X)$. Here, we need to assume that all the functions in $\mathscr{H}_X$ are square-integrable, which can be written as following assumption.

Assumption 3. $E[k(X, X)] < \infty$.

This condition can be easily satisfied such as Gaussian radial basis kernel. Furthermore, let $\mathscr{H}_X$ and $\mathscr{H}_Y$ be RKHS of functions of X and Y, respectively.

Assumption 4. *$\mathscr{H}_Y$ and $\mathscr{H}_X$ are dense in $L_2(P_Y)$ and $L_2(P_X)$ modulo constant.*

Here we define basic notations. For bounded operator T, $\mathrm{ran}(T)$ represents the range of T and $\mathrm{null}(T)$ the null space of T. It can be written as

$$\mathrm{ran}(T) = \{Th : h \in \mathscr{H}\}, \quad \mathrm{null}(T) = \{h \in \mathscr{H} : T(h) = 0\}$$

where $\mathscr{H}$ is Hilbert space. Furthermore, $\Sigma_Y^{\dagger}$ is a Moore - Penrose inverse of Σ_Y.

Following assumption is required to guarantee boundedness of covariance operator.

Assumption 5. *There are constants $C_1 > 0$ and $C_2 > 0$ such that* $\mathrm{var}[g(X)] \leq C_1\|g\|_{\mathscr{H}_X}$ *and* $\mathrm{var}[f(Y)] \leq C_2\|f\|_{\mathscr{H}_Y}$

Let $\Sigma_{XY} : \mathscr{H}_X \to \mathscr{H}_Y$ and $\Sigma_{YX} : \mathscr{H}_Y \to \mathscr{H}_X$ be bounded linear operators defined by the relation

$$\left\langle f, \Sigma_{XY} g\right\rangle_{\mathscr{H}_X} = \left\langle \Sigma_{YX} f, g\right\rangle_{\mathscr{H}_X} = \mathrm{cov}[g(X), f(Y)],$$

where $g \in \mathscr{H}_X$, $f \in \mathscr{H}_Y$.

Suppose that there is a compact operator $E_{XY} : \mathscr{H}_Y \to \mathscr{H}_X$ such that

$$\Sigma_{XY} = \Sigma_{XX} E_{XY}.$$

We denote the operator $E_{XY} : \mathscr{H}_Y \to \mathscr{H}_X$ is conditional mean operator. Intuitively, the operator E_{XY} can be understood in the sense that it sends any function $f \in \mathscr{H}_Y$ to its conditional expectation given X. Then by using similar logic to Lee, Li, and Chiaromonte (2013) we can develop relationship between conditional mean operator and conditional mean. However, the following assumption is required.

Suppose assumption 5 is hold. If $\mathscr{H}_Y, \mathscr{H}_X$ are dense in $L_2(Y)$, $L_2(X)$, respectively. Then we have:

$$\begin{aligned} \forall f \in \mathscr{H}_Y, \quad then \quad E_{XY} f = E[f(Y)|X] \\ \forall g \in \mathscr{H}_X, \quad then \quad E^*_{XY} g = E[g(X)|Y] \end{aligned}$$

Furthermore, under some assumptions, for any f,g $\in \mathscr{H}_Y$,

$$\langle g, E_{XY} E^*_{XY} f \rangle_{\mathscr{H}_X} = \mathrm{cov}[E(g(X|Y)), E(f(X|Y))]$$

Based on the above relationship, we can induce quadratic form

$$f \mapsto \langle f, E_{XY} E^*_{XY} \rangle_{\mathscr{H}_X}, \quad \mathscr{H}_X \times \mathscr{H}_X \to \mathbb{R}$$

generalizes the matrix $\mathrm{Var}(E[X|Y])$ of the linear case, which is core of SIR for linear SDR. As mentioned in Lee, Li, and Chiaromonte (2013), this is why in the nonlinear SDR setting, the sample estimates of a similar operator is called the generalized sliced inverse regression(GSIR).

Thus, we can recover the central class $\mathscr{H}_{\mathcal{G}^{-(i,j)}}$ by computing the nonzero eigenfunctions of the self-adjoint operator

$$E_{XY} E^*_{XY}. \tag{2.1.3}$$

Based on the Lee, Li, and Zhao (2016a), $E_{XY} E^*_{XY}$ can be closely connected to the central class for nonlinear SDR.

Finally, under Assumptions and if $\mathscr{H}_{\mathcal{G}^{-(i,j)}}$ is complete, then we can induce

$$\overline{\text{ran}}(E_{XY}E_{XY}^*) = \mathscr{H}_{\mathcal{G}^{-(i,j)}}.$$

2.2 Graphical Models

Let $\mathscr{G} = (\Gamma, \mathcal{E})$ be an undirected graph consisting of a finite set of nodes $\Gamma = \{1, \ldots, p\}$ and set of edges

$$\mathcal{E} = \{(i,j) \in \Gamma \times \Gamma : i \neq j\}.$$

Since (i,j) and (j,i) represent the same edge in an undirected graph, we can assume without loss of generality that $i > j$. Thus the cardinality of $\mathcal{E}$ is $\binom{p}{2}$. A statistical graphical model links the graph $\mathscr{G}$ with a p-dimensional random vector X by the conditional independence among the components of X. Specifically, let $X = (X^1, \ldots, X^p)$ be a random vector with each component representing a node, and $X^{-(i,j)}$ the subvector of X with its ith and jth components removed. We define the edge set $\mathcal{E}$ by

$$(i,j) \notin \mathcal{E} \Leftrightarrow X^i \perp\!\!\!\perp X^j | X^{-(i,j)} \tag{2.2.1}$$

where $A \perp\!\!\!\perp B|C$ means "A and B are independent given C." In other words, nodes i and j are connected by an edge if and only if X^i and X^j are dependent given $X^{-(i,j)}$. Our goal is to estimate the set $\mathcal{E}$ based on an i.i.d. sample $X_1, \ldots, X_n$ of X and the defining relation (7.1.1). For a comprehensive exposition of graphical models, see (Lauritzen, 1996).

One of the most popular statistical graphical models is the Gaussian graphical model (GGM), which assumes that $X \sim N(\mu, \Sigma)$. This model is especially attractive because, under the multivariate Gaussian assumption, conditional independence in (7.1.1) is encoded in the precision matrix $\Theta = \Sigma^{-1}$ in the following sense

$$X^i \perp\!\!\!\perp X^j | X^{-(i,j)} \Leftrightarrow \theta_{ij} = 0, \tag{2.2.2}$$

where θ_{ij} is the (i,j)th entry of the precision matrix Θ. By this equivalence, estimating $\mathcal{E}$ amounts to identifying the positions of the zero entries of the precision matrix, which can be achieved by sparse estimation methods such as the LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), and the adaptive LASSO (Zou, 2006). A variety of methods have been developed for estimating the GGM, which include, for example, sparse-regulated linear regression (Meinshausen and Bühlmann, 2006), sparse-regulated MLE (Yuan and Lin, 2007),
thresholding (Bickel and Levina, 2008), and sparse partial correlation estimate (Peng et al., 2009). See also Friedman et al. (2008), Guo et al. (2010), and Lam and Fan (2009).

Because the Gaussian distribution assumption is rather restrictive, many recent advances have focussed on relaxing this assumption. A main challenge in developing a more flexible statistical graphical model is to avoid the curse of dimensionality: a straightforward nonparametric extension would resort to a high-dimensional kernel – as high as the size of the graph p. Such kernels are known to be ineffective because a neighborhood in a high-dimensional space contains very few data points, a phenomenon known as the curse of dimensionality (Bellman, 1961).

One way to relax the Gaussian assumption without evoking a high dimensional kernel is to replace the Gaussian distribution with the copula Gaussian distribution. This is the approach taken by Liu et al. (2009), Liu et al. (2012a), and Xue and Zou (2012), and is further extended to a class of transelliptical models by Liu et al. (2012c).

However, the copula Gaussian assumption could still be restrictive: for example, if (A, B) is a random vector with $B = A^2 + \epsilon$, where A and ϵ are i.i.d. $N(0,1)$, then no one-to-one transformation can render $(f_1(A), f_2(B))$ bivariate Gaussian. To further relax the distributional assumption, Li et al. (2014) proposed a new statistical three-way relation, *the additive conditional independence*, as an alternative criterion for constructing the graphical model. Additive conditional independence has the advantage of achieving nonparametric model flexibility without using a high-dimensional kernel. At the same time, additive conditional independence satisfies the same set of semi-graphoid axioms that govern the conditional independence (Dawid, 1979; Pearl and Verma, 1987), and is therefore a valid criterion for constructing a graphical model. In a related setting, Lee et al. (2016a) pro-

posed the additive partial correlation operator to achieve better scaling, leading to enhance estimation accuracy when characterizing additive conditional independence. More recently, Li and Solea (2018a) extend the additive graphical model to functional data. Other approaches to nonparametric graphical models include Fellinghauer et al. (2013) and Voorman et al. (2013).

Chapter 3

On Post Dimension Reduction Statistical Inference

3.1 Introduction

Despite the rapid advances of sufficient dimension reduction methodologies, however, there has been a lack of development on post dimension reduction inference. The outcome of SDR is a vector of sufficient predictors, but this is not the end of a typical data analysis. In most applications, the end product is an estimated statistical model, furnished with confidence intervals and p-values for statistical significance. Currently, the common practice is to feed the sufficient predictors obtained from SDR to the subsequent modeling as if they were the true predictors. It then proceeds with the usual model estimation and inference procedures, which completely ignores the estimation error incurred in the dimension reduction step, and thus tends to produce overly optimistic confidence intervals and p-values. More specifically, sufficient dimension reduction produces an estimate $\hat{\eta}$ of the η in (2.1.1), which, under mild regularity conditions, converges to η at the $n^{-1/2}$ rate. A subsequent modeling step builds a parametric probability model, say $f_\theta(\hat{\eta}^\intercal X, Y)$, which treats $\hat{\eta}^\intercal X$ as the new predictor, and from which an estimate $\hat{\theta}$ of θ is derived. In this process, the error in $\hat{\eta}$ contributes to the error in $\hat{\theta}$, and the contribution is in the same order of magnitude, i.e. $O_P(n^{-1/2})$, as the error in $\hat{\theta}$ when η is known. If we ignore the error propagated from $\hat{\eta}$, as the current solutions do, then the confidence interval for θ will be significantly narrower than the

true confidence interval, and the p-value for testing θ will be significantly smaller than the true p-value. Indeed, our data example in Section 7 shows that in some cases an inference method ignoring the error in $\hat{\eta}$ leads to a statistically significant conclusion, whereas an inference method that takes into account of the error in $\hat{\eta}$ leads to a statistically insignificant one. This lack of formal and rigorous post dimension reduction inference has seriously hindered the applications of sufficient dimension reduction.

In this article, we fill this gap by developing a general and comprehensive framework for post dimension reduction inference. The central issue for post reduction inference is to track how the error induced by dimension reduction propagates into the subsequent model estimation. To do so, we face the challenges that there are a large variety of dimension reduction methods, and as many different methods of estimating a statistical model. A useful post dimension reduction inference framework should be an open system that is capable of adapting to different dimension reduction and model estimation methods. Our idea is to use the influence functions of statistical functionals as a vehicle to achieve this generality. Many SDR methods can be expressed as eigenvectors of matrix-valued statistical functionals. As such, they can be expanded as asymptotic linear forms under mild regularity conditions (Bickel et al., 1993). Likewise, many estimation methods can also be expressed as vector-valued statistical functionals, which again can be expanded as asymptotic linear forms. These two asymptotic linear forms are uniquely determined by the influence functions of the statistical functionals for dimension reduction and estimation, and together would uniquely determine the post dimension reduction asymptotic distribution. Our post reduction framework is designed in such a way that one can input the influence functions of any dimension reduction method and any estimation method to produce the post reduction asymptotic distribution that takes both processes into account.

Within this general framework, we derive explicitly the influence functions for five popular SDR methods and three commonly used model estimation methods. The SDR methods include sliced inverse regression (SIR) (Li, 1991a), sliced average variance estimation (SAVE) (Cook and Weisberg, 1991), two forms of principal Hessian directions (y-PHD and r-PHD) (Li, 1992; Cook, 1998a), and directional regression (DR) (Li and Wang, 2007). The model estimation methods include

differentiable estimating equations, non-differentiable estimating equations, and generalized method of moments (GMM). We note that differentiable estimating equations include generalized linear model (McCullagh and Nelder, 1989) as a special case, whereas non-differentiable estimating equations include median and quantile regression as special cases. Moreover, generalized method of moments (Hansen, 1982; Hansen et al., 1996) have been widely used in econometrics. These 5×3 combinations of SDR and estimation methods cover a wide range of statistical modeling and applications. They also serve as an illustration on how to derive the influence functions and how to plug them into our post dimension reduction inference framework to obtain the desired post reduction asymptotic distribution. As such, more SDR and estimation methods can be incorporated into this framework.

Based on the derived post dimension reduction asymptotic distribution, we proceed further to develop specific methods for conducting statistical inference: constructing confidence intervals and test statistics, and computing the asymptotic null and local alternative distributions of the test statistics. It is our hope that the materials developed in this project can serve as a first step towards incorporating sufficient dimension reduction and post reduction inference into a systematic and comprehensive statistical method.

3.2 General framework post reduction inference

We begin with introduction of two statistical functionals: one for sufficient dimension reduction, which we call the *reduction functional*, and one for model estimation, which we call the *estimation functional*. We then define the composite functional and derive its influence function, from which we obtain the post dimension reduction asymptotic distribution. Finally, we explicitly compare the asymptotic covariance of the estimated parameter with and without taking into account the error induced by dimension reduction.

3.2.1 Reduction, estimation and composite functionals

Let (X, Y) be random vectors in $\mathbb{R}^p \times \mathbb{R}$ that take values in the measurable space $(\Omega_{XY}, \mathcal{F}_{XY})$. Let $\mathcal{P}$ the class of all probability distributions of (X, Y). Let $\mathcal{S}$ be a metric space, which in our context is taken as a space of matrices. A statistical

functional is a mapping R from $\mathcal{P}$ to $\mathcal{S}$. Let F_0 be the true distribution of (X, Y), let (x, y) be a fixed point in Ω_{XY}, and let δ_{xy} be the Dirac measure at (x, y). The *influence function* of the functional R is defined as

$$R^\star(x, y) = \frac{\partial}{\partial \epsilon} R\left[(1-\epsilon)F_0 + \epsilon\delta_{xy}\right]\big|_{\epsilon=0}.$$

For more details about influence functions, see Bickel et al. (1993). Throughout this project, we assume that $R^\star$ satisfies the following conditions.

Assumption 6.

(1) $E[R^\star(X, Y)] = 0$.

(2) $R^\star(X, Y)$ *has finite variance; if* $R^\star(X, Y)$ *is a random vector or a random matrix, then its entries have finite variances.*

These assumptions are mild and hold for all the SDR methods considered in this project. For a set of sufficient conditions for these assumptions, see Bickel et al (1993), page 19. When there is no ambiguity, we abbreviate $R^\star(X, Y)$ by $R^\star$. In the following, an asterisk on a symbol always indicates the influence function of a statistical functional represented by that symbol. For example, for the statistical functionals $\Phi(F, \eta)$ and $\Lambda(F)$ discussed below, $\Phi^\star$ and $\Lambda^\star$ represent their respective influence functions.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. samples of (X, Y). Let F_n be the empirical distribution based on this sample. It is well known that, if R is Hadamard differentiable, then $R(F_n)$ has the following expansion,

$$R(F_n) = R(F_0) + E_n(R^\star) + o_p(n^{-\frac{1}{2}}), \tag{3.2.1}$$

where $E_n(R^\star)$ denotes the sample average $n^{-1}\sum_{i=1}^n R^\star(X_i, Y_i)$. Consequently, by the central limit theorem,

$$\sqrt{n}[R(F_n) - R(F_0)] \xrightarrow{\mathcal{D}} N(0, \mathrm{var}(R^\star)). \tag{3.2.2}$$

Thus, the influence function $R^\star$ uniquely determines the asymptotic distribution of $R(F_n)$. Conventionally, $R(F_n)$ represents a statistic, and $R(F_0)$ the parame-

ter it estimates. For more information about statistical functionals and influence functions, see, e.g., Fernholz (2012), Bickel et al. (1993), and Li (2018b).

We first define the reduction functional. Most SDR methods can be written in the form of a generalized eigen-decomposition problem. That is, there is a statistical functional $\Lambda : \mathcal{P} \to \mathbb{R}^{p\times p}$ satisfying that

$$\Sigma(F_0)^{-1}\text{span}[\Lambda(F_0)] \subseteq \mathcal{S}_{Y|X}, \tag{3.2.3}$$

where $\Sigma(F_0)$ denotes the covariance matrix of X. The relation (3.2.3) implies that the central subspace $\mathcal{S}_{Y|X}$ can be recovered by solving the generalized eigenvalue problem

$$\Lambda(F_0)v = \lambda\Sigma(F_0)v. \tag{3.2.4}$$

Let $\eta = (\eta_1, \ldots, \eta_r)$ denote its first r eigenvectors, where r is the rank of $\Lambda(F_0)$ and $r \leq q$. For many SDR methods, the equality in (3.2.3) holds, and correspondingly, $r = q$. In this case, we say the SDR method is exhaustive. See Li et al. (2005) and Li and Wang (2007) for sufficient conditions for exhaustiveness. For simplicity, we assume the SDR method is exhaustive in this article; i.e., $\mathcal{S}_{Y|X}$ can be fully recovered by $\text{span}(\eta_{01}, \ldots, \eta_{0r})$. We also note that, the generalized eigenvalue problem in (3.2.4) can be solved by transforming it into a standard eigenvalue problem. That is, if $\{\beta_{0i}\}_{i=1}^r$ are the first r eigenvectors of $\Sigma(F_0)^{-1/2}\Lambda(F_0)\Sigma(F_0)^{-1/2}$, then $\eta_{0i} = \Sigma(F_0)^{-\frac{1}{2}}\beta_{0i}$, $i = 1, \ldots, r$, are the first r eigenvectors of the generalized eigenvalue problem (3.2.4). Given i.i.d. samples of (X, Y), the corresponding sample version of (3.2.4) is $\Lambda(F_n)v = \lambda\Sigma(F_n)v$, where $\Sigma(F_n)$ is the sample covariance matrix of X. We define $\{\hat{\eta}_i\}_{i=1}^r$ and $\{\hat{\beta}_i\}_{i=1}^r$ accordingly.

We call the functional $\Lambda(F)$ the *reduction functional*, and assume it is Hadamard differentiable with the influence function $\Lambda^\star$. Correspondingly, we use $\eta(F)$ to denote the $\mathbb{R}^{p\times q}$-valued statistical functional of the first r eigenvectors of $\Lambda(F)$.

We next define the estimation functional. We start with a set of fixed eigenvectors $(\eta_1, \ldots, \eta_q)$ that form an orthonormal set in $\mathbb{R}^p$. Suppose we replace the original p-dimensional predictor vector X with the q-dimensional sufficient predictor $\eta^T X$, then fit some parametric regression model with the model parameter

θ. Assume, for a fixed η, the estimate of θ takes the following general form of a statistical functional

$$\Phi : \mathcal{P} \times \mathbb{R}^{p\times q} \to \Theta \subseteq \mathbb{R}^s,$$

where Θ is the parameter space for the parametric regression model. We call the functional Φ the *estimation functional*, and assume that, for each fixed η, the mapping $F \mapsto \Phi(F, \eta)$ is Hadamard differentiable with the influence function $\Phi^\star$. Since we treat η as fixed, this functional corresponds to the naive estimator as if η is known.

Now we replace the fixed η with the estimate $\hat{\eta} = \eta(F_n)$ from a given SDR method, which leads to an estimate of θ, $T(F_n) = \Phi[F_n, \eta(F_n)]$, and the functional

$$T : \ \mathcal{P} \to \Theta, \quad F \mapsto \Phi[F, \eta(F)].$$

We call it the *composite functional*, as it is a composition of the reduction functional $\Lambda(F)$, which is implicitly contained in $\eta(F)$, and the estimation functional $\Phi(F, \eta)$. The functional T accounts for the variations in both dimension reduction and estimation, and its influence function determines the post dimension reduction asymptotic distribution. It corresponds to the inference procedure that does not pretend η is known.

3.2.2 Influence function and asymptotic distribution

Next we derive the influence function of $T(F)$ given the influence functions $\Lambda^\star$ and $\Phi^\star$. We derive the influence functions $\Lambda^\star$ and $\Phi^\star$ for a variety of dimension reduction and estimation methods in Sections 3.3 and 3.4, respectively. In the following, we use $\otimes$ to denote the Kronecker product. We denote $\Sigma(F_n), \Sigma(F_0), \Sigma(F)$ by $\hat{\Sigma}, \Sigma_0, \Sigma$, and denote $\Lambda(F_n), \Lambda(F_0), \Lambda(F)$ by $\hat{\Lambda}, \Lambda_0, \Lambda$, respectively. We first need the following lemma, whose proof can be found in Li (2018b).

Lemma 1. *Suppose all moments involved are finite. Then*

(1) $\mathrm{vec}(\Sigma^\star) = X \otimes X - E(X \otimes X) - [X - E(X)] \otimes E(X) - E\{X \otimes [X - E(X)]\}$;

(2) $\mathrm{vec}[(\Sigma^{-\frac{1}{2}})^\star] = -(\Sigma_0^{1/2} \otimes \Sigma_0 + \Sigma_0 \otimes \Sigma_0^{1/2})^{-1}\mathrm{vec}(\Sigma^\star)$;

(3) $(\Sigma^{-1})^\star = -\Sigma^{-1}\Sigma^\star\Sigma^{-1}$

Theorem 1. *Suppose the following conditions are satisfied.*

(C1) The statistical functionals $F \mapsto \Lambda(F)$ *and* $F \mapsto \Phi(F, \eta)$ *are Hadamard differentiable with influence functions* $\Lambda^\star(X, Y)$ *and* $\Phi^\star(X, Y, \eta)$*. Both* $\Lambda^\star$ *and* $\Phi^\star$ *satisfy Assumption 6.*

(C2) The function $\eta \mapsto \Phi(F_0, \eta)$ *is differentiable.*

(C3) All the nonzero eigenvalues of $\Sigma_0^{-1/2}\Lambda_0\Sigma_0^{-1/2}$ *are distinct.*

Then the influence function of $T(F)$ *is*

$$T^\star(X, Y) = \Phi^\star(X, Y, \eta_0) + DC \begin{pmatrix} \mathrm{vec}[\Sigma^\star(X, Y)] \\ \mathrm{vec}[\Lambda^\star(X, Y)] \end{pmatrix},$$

where $D = \partial\Phi(F_0, \eta_0)/\partial\mathrm{vec}(\eta)^\intercal$ *and* $C = (A, B)$*, in which*

$$\begin{aligned} A =& -[\beta_0^\intercal \otimes I_p + (I_q \otimes \Sigma_0^{-1/2})H(\Sigma_0^{-1/2}\Lambda_0 \otimes I_p + I_p \otimes \Lambda_0\Sigma_0^{-1/2})] \\ & (\Sigma_0 \otimes \Sigma_0^{\frac{1}{2}} + \Sigma_0^{\frac{1}{2}} \otimes \Sigma_0)^{-1}, \\ B =& (I_q \otimes \Sigma_0^{-1/2})H(\Sigma_0^{-1/2} \otimes \Sigma_0^{-1/2}) \\ H =& (H_1^\intercal, \ldots, H_q^\intercal)^\intercal, \\ H_i =& \beta_{0i}^\intercal \otimes \left[\sum_{j=1, j\neq i}^{p} (\lambda_{0i} - \lambda_{0j})^{-1}(\beta_{0j}\beta_{0j}^\intercal) \right], \ i = 1, \ldots, q. \end{aligned}$$

Proof. Recall that the sample estimator of η_{0i} is $\hat{\eta}_i = \hat{\Sigma}^{-1/2}\hat{\beta}_i$, where $\hat{\beta}_i$ is ith eigenvector of $\hat{\Sigma}^{-1/2}\hat{\Lambda}\hat{\Sigma}^{-1/2}$, $i = 1, \ldots, q$. Thus the influence function of $\hat{\eta}_i$ is

$$\eta_i^\star = (\Sigma^{-1/2})^\star\beta_{0i} + \Sigma_0^{-1/2}\beta_i^\star.$$

Furthermore, by Zhu and Fang (1996), the influence function of $\hat{\beta}_i$ is

$$\beta_i^\star = \sum_{j=1, j\neq i}^{p} \frac{\beta_{0j}\beta_{0j}^\intercal(\Sigma^{-1/2}\Lambda\Sigma^{-1/2})^\star\beta_{0i}}{\lambda_{0i} - \lambda_{0j}} = H_i\mathrm{vec}[(\Sigma^{-1/2}\Lambda\Sigma^{-1/2})^\star], \tag{3.2.5}$$

where

$$H_i = \beta_{0i}^{\mathsf{T}} \otimes \left[\sum_{j=1, j\neq i}^{p} (\lambda_{0i} - \lambda_{0j})^{-1} (\beta_{0j}\beta_{0j}^{\mathsf{T}}) \right].$$

By Lemma 1 and some simple calculation,

$$\begin{aligned}
&\mathrm{vec}[(\Sigma^{-1/2}\Lambda\Sigma^{-1/2})^\star] \\
&= -(\Sigma_0^{-1/2}\Lambda_0 \otimes I_p + I_p \otimes \Lambda_0\Sigma_0^{-1/2})(\Sigma_0^{1/2} \otimes \Sigma_0 + \Sigma_0 \otimes \Sigma_0^{1/2})^{-1}\mathrm{vec}(\Sigma^\star) \\
&\quad + (\Sigma_0^{-1/2} \otimes \Sigma_0^{-1/2})\mathrm{vec}(\Lambda^\star).
\end{aligned}$$

Combination of (3.2.5) and the above equality yields

$$\begin{aligned}
&\mathrm{vec}(\beta^\star) \\
&= -H(\Sigma_0^{-1/2}\Lambda_0 \otimes I_p + I_p \otimes \Lambda_0\Sigma_0^{-1/2})(\Sigma_0^{1/2} \otimes \Sigma_0 + \Sigma_0 \otimes \Sigma^{1/2})^{-1}\mathrm{vec}(\Sigma^\star) \\
&\quad + H(\Sigma_0^{-1/2} \otimes \Sigma_0^{-1/2})\mathrm{vec}(\Lambda^\star),
\end{aligned}$$

where $H = (H_1^{\mathsf{T}}, \ldots, H_q^{\mathsf{T}})^{\mathsf{T}}$. Hence

$$\begin{aligned}
\mathrm{vec}(\eta^\star) &= \mathrm{vec}[(\Sigma^{-\frac{1}{2}})^\star\beta_0 + \Sigma_0^{-\frac{1}{2}}\beta^\star] \\
&= -(\beta_0^{\mathsf{T}} \otimes I_p)(\Sigma_0^{\frac{1}{2}} \otimes \Sigma_0 + \Sigma_0 \otimes \Sigma_0^{\frac{1}{2}})^{-1}\mathrm{vec}(\Sigma^\star) + (I_q \otimes \Sigma_0^{-\frac{1}{2}})\mathrm{vec}(\beta^\star) \\
&= C \begin{pmatrix} \mathrm{vec}(\Sigma^\star) \\ \mathrm{vec}(\Lambda^\star) \end{pmatrix},
\end{aligned}$$

where C is as defined in the theorem. By condition (C2) and the chain rule for differentiation, we have

$$T^\star(X, Y) = \Phi^\star(X, Y, \eta_0) + D\,\mathrm{vec}[\eta^\star(X, Y)],$$

which completes the proof. □

Condition (C1) of Theorem 1 is mild as most Λ matrices in SDR are functions of sample moments, which are Hadamard differentiable if the moments of X and Y up to a certain order are finite. Condition (C2) is also mild and is easy to verify. As we will see in Section 3.2.4, Condition (C3) is also satisfied by numerous

SDR methods and statistical models. Based on Theorem 1, we next derive the asymptotic distribution of $\hat{\theta} = \Phi(F_n, \hat{\eta})$.

Corollary 1. *Suppose the conditions in Theorem 1 are satisfied. Then*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} N(0, \Gamma),$$

where $\Gamma = (I_p, DC)B(I_p, DC)^\intercal$ *and*

$$B = \begin{pmatrix} E(\Phi^\star \Phi^{\star\intercal}) & E[\Phi^\star \mathrm{vec}(\Sigma^\star)^\intercal] & E[\Phi^\star \mathrm{vec}(\Lambda^\star)^\intercal] \\ E[\mathrm{vec}(\Sigma^\star)\Phi^{\star\intercal}] & E[\mathrm{vec}(\Sigma^\star)\mathrm{vec}(\Sigma^\star)^\intercal] & E[\mathrm{vec}(\Sigma^\star)\mathrm{vec}(\Lambda^\star)^\intercal] \\ E[\mathrm{vec}(\Lambda^\star)\Phi^{\star\intercal}] & E[\mathrm{vec}(\Lambda^\star)\mathrm{vec}(\Sigma^\star)^\intercal] & E[\mathrm{vec}(\Lambda^\star)\mathrm{vec}(\Lambda^\star)^\intercal] \end{pmatrix}.$$

Proof. By Theorem 1 and the relation (3.2.1) between the influence function and its asymptotic linear form, we have

$$\hat{\theta} = \theta_0 + \Big(I_p,\ DC\Big)\, E \begin{pmatrix} \Phi^\star(X, Y, \eta_0) \\ \mathrm{vec}[\Sigma^\star(X, Y)] \\ \mathrm{vec}[\Lambda^\star(X, Y)] \end{pmatrix} + o_p(n^{-1/2}).$$

Then applying (3.2.2) completes the proof. □

At the sample level, Σ_0, λ_{0i}, β_0 in the matrix C are estimated by $\hat{\Sigma}$, $\hat{\lambda}_i$ and $\hat{\beta}$. The matrix D is estimated by $\partial\Phi(F_n, \eta_0)/\partial\mathrm{vec}(\eta)^\intercal$. This is justified by

$$\frac{\partial\Phi(F_n, \eta_0)}{\partial\mathrm{vec}(\eta)^\intercal} \xrightarrow{P} D,$$

which holds under mild regularity conditions.

3.2.3 Asymptotic comparison of naive and objective inference

We compare the asymptotic covariance of the parameter estimate $\hat{\theta} = T(F_n) = \Phi(F_n, \eta(F_n))$ that takes into account the estimation error induced by dimension reduction, and the naive estimate $\tilde{\theta}(\eta_0) = \Phi(F_n, \eta_0)$ that does not. We denote their corresponding asymptotic covariance matrix by $\Gamma(\eta_0, \theta_0)$ and $\tilde{\Gamma}(\eta_0, \theta_0)$, respectively. Given the data, $\Gamma(\eta_0, \theta_0)$ and $\tilde{\Gamma}(\eta_0, \theta_0)$ are estimated by $\Gamma(\hat{\eta}, \hat{\theta})$ and $\tilde{\Gamma}(\hat{\eta}, \hat{\theta})$. Since $\hat{\eta}$

and $\hat{\theta}$ are root-n consistent and Γ and $\tilde{\Gamma}$ are differentiable, the differences, $\Gamma(\hat{\eta}, \hat{\theta}) - \Gamma(\eta_0, \theta_0)$ and $\tilde{\Gamma}(\hat{\eta}, \hat{\theta}) - \tilde{\Gamma}(\eta_0, \theta_0)$, are both of the order $O_P(n^{-1/2})$. Thus it suffices to compare $\Gamma(\eta_0, \theta_0)$ with $\tilde{\Gamma}(\eta_0, \theta_0)$. The next theorem characterizes the amount of the asymptotic variance increase after taking the dimension reduction error into account.

Theorem 2. *Suppose the conditions in Theorem 1 are satisfied. Moreover, suppose when η_0 is known, $\tilde{\theta}(\eta_0)$ is an efficient estimator of θ_0. Then*

$$\begin{aligned}&\Gamma(\eta_0, \theta_0) - \tilde{\Gamma}(\eta_0, \theta_0) \\ &\qquad = DC \begin{pmatrix} E[\mathrm{vec}(\Sigma^\star)\mathrm{vec}(\Sigma^\star)^\intercal] & E[\mathrm{vec}(\Sigma^\star)\mathrm{vec}(\Lambda^\star)^\intercal] \\ E[\mathrm{vec}(\Lambda^\star)\mathrm{vec}(\Sigma^\star)^\intercal] & E[\mathrm{vec}(\Lambda^\star)\mathrm{vec}(\Lambda^\star)^\intercal] \end{pmatrix} C^\intercal D^\intercal\end{aligned}$$

Proof. The proof echoes the Hajek-LeCam convolution theorem of regular estimators (Bickel et al., 1993). Since, when η_0 is given, both $\hat{\theta}$ and $\tilde{\theta}(\eta_0)$ are regular estimators of θ_0, and $\tilde{\theta}(\eta_0)$ is efficient, by the LeCam-Hajek convolution theorem, $\sqrt{n}(\hat{\theta}(\eta_0) - \theta_0)$ can be decomposed into the sum of two asymptotically independent terms

$$\begin{aligned}&\sqrt{n}(\tilde{\theta}(\eta_0) - \theta_0) + [\sqrt{n}(\hat{\theta} - \theta_0) - \sqrt{n}(\tilde{\theta}(\eta_0) - \theta_0)] \\ &\qquad\qquad = \sqrt{n}E(\Phi^\star) + \sqrt{n}E(T^\star - \Phi^\star) + o_P(1),\end{aligned}$$

which implies that $E[\Phi^\star(T^\star - \Phi^\star)^\intercal] = 0$. Hence

$$\mathrm{var}[T^\star(X, Y)] = \mathrm{var}[\Phi^\star(X, Y, \eta_0)] + D\ \mathrm{var}\{\mathrm{vec}[\eta^\star(X, Y)]\}\ D^\intercal.$$

Substituting the form of $\mathrm{vec}(\eta^\star)$ into this equation completes the proof. □

3.2.4 Identifiability of reduction parameter

Here we briefly discuss the subtle issue of the identifiability for the reduction parameters. In the framework of SDR with the structural dimension $q > 1$, the basis $(\gamma_1, \ldots, \gamma_q)$ of $\mathcal{S}_{Y|X}$ is not identifiable. However, in practice, we always use a specific SDR method, say SIR, to estimate $\mathcal{S}_{Y|X}$. A specific SDR method, when applied

to a specific statistical model, almost always yields a fixed set of eigenvectors in $\mathcal{S}_{Y|X}$ up to a sign. Thus, if we agree to take, for example, the first nonzero component of the relevant eigenvectors to be positive, then we have a well identified set of reduction dimension parameters. As an example, for Model III and Model IV in Section 6.1, the structural dimension $q = 2$ and the first two eigenvalues of $\Sigma_0^{-1/2}\Lambda_0\Sigma_0^{-1/2}$ for DR are, respectively, 1.30, 1.25 and 1.54, 1.35. These distinct population-level eigenvalues give rise to well identified reduction parameters β_1 and β_2. A parametric statistical model can then be imposed upon the predictors $\beta_1^\intercal X$ and $\beta_2^\intercal X$ without ambiguity.

3.3 Influence functions for estimation functionals

The asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$ relies on the reduction influence function $\Lambda^\star(X, Y)$, the estimation influence function $\Phi^\star(X, Y, \eta)$, and the form of $D = \partial\Phi(F_0, \eta_0)/\partial\text{vec}(\eta)^\intercal$. In this section, we derive the explicit forms of the influence function $\Phi^\star(X, Y, \eta)$ and the derivative D for three estimation methods: the differentiable estimating equations, the non-differentiable estimating equations, and the generalized method of moments. They cover a wide variety of regression methods, including generalized linear model, nonlinear mean regression, and nonlinear median and quantile regression, among others.

3.3.1 Differentiable estimating equations

Many commonly used parametric models can be formulated as special cases of a general class of estimator of θ, which is defined as the solution to the estimating equations

$$E[g(\theta, X, Y)] = 0, \tag{3.3.1}$$

where $E_\theta[g(\theta, X, Y)] = 0$, $\text{var}_\theta[g(\theta, X, Y)]$ is a matrix with finite entries, and the dimension of g is the same as the dimension of θ. One example is generalized linear model, which can be expressed as the solution to the estimating equations

$$E\left\{\frac{\partial\mu(\theta^\intercal X)}{\partial\theta^\intercal}V^{-1}(\theta^\intercal X)[Y - \mu(\theta^\intercal X)]\right\} = 0,$$

where $\mu(\theta^\tau X) = E(Y|\theta^\tau X)$, and $V(\theta^\tau X) = \text{var}(Y|\theta^\tau X)$. See, for example, McCullagh and Nelder (1989) and Li (1993). Another example is the parametric nonlinear regression, where we minimize the objective function $E[Y - h(\theta^\tau X)]^2$, and h can take a polynomial form, $h(u_1, \ldots, u_k) = \sum_{i=1}^k \theta_i u_i + \sum_{i,j=1}^k \theta_{ij} u_i u_j$. Correspondingly, the parameter θ can be expressed as the solution to the estimating equations

$$E\left\{2\frac{\partial h(\theta^\tau X)}{\partial \theta}[Y - h(\theta^\tau X)]\right\} = 0.$$

In our context of SDR based parametric modeling, the predictor vector X is replaced by the sufficient predictor $\eta^\tau X$. The statistical functional of the estimator θ in (3.3.1) is $\Phi(F, \eta)$, which is implicitly defined by the equation $\int g[\Phi(F,\eta), \eta^\tau X, Y]dF = 0$. We next derive the explicit forms of the corresponding influence function $\Phi^\star$ and the derivative D, and summarize the results in the next proposition.

Proposition 1. *For the estimating equations (3.3.1), we have*

$$\Phi^\star(X, Y, \eta) = -\left\{E\left[\frac{\partial g(\theta_0, \eta^\tau X, Y)}{\partial \theta^\tau}\right]\right\}^{-1} g(\theta_0, \eta^\tau X, Y),$$
$$D = -\left\{E\left[\frac{\partial g(\theta_0, \eta_0^\tau X, Y)}{\partial \theta^\tau}\right]\right\}^{-1} E\left[\frac{\partial g(\theta_0, \eta_0^\tau X, Y)}{\partial u}(I_q \otimes X^\tau)\right].$$

Proof. Let $F_\epsilon = (1-\epsilon)F_0 + \epsilon\delta_{XY}$. Then, for all $\epsilon \in [0,1]$, we have $\int g[\Phi(F_\epsilon, \eta), \eta^\tau X, Y]dF_\epsilon = 0$. Differentiating (3.3.1) with respect to ϵ, and evaluating the derivatives at $\epsilon = 0$, we have

$$\frac{\partial}{\partial \epsilon}\int g(\Phi(F_\epsilon, \eta), \eta^\tau X, Y)dF_\epsilon\Big|_{\epsilon=0}$$
$$= \left[\int \frac{\partial g(\theta_0, \eta^\tau X, Y)}{\partial \theta^\tau} dF_0\right]\Phi^\star + \int g(\theta_0, \eta^\tau X, Y)d(\delta_{XY} - F_0) = 0.$$

Since $Eg(\theta_0, \eta^\tau, Y) = 0$, the second term on the right-hand side is simply $g(\theta_0, \eta^\tau X, Y)$, which leads to the desired form for $\Phi^\star(X, Y, \eta)$.

Next, we note that $\Phi(F_0, \eta)$ satisfies $E[g(\Phi(F_0, \eta), \eta^\tau X, Y)] = 0$. Differentiating

this equation with respect to $\mathrm{vec}(\eta)$, we have

$$\left[\int \frac{\partial}{\partial \theta^\intercal} g(\theta_0, \eta^\intercal X, Y) dF_0\right] \frac{\partial \Phi(F_0, \eta)}{\partial \mathrm{vec}(\eta)^\intercal} + \int \frac{\partial}{\partial u} g(\theta_0, \eta^\intercal X, Y) \frac{\partial(\eta^\intercal X)}{\partial \mathrm{vec}(\eta)^\intercal} dF_0 = 0,$$

where $\partial g/\partial u$ denotes the partial derivative with respect to the second argument of g, which is $\eta^\intercal X$. Since $\eta^\intercal X = \mathrm{vec}(X^\intercal \eta) = \mathrm{vec}(X^\intercal \eta I_q) = (I_q \otimes X^\intercal)\mathrm{vec}(\eta)$, we have $\partial(\eta^\intercal X)/\partial \mathrm{vec}(\eta)^\intercal = I_q \otimes X^\intercal$. Henceforth,

$$E\left[\frac{\partial g(\theta_0, \eta_0^\intercal X, Y)}{\partial \theta^\intercal}\right] D + E\left[\frac{\partial g(\theta_0, \eta_0^\intercal X, Y)}{\partial u}(I_q \otimes X^\intercal)\right] = 0.$$

Solving this equation yields the desired form for D. □

3.3.2 Non-differentiable estimating equations

Another family of popular models can be formulated as solving a set of non-differentiable estimating equations. Examples include nonlinear quantile regression (He et al., 2003; Wang and Wang, 2009) and support vector regression (Smola and Scholkopt, 2004). In this section, we use nonlinear quantile regression as an illustration. The derivation of the estimation functional for other models follow in a similar fashion.

For a number $\tau \in [0, 1]$, define the function $\rho : \mathbb{R} \to \mathbb{R}^+$ as $\rho_\tau(u) = \tau u$ if $u > 0$, and $-(1-\tau)u$ if $u < 0$. Let $m(\eta^\intercal X, \theta)$ be a function such that, for the true value (η_0, θ_0) of (η, θ), it is the τ-th conditional quantile, $P[Y \leq m(\eta_0^\intercal X, \theta_0)|X] = \tau$. At the population level, nonlinear quantile regression is defined as minimizing the objective function $E\{\rho_\tau[Y - m(\eta^\intercal X, \theta)]\}$ over $\theta \in \mathbb{R}^d$, which amounts to solving the estimating equations

$$E\left\{\dot{\rho}_\tau[Y - m(\eta^\intercal X, \theta)] \frac{\partial m(\eta^\intercal X, \theta)}{\partial \theta}\right\} = 0, \tag{3.3.2}$$

where $\dot{\rho}_\tau(u) = \tau I(u > 0) - (1-\tau) I(u \leq 0) = \tau - I(u \leq 0)$. Rigorously speaking, $\dot{\rho}_\tau$ is not defined at $u = 0$. But since $u = 0$ has measure 0, we can assign any value to $\dot{\rho}(0)$; in our case, we set $\dot{\rho}(0)$ equal to $-(1-\tau)$.

Next we write the first argument $\eta^\intercal X$ of $m(\eta^\intercal X, \theta)$ as u, and use the following

notations for partial derivatives, $\dot{m}_u = \partial m/\partial u, \dot{m}_\theta = \partial m/\partial\theta, \ddot{m}_{uu} = \partial^2 m/\partial u\partial u^\intercal$, $\ddot{m}_{u\theta} = \partial^2 m/\partial u\partial\theta^\intercal$, and $\ddot{m}_{\theta\theta} = \partial^2 m/\partial\theta\partial\theta^\intercal$. We derive the influence function $\Phi^\star$ and the derivative D in the next proposition.

Proposition 2. *For the estimating equations (3.3.2), we have*

$$\begin{aligned}
\Phi^\star(X,Y,\eta) =& \left(E\left\{f_{Y|X}[m(\eta_0^\intercal X,\theta_0)|X]\dot{m}_\theta(\eta_0^\intercal X,\theta_0)\dot{m}_\theta^\intercal(\eta_0^\intercal X,\theta_0)\right\}\right)^{-1} \\
& \left\{\tau - I[Y \le m(\eta_0^\intercal X,\theta_0)]\right\}\dot{m}_\theta(\eta_0^\intercal X,\theta_0), \\
D =& -\left(E\{\dot{m}_\theta(\eta_0^\intercal X,\theta_0)\dot{m}_\theta^\intercal(\eta_0^\intercal X,\theta_0)f_{Y|X}[m(\eta_0^\intercal,\theta_0)|x]\}\right)^{-1} \\
& E\left\{\dot{m}_\theta(\eta_0^\intercal X,\theta_0)\dot{m}_u^\intercal(\eta_0^\intercal X,\theta_0)(I_q\otimes X^\intercal)f_{Y|X}[m(\eta_0^\intercal X,\theta_0)|x]\right\}.
\end{aligned}$$

Proof. Denote $A(F,\eta_0) = \int \dot{\rho}_\tau\{Y - m[\eta_0^\intercal X,\Phi(F,\eta_0)]\}\dot{m}_\theta[\eta_0^\intercal X,\Phi(F,\eta_0)]dF$. The influence function $\Phi^\star(X,Y,\eta_0)$ can be obtained from the equation

$$\frac{\partial}{\partial\epsilon}A(F_\epsilon,\eta_0)\Big|_{\epsilon=0} = 0.$$

In the following, we abbreviate $\partial f(\epsilon)/\partial\epsilon\big|_{\epsilon=0}$ by $\partial f(\epsilon)/\partial\epsilon$. By the chain rule, we decompose the above derivative into three terms:

$$\frac{\partial}{\partial\epsilon}A(F_\epsilon,\eta_0) = \frac{\partial}{\partial\epsilon}A_1(F_\epsilon,\eta_0) + \frac{\partial}{\partial\epsilon}A_2(F_\epsilon,\eta_0) + \frac{\partial}{\partial\epsilon}A_3(F_\epsilon,\eta_0), \quad \text{where} \tag{3.3.3}$$

$$\begin{aligned}
A_1(F_\epsilon,\eta_0) &= \int \dot{\rho}_\tau\{Y - m[\eta_0^\intercal X,\Phi(F_\epsilon,\eta_0)]\}\dot{m}_\theta[\eta_0^\intercal X,\Phi(F_0,\eta_0)]dF_0, \\
A_2(F_\epsilon,\eta_0) &= \int \dot{\rho}_\tau\{Y - m[\eta_0^\intercal X,\Phi(F_0,\eta_0)]\}\dot{m}_\theta[\eta_0^\intercal X,\Phi(F_\epsilon,\eta_0)]dF_0, \\
A_3(F_\epsilon,\eta_0) &= \int \dot{\rho}_\tau\{Y - m[\eta_0^\intercal X,\Phi(F_0,\eta_0)]\}\dot{m}_\theta[\eta_0^\intercal X,\Phi(F_0,\eta_0)]dF_\epsilon.
\end{aligned}$$

The term $\partial A_1(F_\epsilon, \eta_0)/\partial\epsilon$ can be written as

$$\begin{aligned}
&\frac{\partial}{\partial\epsilon}A_1(F_\epsilon, \eta_0) \\
&= \frac{\partial}{\partial\epsilon}\int \left(\tau - I\{Y \le m[\eta_0^\intercal X, \Phi(F_\epsilon, \eta_0)]\}\right) \dot{m}_\theta[\eta_0^\intercal X, \Phi(F_0, \eta_0)]dF_0 \\
&= -\frac{\partial}{\partial\epsilon}\int I\{Y \le m[\eta_0^\intercal X, \Phi(F_\epsilon, \eta_0)]\}\dot{m}_\theta[\eta_0^\intercal X, \Phi(F_0, \eta_0)]dF_0 \\
&= -\int_{\Omega_X}\frac{\partial}{\partial\epsilon}\int_{-\infty}^{m[\eta_0^\intercal X, \Phi(F_\epsilon, \eta_0)]} f_{Y|X}(y|x)dy\, \dot{m}_\theta[\eta_0^\intercal X, \Phi(F_0, \eta_0)]f_X(x)dx \\
&= -\left\{\int_{\Omega_X} f_{Y|X}[m(\eta_0^\intercal X, \theta_0)|x]\dot{m}_\theta(\eta_0^\intercal X, \theta_0)\dot{m}_\theta^\intercal(\eta_0^\intercal X, \theta_0)f_X(x)dx\right\}\Phi^\star \\
&= -E\left\{f_{Y|X}[m(\eta_0^\intercal X, \theta_0)|x]\dot{m}_\theta(\eta_0^\intercal X, \theta_0)\dot{m}_\theta^\intercal(\eta_0^\intercal X, \theta_0)\right\}\Phi^\star,
\end{aligned} \tag{3.3.4}$$

where the first equality is by the definition of $\dot{\rho}_\tau(u)$, the second equality is because $\tau\int \dot{m}_\theta(\eta_0^\intercal X, \Phi(F_0, \eta_0))dF_0$ does not depend on ϵ, and the fourth equality is because $\Phi(F_0, \eta_0) = \theta_0$.

The term $\partial A_2(\epsilon, \eta)/\partial\epsilon$ can be written as

$$\begin{aligned}
\frac{\partial}{\partial\epsilon}A_2(\epsilon, \eta) &= \int \dot{\rho}_\tau[Y - m(\eta_0^\intercal X, \theta_0)]\ddot{m}_{\theta\theta}(\eta_0^\intercal X, \theta_0)dF_0\Phi^\star \\
&= E\left[E\big\{\dot{\rho}_\tau[Y - m(\eta_0^\intercal X, \theta_0)]\big|X\big\}\ddot{m}_{\theta\theta}(\eta_0^\intercal X, \theta_0)\right]\Phi^\star \\
&= 0,
\end{aligned} \tag{3.3.5}$$

where the last equality is due to that, since $m(\eta_0^\intercal X, \theta_0)$ is the τ-th conditional quantile,

$$E\{\dot{\rho}_\tau[Y - m(\eta_0^\intercal X, \theta_0)]\big|X\} = E\{\tau - I[Y \le m(\eta_0^\intercal X, \theta_0)]\big|X\} = 0.$$

The term $\partial A_3(F_\epsilon, \eta)/\partial\epsilon$ can be written as

$$\begin{aligned}
&\frac{\partial}{\partial\epsilon}A_3(\epsilon, \eta) \\
&= \dot{\rho}_\tau[Y - m(\eta_0^\intercal X, \theta_0)]\dot{m}_\theta(\eta_0^\intercal X, \theta_0) - E\{\dot{\rho}_\tau[Y - m(\eta_0^\intercal X, \theta_0)]\dot{m}_\theta(\eta_0^\intercal X, \theta_0)\}.
\end{aligned}$$

By the fact that $A(F, \eta_0) = 0$, the second term above is 0, leading to

$$\partial A_3(F_\epsilon, \eta_0)/\partial\epsilon = \dot{\rho}_\tau[Y - m(\eta_0^\intercal X, \theta_0)]\dot{m}_\theta(\eta_0^\intercal X, \theta_0). \tag{3.3.6}$$

Substituting (3.3.4), (3.3.5), and (3.3.6) into (3.3.3), we obtain

$$\begin{aligned} - E\left\{f_{Y|X}[m(\eta_0^\intercal X, \theta_0)|X]\dot{m}_\theta(\eta_0^\intercal X, \theta_0)\dot{m}_\theta^\intercal(\eta_0^\intercal X, \theta_0)\right\}\Phi^\star \\ + \dot{\rho}_\tau[Y - m(\eta_0^\intercal X, \theta_0)]\dot{m}_\theta(\eta_0^\intercal X, \theta_0) = 0. \end{aligned}$$

This yields the desired form for $\Phi^\star$.

Next, we note that $\eta \mapsto \Phi(F_0, \eta)$ is defined by the equation

$$\int \dot{\rho}_\tau\{Y - m[\eta^\intercal X, \Phi(F_0, \eta)]\}\dot{m}_\theta[\eta^\intercal X, \Phi(F_0, \eta)]dF_0 = 0.$$

Denote the left hand side by $B(\eta)$, we have

$$\frac{\partial}{\partial \mathrm{vec}(\eta)^\intercal}B(\eta_0) = \frac{\partial}{\partial \mathrm{vec}(\eta)^\intercal}B_1(\eta_0) + \frac{\partial}{\partial \mathrm{vec}(\eta)^\intercal}B_2(\eta_0),$$

where

$$\begin{aligned} B_1(\eta) &= \int \dot{\rho}_\tau\{Y - m[\eta^\intercal X, \Phi(F_0, \eta)]\}\dot{m}[\eta_0^\intercal X, \Phi(F_0, \eta_0)]dF_0 \\ B_2(\eta) &= \int \dot{\rho}_\tau\{Y - m[\eta_0^\intercal X, \Phi(F_0, \eta_0)]\}\dot{m}[\eta^\intercal X, \Phi(F_0, \eta)]dF_0. \end{aligned}$$

Since $E[\dot{\rho}_\tau(Y - m(\eta_0^\intercal X, \theta_0))\big|X] = 0$, we have

$$\frac{\partial}{\partial \mathrm{vec}(\eta)^\intercal}B_2(\eta_0) = \int \dot{\rho}_\tau(Y - m(\eta_0^\intercal X, \theta_0))\frac{\partial \dot{m}(\eta_0^\intercal X, \Phi(F_0, \eta_0))}{\partial \mathrm{vec}(\eta)^\intercal}dF_0 = 0.$$

The term $\partial B_1(\eta)/\partial \mathrm{vec}(\eta)^\intercal$ can be written as

$$\begin{aligned} &\frac{\partial}{\partial \mathrm{vec}(\eta)^\intercal}\int_{\Omega_X}\int_{-\infty}^{m[\eta^\intercal X, \Phi(F_0, \eta)]} f_{Y|X}(y|x)dy\, \dot{m}[\eta_0^\intercal X, \Phi(F_0, \eta_0)]f_X(x)dx \\ &= \int_{\Omega_X} \dot{m}_\theta(\eta_0^\intercal X, \theta_0)f_{Y|X}[m(\eta_0^\intercal X, \theta_0)|x] \end{aligned}$$

$$\left[\dot{m}_u^\intercal(\eta_0^\intercal X, \theta_0)\frac{\partial \eta^\intercal X}{\partial \mathrm{vec}(\eta)^\intercal} + \dot{m}_\theta(\eta_0^\intercal X, \theta_0)D\right] f_X(x)dx.$$

Recall that $\partial(\eta^\intercal X)/\partial \mathrm{vec}(\eta)^\intercal = I_q \otimes X^\intercal$. So the above term can be written as

$$E\left\{\dot{m}_\theta(\eta_0^\intercal X, \theta_0) f_{Y|X}(m(\eta_0^\intercal X, \theta_0)|x)\left[\dot{m}_u^\intercal(\eta_0^\intercal X, \theta_0)(I_q \otimes X^\intercal) + \dot{m}_\theta^\intercal(\eta_0^\intercal X, \theta_0)D\right]\right\}.$$

Equating it to 0 and solving for D lead to the desired form for D. □

3.3.3 Generalized method of moments

Generalized method of moments (Hansen, 1982, GMM) is another popular parametric method in both econometrics and statistics. For instance, it is used to construct optimal estimation and inference procedures based on generalized estimating equations (Qu et al., 2000), or to combine efficient and robust estimators (Park and Lindsay, 1999). We next derive the influence function $\Phi^\star(X, Y, \eta_0)$ and D for this approach.

In GMM, we have more estimating equations than the number of parameters. That is, we estimate the p-dimensional parameter vector θ by $k > p$ estimating equations $E[g(\theta, X, Y)] = 0$, where

$$g(\theta, \eta^\intercal X, Y) = [g_1(\theta, \eta^\intercal X, Y), \ldots, g_k(\theta, \eta^\intercal X, Y)]^\intercal,$$

and again we assume $E_{\theta,\eta}[g(\theta, \eta^\intercal X, Y)] = 0$ and $\mathrm{var}_{\theta,\eta}[g(\theta, \eta^\intercal X, Y)] < \infty$. For a given η, $\tilde{\theta}(\eta) = \Phi(F_n, \eta)$ in the optimal version of GMM is defined as the minimizer of the function

$$L(F_n, \theta, \eta) = E_n g(\theta, \eta^\intercal X, Y)^\intercal [E_n g(\theta, \eta^\intercal X, Y) g^\intercal(\theta, \eta^\intercal X, Y)]^{-1} E_n g(\theta, \eta^\intercal X, Y).$$

Thus, the functional $\Phi(F, \eta)$ is the minimizer of

$$L(F, \theta, \eta) = V(F, \theta, \eta)^\intercal W(F, \theta, \eta) V(F, \theta, \eta),$$

where $V(F,\theta,\eta)=\int g(\theta,\eta^{\tau}X,Y)dF$, and

$$W(F,\theta,\eta)=\left(\int g(\theta,\eta^{\tau}X,Y)g^{\tau}(\theta,\eta^{\tau}X,Y)dF\right)^{-1}.$$

Proposition 3. *For the generalized method of moments, we have*

$$\begin{aligned}\Phi^{\star}(X,Y,\eta)=&-\left\{E\left(\frac{\partial g^{\tau}}{\partial\theta}\right)[E(gg^{\tau})]^{-1}E\left(\frac{\partial g}{\partial\theta^{\tau}}\right)\right\}^{-1}E\left(\frac{\partial g^{\tau}}{\partial\theta}\right)[E(gg^{\tau})]^{-1}g,\\ D=&-\left\{E\left(\frac{\partial g^{\tau}}{\partial\theta}\right)[E(gg^{\tau})]^{-1}E\left(\frac{\partial g}{\partial\theta^{\tau}}\right)\right\}^{-1}E\left(\frac{\partial g^{\tau}}{\partial\theta}\right)[E(gg^{\tau})^{-1}]\\ &E\left(\frac{\partial g}{\partial u^{\tau}}\right)(I_q\otimes X^{\tau}),\end{aligned}$$

where $g=g(\theta_0,\eta_0^{\tau}X,Y)$.

Proof. Let $H(F,\theta,\eta_0)=\partial L(F,\theta,\eta_0)/\partial\theta$. Then $\Phi(F,\eta_0)$ satisfies

$$H(F,\Phi(F,\eta_0),\eta_0)=0.$$

Hence the influence function $\Phi^{\star}(X,Y,\eta_0)$ can be solved from the equation

$$\frac{\partial}{\partial\epsilon}H(F_{\epsilon},\Phi(F_{\epsilon},\eta_0),\eta_0)=0,$$

which, by the chain rule, yields

$$\Phi^{\star}=-\left[\frac{\partial}{\partial\theta^{\tau}}H(F_0,\theta_0,\eta_0)\right]^{-1}\frac{\partial}{\partial\epsilon}H\left[F_0,\Phi(F_{\epsilon},\eta_0),\eta_0\right]\Big|_{\epsilon=0}.$$

We now express the above derivatives in terms of $V(F,\theta,\eta)$ and $W(F,\theta,\eta)$. By definition,

$$\begin{aligned}\frac{\partial L(F,\theta,\eta_0)}{\partial\theta}=&\frac{\partial V^{\tau}(F,\theta,\eta_0)}{\partial\theta}W(F,\theta,\eta_0)V(F,\theta,\eta_0)\\ &+\ V^{\tau}(F,\theta,\eta_0)\frac{\partial W(F,\theta,\eta_0)}{\partial\theta}V(F,\theta,\eta_0)\\ &+\ V^{\tau}(F,\theta,\eta_0)W(F,\theta,\eta_0)\frac{\partial V(F,\theta,\eta_0)}{\partial\theta},\end{aligned}\tag{3.3.7}$$

Differentiating (3.3.7) with respect to θ, and evaluating the derivative at θ_0, we

obtain

$$\begin{aligned}\frac{\partial H}{\partial \theta^\intercal} =& \frac{\partial^2 V}{\partial \theta \partial \theta^\intercal} WV + \frac{\partial V^\intercal}{\partial \theta}\frac{\partial W}{\partial \theta^\intercal}V + \frac{\partial V^\intercal}{\partial \theta}W\frac{\partial V}{\partial \theta^\intercal} \\ &+ \frac{\partial V^\intercal}{\partial \theta^\intercal}\frac{\partial W}{\partial \theta}V + V^\intercal\frac{\partial^2 W}{\partial \theta \partial \theta^\intercal}V + V^\intercal\frac{\partial W}{\partial \theta}\frac{\partial V}{\partial \theta^\intercal} \\ &+ \frac{\partial V^\intercal}{\partial \theta^\intercal}W\frac{\partial V}{\partial \theta} + V^\intercal\frac{\partial W}{\partial \theta^\intercal}\frac{\partial V}{\partial \theta} + V^\intercal W\frac{\partial^2 V}{\partial \theta \partial \theta^\intercal}.\end{aligned} \tag{3.3.8}$$

Since, by construction, $V(F_0, \theta_0, \eta_0) = \int g(\theta_0, \eta_0^\intercal X, Y)dF_0 = 0$, all the terms in (3.3.8) that involve V vanish, resulting in

$$\frac{\partial H}{\partial \theta^\intercal} = 2\frac{\partial V^\intercal}{\partial \theta^\intercal}W\frac{\partial V}{\partial \theta^\intercal}. \tag{3.3.9}$$

Similarly, we have

$$\frac{\partial H}{\partial \epsilon} = \frac{\partial V^\intercal}{\partial \theta}W\frac{\partial V}{\partial \epsilon} + \frac{\partial V^\intercal}{\partial \epsilon}W\frac{\partial V}{\partial \theta} = 2\frac{\partial V^\intercal}{\partial \epsilon}W\frac{\partial V}{\partial \theta}.$$

Due to the fact that

$$\frac{\partial V^\intercal}{\partial \theta} = \frac{\partial}{\partial \theta}\int g(\theta_0, \eta_0^\intercal X, Y)dF_0 = E\left[\frac{\partial}{\partial \theta}g^\intercal(\theta_0, \eta_0^\intercal X, Y)\right],$$

we obtain the desired form for $\Phi^\star(X, Y, \eta_0)$.

Next, we note that $H[F_0, \Phi(F_0, \eta), \eta] = 0$ for all η. Hence,

$$\frac{\partial}{\partial \theta^\intercal}H(F_0, \theta_0, \eta)\frac{\partial \Phi(F_0, \eta)}{\partial \mathrm{vec}(\eta)^\intercal} + \frac{\partial H(F_0, \theta_0, \eta)}{\partial \mathrm{vec}(\eta)^\intercal} = 0.$$

Solving this equation, we have

$$D = -\left(\frac{\partial H}{\partial \theta^\intercal}\right)^{-1}\frac{\partial H}{\partial \mathrm{vec}(\eta)^\intercal}. \tag{3.3.10}$$

The computation of $\partial H/\partial \mathrm{vec}(\eta)^\intercal$ is similar to that of $\partial H/\partial \theta^\intercal$: there are q terms

in total, and all the terms that involve V vanish, resulting in

$$\begin{aligned}\frac{\partial H(F_0,\theta_0,\eta_0)}{\partial \mathrm{vec}(\eta)^\intercal} &= \frac{\partial V^\intercal}{\partial\theta} W \frac{\partial V}{\partial \mathrm{vec}(\eta)^\intercal} + \frac{\partial V^\intercal}{\partial \mathrm{vec}(\eta)^\intercal} W \frac{\partial V}{\partial\theta} \\ &= 2\frac{\partial V^\intercal}{\partial\theta} W \frac{\partial V}{\partial \mathrm{vec}(\eta)^\intercal}.\end{aligned} \tag{3.3.11}$$

Furthermore,

$$\begin{aligned}\frac{\partial V}{\mathrm{vec}(\eta)^\intercal} &= E\left[\frac{\partial}{\partial u^\intercal} g(\theta_0, \eta_0^\intercal X, Y)\right] \frac{\partial \mathrm{vec}(\eta^\intercal X)}{\partial \mathrm{vec}(\eta)^\intercal} \\ &= E\left[\frac{\partial}{\partial u^\intercal} g(\theta_0, \eta_0^\intercal X, Y)\right] (I_q \otimes X^\intercal).\end{aligned} \tag{3.3.12}$$

Substituting (3.3.9), (3.3.11), (3.3.12) into (3.3.10), we obtain the desired form of D. □

3.4 Influence functions for reduction functionals

In this section, we derive the influence function $\Lambda^\star(X,Y)$ for some popular SDR methods, including SIR, SAVE, DR, and two forms of PHD. Although some forms of asymptotic expansions exist in the SDR literature (Li, 1991a, 1992; Li and Wang, 2007; Shao et al., 2007; Li, 2018b), they have all been developed for sequential tests, and none was in the form suitable for post reduction inference. Also, the development here can be extended to other regression-based SDR methods, e.g., the minimal discrepancy method (Cook and Ni, 2005), in a similar fashion.

Many SDR methods begin with slicing the range of the response to a fixed number of non-overlapping intervals; let $\{J_k : k = 1, \ldots, H\}$ be a set of intervals that partition Ω_Y. Let $D_k = I(Y \in J_k)$, $p_k = E(D_k)$, $\mu_k = E(X|Y \in J_k)$, and $\Sigma_k = \mathrm{var}(X|Y \in J_k)$. Let $\mu = E(X)$, $\nu = E(Y)$. The specific form of Λ for the above SDR methods are as follows.

(1) For SIR (Li, 1991a), $\Lambda_{\mathrm{SIR}}(F) = \sum_{k=1}^H p_k(\mu_k - \mu)(\mu_k - \mu)^\intercal$.

(2) For SAVE (Cook and Weisberg, 1991),

$$\Lambda_{\text{SAVE}}(F) = \sum_{k=1}^{H} p_k(\Sigma - \Sigma_k)\Sigma^{-1}(\Sigma - \Sigma_k)^{\mathsf{T}}.$$

(3) For DR (Li and Wang, 2007), $\Lambda_{\text{DR}}(F) = 2\Lambda_{\text{DR},1}(F) + 2\Lambda_{\text{DR},2}(F) + 2\Lambda_{\text{DR},\,3}(F)$, where

$$\begin{aligned}
&\Lambda_{\text{DR},1}(F)\\
&= E\{E[(X-\mu)(X-\mu)^{\mathsf{T}} - \Sigma|\tilde{Y}]\Sigma^{-1}E[(X-\mu)(X-\mu)^{\mathsf{T}} - \Sigma|\tilde{Y}]\},\\
&\Lambda_{\text{DR},2}(F)\\
&= E[E(X-\mu|\tilde{Y})E((X-\mu)^{\mathsf{T}}|\tilde{Y})]\Sigma^{-1}E[E(X-\mu|\tilde{Y})E((X-\mu)^{\mathsf{T}}|\tilde{Y})],\\
&\Lambda_{\text{DR},3}(F)\\
&= E[E((X-\mu)^{\mathsf{T}}|\tilde{Y})\Sigma^{-1}E(X-\mu|\tilde{Y})]E[E(X-\mu|\tilde{Y})E((X-\mu)^{\mathsf{T}}|\tilde{Y})],
\end{aligned}$$

with $\tilde{Y}$ being the discretized Y according to the partition $(J_1, \ldots, J_h)$; that is, $\tilde{Y} = \sum_{k=1}^{h} kI(Y \in J_k)$.

(4) For y-based PHD (Li, 1992), $\Lambda_{y\text{-PHD}}(F) = \Sigma_{YXX}\Sigma^{-1}\Sigma_{YXX}$, where

$$\Sigma_{YXX} = E((Y-\nu)(X-\mu)(X-\mu)^{\mathsf{T}}).$$

(5) For r-based PHD (Li, 1992; Cook, 1998a), $\Lambda_{r\text{-PHD}}(F) = \Sigma_{RXX}\Sigma^{-1}\Sigma_{RXX}$, where

$$\Sigma_{RXX} = E\{[(Y-\nu) - \beta^{\mathsf{T}}(X-\mu)](X-\mu)(X-\mu)^{\mathsf{T}}\},$$

and β is the regression coefficient vector $\Sigma^{-1}\Sigma_{XY}$, with $\Sigma_{XY} = \text{cov}(X, Y)$.

The next proposition gives the explicit forms of $\text{vec}(\Lambda^\star)$ for these SDR methods. The derivations are tedious but straightforward; the details are omitted here. We first write down some simple influence functions:

$$\begin{aligned}
&p_k^\star = D_k - p_k,\ \mu^\star = X - \mu,\ \text{and},\ \nu^\star = Y - E(Y)\\
&\mu_k^\star = -p_k^{-2}p_k^\star E(XD_k) + p_k^{-1}[XD_k - E(XD_k)]
\end{aligned}$$

$$\Sigma_k^\star = -p_k^{-2}p_k^\star E(XX^\tau D_k) - p_k^{-1}[XX^\tau D_k - E(XX^\tau D_k)] - \mu_k^\star\mu_k^\tau - \mu_k(\mu_k^\star)^\tau.$$

The influence function of β is

$$\beta^\star = (\Sigma^{-1})^\star\Sigma_{XY} + (\Sigma^{-1})\Sigma_{XY}^\star,$$

where $\Sigma_{XY}^\star = XY - E(XY) - (X-\mu)\nu - \mu(Y-\nu)$.

Proposition 4. *The influence functions for the above five reduction functionals are given by the following formulas.*

(1) For SIR,

$$\begin{aligned}\mathrm{vec}(\Lambda_{\mathrm{SIR}}^\star) = &\sum_{k=1}^H(\mu_k-\mu)\otimes(\mu_k-\mu)p_k^\star \\ &+ [p_k(\mu_k-\mu)\otimes I_p + I_p\otimes p_k(\mu_k-\mu)](\mu_k^\star-\mu^\star).\end{aligned}$$

(2) For SAVE,

$$\begin{aligned}\mathrm{vec}(\Lambda_{\mathrm{SAVE}}^\star) = &\sum_{k=1}^H[(\Sigma-\Sigma_k)\otimes(\Sigma-\Sigma_k)]\mathrm{vec}(\Sigma^{-1})p_k^\star \\ &+ p_k[(\Sigma-\Sigma_k)\otimes(\Sigma-\Sigma_k)]\mathrm{vec}[(\Sigma^{-1})^\star] \\ &+ \ p_k[(\Sigma-\Sigma_k)\Sigma^{-1}\otimes I_p + I_p\otimes(\Sigma-\Sigma_k)\Sigma^{-1}]\mathrm{vec}(\Sigma^\star-\Sigma_k^\star),\end{aligned}$$

where $\Sigma^\star$ and $(\Sigma^{-1})^\star$ are as given in Lemma 1.

(3) For DR,

$$\mathrm{vec}(\Lambda_{\mathrm{DR}}^\star) = 2\mathrm{vec}(\Lambda_{\mathrm{DR},1}^\star) + 2\mathrm{vec}(\Lambda_{\mathrm{DR},2}^\star) + 2\mathrm{vec}(\Lambda_{\mathrm{DR},3}^\star),$$

where

$$\begin{aligned}\mathrm{vec}(\Lambda_{\mathrm{DR},1}^\star) = &\sum_{k=1}^H(A_k\otimes A_k)\mathrm{vec}(\Sigma^{-1})p_k^\star + \ p_k(A_k\otimes A_k)\mathrm{vec}[(\Sigma^{-1})^\star] \\ &+ p_k(A_k\Sigma^{-1}\otimes I_p + I_p\otimes A_k\Sigma^{-1})\mathrm{vec}(A_k^\star), \\ \mathrm{vec}(\Lambda_{\mathrm{DR},2}^\star) = &(B\Sigma^{-1}\otimes I_p + I_p\otimes B\Sigma^{-1})\mathrm{vec}(B^\star) + (B\otimes B)\mathrm{vec}((\Sigma^{-1})^\star),\end{aligned}$$

$$\mathrm{vec}(\Lambda^{\star}_{\mathrm{DR},3}) = C^{\star}\mathrm{vec}(B) + C\mathrm{vec}(B^{\star}),$$

in which $A_k = E[(X-\mu)(X-\mu)^{\mathsf{T}} - \Sigma \mid Y \in J_k]$, $B = \sum_{k=1}^{H} p_k(\mu_k - \mu)$, *and* $C = \sum_{k=1}^{H} p_k(\mu_k - \mu)^{\mathsf{T}}\Sigma^{-1}(\mu_k - \mu)$, *with the influence functions*

$$\begin{aligned}
A_k^{\star} =& -p_k^{-2}p_k^{\star}E(XX^{\mathsf{T}}D_k) + p_k^{-1}[XX^{\mathsf{T}}D_k - E(XX^{\mathsf{T}}D_k)] \\
& - \mu_k^{\star}\mu^{\mathsf{T}} - \mu_k\mu^{*\mathsf{T}} + \mu^{\star}\mu^{\mathsf{T}} + \mu\mu^{*\mathsf{T}} - \Sigma^{\star}, \\
B^{\star} =& \sum_{k=1}^{H} p_k^{\star}(\mu_k - \mu)(\mu_k - \mu)^{\mathsf{T}} + p_k(\mu_k^{\star} - \mu^{\star})(\mu_k - \mu)^{\mathsf{T}} \\
& + p_k(\mu_k - \mu)(\mu_k^{\star} - \mu^{\star})^{\mathsf{T}}, \\
C^{\star} =& \sum_{k=1}^{H} p_k^{\star}(\mu_k - \mu)^{\mathsf{T}}\Sigma^{-1}(\mu_k - \mu) + p_k(\mu_k^{\star} - \mu^{\star})^{\mathsf{T}}\Sigma^{-1}(\mu_k - \mu) \\
& + \ p_k(\mu_k - \mu)^{\mathsf{T}}\Sigma^{-1*}(\mu_k - \mu) + p_k(\mu_k - \mu)^{\mathsf{T}}\Sigma^{-1}(\mu_k^{\star} - \mu^{\star}).
\end{aligned}$$

(4) For y-based PHD,

$$\begin{aligned}
\mathrm{vec}(\Lambda^{\star}_{\mathrm{y-PHD}}) =& (\Sigma_{YXX}\Sigma^{-1} \otimes I_p + I_P \otimes \Sigma_{YXX}\Sigma^{-1})\mathrm{vec}(\Sigma^{\star}_{YXX}) \\
& + (\Sigma_{YXX} \otimes \Sigma_{YXX})\mathrm{vec}(\Sigma^{-1*}),
\end{aligned}$$

where

$$\begin{aligned}
\Sigma_{YXX} =& E[(Y-\nu)(X-\mu)(X-\mu)^{\mathsf{T}}] \\
\Sigma^{\star}_{YXX} =& YXX^{\mathsf{T}} - E(YXX^{\mathsf{T}}) - \nu^{\star}E(XX^{\mathsf{T}}) - \nu[XX^{\mathsf{T}} - E(XX^{\mathsf{T}})] \\
& - \mu^{\star}E(YX^{\mathsf{T}}) - \mu[YX^{\mathsf{T}} - E(YX^{\mathsf{T}})] - [YX - E(YX)]\mu^{\mathsf{T}} \\
& - E(YX)\mu^{*\mathsf{T}} + \nu^{\star}\mu\mu^{\mathsf{T}} + \nu\mu^{\star}\mu^{\mathsf{T}} + \nu\mu\mu^{*\mathsf{T}}.
\end{aligned}$$

(5) For r-based PHD,

$$\begin{aligned}
\mathrm{vec}(\Lambda^{\star}_{\mathrm{r-PHD}}) =& (\Sigma_{RXX}\Sigma^{-1} \otimes I_p + I_p \otimes \Sigma_{RXX}\Sigma^{-1})\mathrm{vec}(\Sigma^{\star}_{RXX}) \\
& + (\Sigma_{RXX} \otimes \Sigma_{RXX})\mathrm{vec}(\Sigma^{-1*}),
\end{aligned}$$

where the matrix Σ_{RXX} is defined as $\Sigma_{YXX} - R$, and

$$\begin{aligned} R = & E(XX^\intercal \beta X^\intercal) - E(XX^\intercal)\beta\mu^\intercal \\ & - \mu\beta^\intercal E(XX^\intercal) - E(X\mu^\intercal \beta X^\intercal) + 2E(X\mu^\intercal \beta \mu^\intercal). \end{aligned}$$

The influence function of Σ_{RXX} is

$$\begin{aligned} \mathrm{vec}(\Sigma^\star_{RXX}) &= \mathrm{vec}(\Sigma^\star_{YXX}) - \mathrm{vec}(R^\star), \\ \mathrm{vec}(R^\star) &= R^\star_1 - R^\star_2 - R^\star_3 - R^\star_4 + R^\star_5, \end{aligned}$$

where

$$\begin{aligned} R^\star_1 &= \{X \otimes (XX^\intercal) - E[X \otimes (XX^\intercal)]\}\,\beta + E(X \otimes XX^\intercal)\beta^\star \\ R^\star_2 &= \{I_p \otimes [XX^\intercal - E(XX^\intercal)]\}\,(\mu \otimes \beta) + [I_p \otimes E(XX^\intercal)]\,\mu^\star \otimes \beta \\ &\quad + [I_p \otimes E(XX^\intercal)]\,\mu \otimes \beta^\star \\ R^\star_3 &= \{[XX^\intercal - E(XX^\intercal)] \otimes I_p\}\,(\beta \otimes \mu) + \{[E(XX^\intercal)] \otimes I_p\}\,(\beta^\star \otimes \mu) \\ &\quad + \{[E(XX^\intercal)] \otimes I_p\}\,(\beta \otimes \mu^\star) \\ R^\star_4 &= [X \otimes X - E(X \otimes X)]\,\mu^\intercal\beta + E(X \otimes X)\mu^{*\intercal}\beta + E(X \otimes X)\mu^\intercal\beta^\star \\ R^\star_5 &= 2[(\mu\beta^\intercal \otimes I_p)\mathrm{vec}(\mu^\star\mu^\intercal) + (\mu\beta^\intercal \otimes I_p)\mathrm{vec}(\mu\mu^{*\intercal}) \\ &\quad + (I_p \otimes \mu\mu^\intercal)\mathrm{vec}(\beta^\star\mu^\intercal) + (I_p \otimes \mu\mu^\intercal)\mathrm{vec}(\beta\mu^{*\intercal})]. \end{aligned}$$

The five influence functions in Proposition 4 can be easily estimated by replacing, whenever applicable, the expectation $E(\cdot)$ with the sample average $E_n(\cdot)$. We can then substitute into the formulas for B and Γ in Corollary 1 to obtain the estimated asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$.

3.5 Post dimension reduction inference

In this section, we develop the formal statistical inference procedures for θ based on the asymptotic distribution of $\hat{\theta} = \Phi(F_n, \hat{\eta})$ derived in Sections 3.2 through 3.4. First, we consider the confidence interval for an arbitrary linear combination of θ. Let $c \in \mathbb{R}^s$ be a vector and let z_α be the $(1-\alpha)$-th percentile of the standard normal distribution. Because $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{D}} N(0, \Gamma)$, the interval

$(c^\intercal\hat{\theta} - z_{\alpha/2}\sqrt{c^\intercal\Gamma c},\ c^\intercal\hat{\theta} + z_{\alpha/2}\sqrt{c^\intercal\Gamma c})$ covers the true parameter θ_0 with probability tending to $1-\alpha$. Therefore, by Slutsky's theorem, the asymptotic $(1-\alpha)$-level confidence interval for θ is

$$\left(c^\intercal\hat{\theta} - z_{\frac{\alpha}{2}}\sqrt{c^\intercal\widehat{\Gamma}c},\ c^\intercal\hat{\theta} + z_{\frac{\alpha}{2}}\sqrt{c^\intercal\widehat{\Gamma}c}\right),$$

where $\widehat{\Gamma} = \Gamma(\hat{\eta}, \hat{\theta})$ is an estimate of Γ as defined in Corollary 1.

Next, we consider testing the null hypothesis

$$H_0 : h(\theta) = h(\theta_0),$$

where $h : \mathbb{R}^s \to \mathbb{R}^k$ is a differentiable function. We use the function h to accommodate the situation where only part of the parameter θ, e.g., the first component of θ, is of interest. For power assessment, we consider the local alternative hypothesis

$$H_{1,n}(\lambda) : h(\theta) = h\left(\theta_0 + \frac{\lambda}{\sqrt{n}}\right),$$

where λ is a fixed vector in $\mathbb{R}^s$. Let $H(\theta) = \partial h^\intercal(\theta)/\partial\theta \in \mathbb{R}^{s\times k}$ be the gradient matrix of h at θ, $\theta_n = \theta_0 + \lambda/\sqrt{n}$, $\widehat{H} = H(\hat{\theta})$, and $H = H(\theta_0)$. We propose the following Wald-type test statistic

$$T = \sqrt{n}\left[h(\hat{\theta}) - h(\theta_0)\right]\left(\widehat{H}^\intercal\widehat{\Gamma}\widehat{H}\right)^{-1}\sqrt{n}\left[h(\hat{\theta}) - h(\theta_0)\right].$$

The next theorem gives the asymptotic distributions of T under the null and the local alternative distribution. In the following, convergence in distribution under the null hypothesis is written as $\xrightarrow[\theta_0]{\mathcal{D}}$, while convergence in distribution under the local alternative hypothesis is written as $\xrightarrow[\theta_n]{\mathcal{D}}$.

Theorem 3. *Suppose the conditions in Theorem 1 are satisfied and the matrices* Γ *and* H *are nonsingular, then*

$$T \xrightarrow[\theta_0]{\mathcal{D}} \chi^2_k. \tag{3.5.1}$$

Suppose, moreover, that $\hat{\theta}$ is a regular estimator, then

$$T \xrightarrow[\theta_n]{\mathcal{D}} \chi^2_k \left(\lambda^\intercal H \Gamma H^\intercal \lambda\right). \tag{3.5.2}$$

Proof. By Corollary 1 and the delta method, we have

$$\sqrt{n}\left[h(\hat{\theta}) - h(\theta_0)\right] \xrightarrow[\theta_0]{\mathcal{D}} N(0, H^\intercal \Gamma H),$$

which implies (3.5.1).

Since $\hat{\theta}$ is a regular estimator and h is differentiable, the asymptotic distribution of $\sqrt{n}[h(\hat{\theta}) - h(\theta_n)]$ under $H_{1,n}(\lambda)$ is the same as the asymptotic distribution of $\sqrt{n}[h(\hat{\theta}) - h(\theta_0)]$ under H_0. Next we decompose $\sqrt{n}[h(\hat{\theta}) - h(\theta_0)]$ as

$$\begin{aligned}\sqrt{n}\left[h(\hat{\theta}) - h(\theta_0)\right] &= \sqrt{n}\left[h(\hat{\theta}) - h(\theta_n)\right] + \sqrt{n}\left[h(\theta_n) - h(\theta_0)\right] \\ &= \sqrt{n}\left[h(\hat{\theta}) - h(\theta_n)\right] + H^\intercal \lambda + o(n^{-1/2}).\end{aligned}$$

By Slutsky's theorem,

$$\sqrt{n}\left[h(\hat{\theta}) - h(\theta_0)\right] \xrightarrow[\theta_n]{\mathcal{D}} N(H^\intercal \lambda, H^\intercal \Gamma H),$$

which implies

$$\sqrt{n}\left[H^\intercal \Gamma H\right]^{-\frac{1}{2}}\left[h(\hat{\theta}) - h(\theta_0)\right] \xrightarrow[\theta_n]{\mathcal{D}} N\left(\left(H^\intercal \Gamma H\right)^{-\frac{1}{2}} H^\intercal \lambda, I_k\right).$$

Together we have

$$\sqrt{n}\left[h(\hat{\theta}) - h(\theta_0)\right]\left[H^\intercal \Gamma H\right]^{-1}\left[h(\hat{\theta}) - h(\theta_0)\right] \xrightarrow[\theta_n]{\mathcal{D}} \chi^2_k\left[\lambda^\intercal H (H^\intercal \Gamma H)^{-1} H^\intercal \lambda\right].$$

Applying Slutsky's theorem again, we obtain (3.5.2). □

We briefly comment that the requirement $\hat{\theta}$ is a regular estimator is rather mild, and is satisfied by most estimators. See Bickel et al. (1993) and Van der Vaart (1998).

3.6 Conclusions

Despite the extensive development of sufficient dimension reduction in the past three decades, the critical step of post dimension reduction inference has never been taken – at least not in a systematic and rigorous manner. SDR is not complete without a proper post reduction inference procedure that takes the estimation error induced in the dimension reduction step into the subsequent model estimation step. We fill this gap by developing a general post dimension reduction inference framework that is adaptive to a multitude of dimension reduction and model estimation methods. We derive the inference procedures for confidence interval and hypothesis testing based on a combination of commonly used SDR and model building methods.

The framework laid out in this project also opens the door for developing objective inference procedures for a much broader class of dimension reduction problems than considered here. Potential extensions include unsupervised dimension reduction methods such as principal components analysis and independent components analysis (Hyvärinen et al., 2004), sparse sufficient dimension reduction methods (Li, 2007; Bondell and Li, 2009; Chen et al., 2010; Wang and Yin, 2008), and nonparametric sufficient dimension reduction methods (Xia et al., 2002; Xia, 2007). A particularly promising direction of extension is to the semiparametrically efficient SDR methods developed in Ma and Zhu (2012, 2013, 2014) and Luo et al. (2014). For these methods, the influence function can be readily developed from the efficient score, and it is plausible that semiparametric efficiency for sufficient dimension reduction can be inherited, to some degree at least, by the post dimension reduction inference procedure.

Beyond the asymptotic normality-based procedures considered in this project, it is also useful to develop nonparametric inference procedures for post dimension reduction inference. For example, it is possible to employ the empirical likelihood approach (Owen, 1988, 1990) to conduct post dimension inference. In this direction, Li et al. (2010) proposed an empirical likelihood inference procedure for the single-index model, and the ideas and techniques there might be adaptable to the current setting. The full potential and scope of the general framework of post dimension reduction inference will be explored in future research.

Chapter 4

Sufficient Graphical Models

4.1 Introduction

In this project, instead of relying on additivity to avoid the curse of dimensionality, we apply the recently developed nonparametric sufficient dimension reduction (Lee et al., 2013; Li, 2018c) to achieve this goal. The estimation proceeds in two steps: first, we use nonlinear sufficient dimension reduction to reduce the random vector $X^{-(i,j)}$ in (7.1.1) to a low-dimensional random vector, say U^{ij}; second, we use the kernel method to construct a nonparametric graphical model based on $X^{(i,j)}$ and the dimension-reduced random vectors U^{ij}. The main differences between this approach and Li et al. (2014) are, first, we are able to retain conditional independence as the criterion for constructing the graphical model, which is a widely accepted criterion with a more direct interpretation, and second, we are no longer restricted by the additive structure in the graphical model. Furthermore, an attractive feature of our method is due to the use of the "kernel trick", which means that the complexity of the proposed methods depend on the sample size rather than the size of the networks. This makes its algorithm computationally economic when handling large networks.

4.2 Sufficient graphical model

To set the stage, we first give an outline of nonlinear sufficient dimension reduction. In its classical form, sufficient dimension reduction (SDR) seeks to reduce the

dimension of a random vector $X \in \mathbb{R}^p$ by projecting it on to a subspace while preserving the information about a response random vector $Y \in \mathbb{R}^q$ contained in X. That is, we seek the smallest subspace $\mathcal{S}$ of $\mathbb{R}^p$ such that $Y \perp\!\!\!\perp X | P_{\mathcal{S}} X$, where $P_{\mathcal{S}}$ is the projection onto the subspace $\mathcal{S}$. This subspace is called the central subspace, written as $\mathcal{S}_{Y|X}$. See, for example, Li (1991b), Cook (1994), and Li (2018c). Li et al. (2011b) and Lee et al. (2013) extended this framework to the nonlinear setting by considering the more general problem: $Y \perp\!\!\!\perp X | \mathcal{G}$, where $\mathcal{G}$ a sub-σ field of the σ-field generated by X. The class of functions in a Hilbert space that are measurable with respect to $\mathcal{G}$ is called the central class, written as $\mathfrak{S}_{Y|X}$. Li et al. (2011b) introduced the Principal Support Vector Machine, and Lee et al. (2013) generalized the Sliced Inverse Regression (Li, 1991b) and the Sliced Average Variance Estimate (Cook and Weisberg, 1991) to estimate the central class. Precursors of this theory include Bach and Jordan (2002), Wu (2008), and Wang (2008).

Returning to the statistical graphical model. Let $(\Omega, \mathcal{F}, P)$ be a probability space, and $(\Omega_X, \mathcal{F}_X)$ a measurable space, where Ω_X is a subset of $\mathbb{R}^p$ and $\mathcal{F}_X$ is the σ-field of Borel sets in Ω_X. Let $X : \Omega \to \Omega_X$ be a random vector, and $P_X = P \circ X^{-1}$ its distribution. The ith component of X is written as X^i and its range written as Ω_{X^i}. We assume $\Omega_X = \Omega_{X^1} \times \cdots \times \Omega_{X^p}$. Let $X^{(i,j)}$ be the vector (X^i, X^j) and $X^{-(i,j)}$ the vector formed by removing the ith and jth components of X. Let $\sigma(X^{-(i,j)})$ be the σ-field generated by $X^{-(i,j)}$.

We assume, for each $(i, j) \in \Gamma \times \Gamma$, there is a proper sub σ-field $\mathcal{G}^{-(i,j)}$ of $\sigma(X^{-(i,j)})$ such that

$$X^{(i,j)} \perp\!\!\!\perp X^{-(i,j)} | \mathcal{G}^{-(i,j)}.$$

Without loss of generality, we assume $\mathcal{G}^{-(i,j)}$ is the smallest sub σ-field of $\sigma(X^{-(i,j)})$ that satisfies the above relation; that is, $\mathcal{G}^{-(i,j)}$ is the central σ-field for $X^{(i,j)}$ versus $X^{-(i,j)}$.

The next proposition shows that, under the above dimension reduction condition, the conditional independence $X^i \perp\!\!\!\perp X^j | X^{-(i,j)}$ is equivalent to $X^i \perp\!\!\!\perp X^j | \mathcal{G}^{-(i,j)}$. The proof of the next proposition relies on the properties of conditional independence developed in Dawid (1979), with a detailed proof given in Li (2018c).

Theorem 4. *If $X^{(i,j)} \perp\!\!\!\perp X^{-(i,j)} | \mathcal{G}^{-(i,j)}$, then the following statements are equivalent:*

1. $X^i \perp\!\!\!\perp X^j | X^{-(i,j)}$;
2. $X^i \perp\!\!\!\perp X^j | \mathcal{G}^{-(i,j)}$.

This equivalence motivates us to use $X^i \perp\!\!\!\perp X^j | \mathcal{G}^{-(i,j)}$ as the criterion to construct the graph $\mathcal{G}$ after performing nonlinear sufficient dimension reduction of $X^{(i,j)}$ versus $X^{-(i,j)}$ for each $(i,j) \in \Gamma \times \Gamma$, $i > j$.

Definition 1. *Suppose, for each $(i,j) \in \Gamma \times \Gamma$, $i > j$, there is a proper smallest sub σ-field of $\sigma(X^{-(i,j)})$ such that $X^{(i,j)} \perp\!\!\!\perp X^{-(i,j)} | \mathcal{G}^{-(i,j)}$. Then the statistical graphical model defined by the following equivalence*

$$(i,j) \notin \mathcal{E} \Leftrightarrow X^i \perp\!\!\!\perp X^j | \mathcal{G}^{-(i,j)}$$

is called the Sufficient Graphical Model, abbreviated by SGM.

Nonlinear sufficient dimension reduction is a particularly natural framework for reducing dimension in a statistical graphical model: it would have been conceptually difficult to use linear sufficient dimension reduction. For example, if we assume

$$X^{(i,j)} \perp\!\!\!\perp X^{-(i,j)} | \beta_{ij}^{\mathsf{T}} X^{-(i,j)}, \tag{4.2.1}$$

and if this conditional independence is realized by the regression model

$$X^{(i,j)} = f_{ij}(\beta_{ij}^{\mathsf{T}} X^{-(i,j)}) + \epsilon_{ij},$$

where $X^{-(i,j)} \perp\!\!\!\perp \epsilon_{ij}$, then this may well rule out the possibility of any model of the form

$$X^{(k,\ell)} = f_{k\ell}(\beta_{k\ell}^{\mathsf{T}} X^{(k,\ell)}) + \epsilon_{k\ell},$$

where $\{k,\ell\} \subseteq \{1,\ldots,p\} \setminus \{i,j\}$. In other words, it may be impossible for (4.2.1) to hold for all $(i,j) \in \Gamma \times \Gamma$, $i > j$. On the other hand, nonlinear sufficient dimension reduction has no such difficulty: $X^i \perp\!\!\!\perp X^j | \mathcal{G}^{-(i,j)}$ imposes no restriction

on the functional form of the dependence of $X^{(i,j)}$ on $X^{-(i,j)}$, and it can hold for all $(i,j) \in \Gamma \times \Gamma$, $i > j$.

4.3 Estimation: population-level development

The estimation of the SMG involves two steps: the first step is to use nonlinear sufficient dimension reduction to estimate $\mathcal{G}^{-(i,j)}$; the second is to construct a graph $\mathscr{G}$ based on reduced data

$$\{(X^{(i,j)}, \mathcal{G}^{-(i,j)}) : (i,j) \in \Gamma \times \Gamma, i > j\}.$$

In this section we describe the two steps at the population level.

4.3.1 Preliminaries

We first briefly describe the mean element and covariance operator in a reproducing kernel Hilbert space (RKHS). A full development of the related theory is given in Li (2018c). For a linear operator A in a Hilbert space $\mathscr{H}$, let $\ker(A) = \{h \in \mathscr{H} : Ah = 0\}$ be the kernel of A, $\text{ran}(A) = \{Ah : h \in \mathscr{H}\}$ the range of A, and $\overline{\text{ran}}(A)$ the closure of $\text{ran}(A)$.

We first introduce the generic notion of the centered RKHS. Suppose $(\Omega, \mathcal{F}, P)$ is a probability space and U is a random element defined on $(\Omega, \mathcal{F}, P)$ taking values in $(\Omega_U, \mathcal{F}_U)$. This random element will be taken to be several different random vectors in the next few subsections. Let $\kappa : \Omega \times \Omega \to \mathbb{R}$ be a positive definite kernel function. The RKHS generated by the kernel κ is the completion of the linear span of the set of functions $\{\kappa(\cdot, u) : u \in \Omega_U\}$ with the inner product between members of the linear span determined by $\langle u_1, u_2 \rangle = \kappa(u_1, u_2)$. Let us denote this RKHS as $\mathscr{K}_U$. Under the assumption that $E\kappa(U,U) < \infty$, the mean element μ_U of U is a well defined element of $\mathscr{K}_U$ and the covariance operator Σ_{UU} of U is a well defined linear operator that maps from $\mathscr{K}_U$ to $\mathscr{K}_U$, and they satisfy:

1. $\langle f, \mu_U \rangle_{\mathscr{K}_U} = Ef(U)$ for each $f \in \mathscr{K}_U$;
2. $\langle f, \Sigma_{UU} g \rangle_{\mathscr{K}_U} = \text{cov}[f(U), g(U)]$ for all $f, g \in \mathscr{K}_U$.

To study statistical relations, we can, without loss of generality, reset $\mathscr{K}_U$ to be $\overline{\text{ran}}(\Sigma_{UU})$. This is because $\overline{\text{ran}}(\Sigma_{UU}) = \ker(\Sigma_{UU})^\perp$, any $f \in \ker(\Sigma_{UU})$ satisfies $\text{var}[f(U)] = 0$. Hence the $\ker(\Sigma_{UU})$ consists of functions of U that are almost surely constants. Such functions can be removed from $\mathscr{K}_U$ without affecting any statistical relation. Thus, for the rest of the project we will only be concerned with the subspace $\overline{\text{ran}}(\Sigma_{UU})$ of $\mathscr{K}_U$. We call this subspace the centered RKHS generated by κ. It can be shown that $\mathscr{H}_U$ is the subspace of $\mathscr{K}_U$ spanned by $\{\kappa_U(\cdot, u) - \mu_U : u \in \Omega_U\}$. Since $\ker(\Sigma_{UU}) = 0$ when Σ_{UU} is restricted on $\mathscr{H}_U$, it is an injective linear operator. We use Σ_{UU}^{-1} to denote the operator from $\text{ran}(\Sigma_{UU})$ to $\overline{\text{ran}}(\Sigma_{UU}) = \mathscr{H}_U$ that sends $\Sigma_{UU}h$ to h. This inverse, however, is not a bounded operator because the operator Σ_{UU}, if it is defined, is a Hilbert-Schmidt operator.

Next, let V be another random element defined on $(\Omega, \mathcal{F}, P)$ taking values in $(\Omega_V, \mathcal{F}_V)$, $\kappa_V : \Omega_V \times \Omega_V \to \mathbb{R}$ a positive definite kernel that satisfies $E\kappa_V(V, V) < \infty$, and $\mathscr{H}_V$ the centered RKHS generated by κ_V. The covariance operator Σ_{UV} is a mapping from $\mathscr{H}_V$ to $\mathscr{H}_U$ that satisfies

$$\langle f, \Sigma_{UV} g \rangle_{\mathscr{H}_U} = \text{cov}[f(U), g(V)]$$

for each $f \in \mathscr{H}_U$ and $g \in \mathscr{H}_V$.

The function μ_U and the linear operators such as Σ_{UU} and Σ_{UV} can be represented explicitly in terms of kernels, as follows:

$$\begin{aligned} \mu_U &= E\kappa_U(\cdot, U), \\ \Sigma_{UU} &= E\{[\kappa_U(\cdot, U) - E\kappa_U(\cdot, U)] \otimes [\kappa_U(\cdot, U) - E\kappa_U(\cdot, U)]\}, \\ \Sigma_{UU} &= E\{[\kappa_U(\cdot, U) - E\kappa_U(\cdot, U)] \otimes [\kappa_V(\cdot, V) - E\kappa_U(\cdot, V)]\}, \end{aligned} \tag{4.3.1}$$

where $\otimes$ is the tensor product, $\kappa_U(\cdot, U)$ is the function $\Omega_U \to \mathbb{R}$, $u \mapsto \kappa_U(u, U)$, and $E\kappa_U(\cdot, U)$ is the function $\Omega_U \to \mathbb{R}$, $\mapsto E[\kappa_U(u, U)]$.

4.3.2 Step 1: nonlinear dimension reduction

We use the Generalized Sliced Inverse Regression (Lee et al., 2013; Li, 2018c, GSIR) to perform the nonlinear dimension reduction. For each pair $(i, j) \in \Gamma \times \Gamma$, $i > j$, let $\Omega_{X^{-(i,j)}}$ be the range of $X^{-(i,j)}$; that is, the Cartesian product of $\Omega_{X^1}, \ldots, \Omega_{X^p}$

with Ω_{X^i} and Ω_{X^j} removed. Let

$$\kappa_X^{-(i,j)} : \Omega_{X^{-(i,j)}} \times \Omega_{X^{-(i,j)}} \to \mathbb{R}$$

be a positive definite kernel function. Let $\mathscr{H}_X^{-(i,j)}$ be the centered RKHS generated by $\kappa_X^{-(i,j)}$. Similarly, let $\Omega_{X^{(i,j)}}$ be the range of the mapping $X^{(i,j)} : \Omega \to \mathbb{R}^2$, let

$$\kappa_X^{(i,j)} : \Omega_{X^{(i,j)}} \times \Omega_{X^{(i,j)}} \to \mathbb{R}$$

be a positive kernel function, and let $\mathscr{H}_X^{(i,j)}$ be the centered RKHS generated by $\kappa_X^{(i,j)}$.

Assumption 7.

$$E[\kappa_X^{-(i,j)}(X^{-(i,j)}, X^{-(i,j)})] < \infty, \quad E[\kappa_X^{(i,j)}(X^{(i,j)}, X^{(i,j)})] < \infty.$$

Under this assumption, the following operators are well defined:

$$\Sigma_{X^{-(i,j)}X^{(i,j)}} : \mathscr{H}_X^{(i,j)} \to \mathscr{H}_X^{-(i,j)}$$
$$\Sigma_{X^{-(i,j)}X^{-(i,j)}} : \mathscr{H}_X^{-(i,j)} \to \mathscr{H}_X^{-(i,j)}.$$

Next, we introduce the regression operator from $\mathscr{H}_X^{(i,j)}$ to $\mathscr{H}_X^{-(i,j)}$. For this purpose we need to make the following assumption.

Assumption 8. $\operatorname{ran}(\Sigma_{X^{-(i,j)}X^{(i,j)}}) \subseteq \operatorname{ran}(\Sigma_{X^{-(i,j)}X^{-(i,j)}})$.

As argued in Li (2018c), this assumption can be interpreted as a type of collective smoothness in the relation between $X^{(i,j)}$ and $X^{-(i,j)}$: it is satisfied if and only if, for any $g \in \operatorname{ran}(\Sigma_{X^{-(i,j)}X^{(i,j)}})$,

$$\sum_{r=1}^{\infty} \frac{\operatorname{cov}^2[g(X^{(i,j)}), f_r(X^{-(i,j)})]}{\operatorname{var}^4[f_r(X^{-(i,j)})]} < \infty,$$

where $f_1, f_2, \ldots$ are the eigenfunctions of the operator $\Sigma_{X^{-(i,j)}X^{-(i,j)}}$. So the output of the covariance operator $\Sigma_{X^{-(i,j)}X^{(i,j)}}$ is concentrated on the low-frequency components of $\Sigma_{X^{-(i,j)}X^{-(i,j)}}$.

Under this assumption, the linear operator

$$\Sigma^{-1}_{X^{-(i,j)}X^{-(i,j)}}\Sigma_{X^{-(i,j)}X^{(i,j)}}$$

is defined, and we call it the regression operator from $\mathscr{H}_X^{(i,j)}$ to $\mathscr{H}_X^{-(i,j)}$. We denote this operator by $R_{X^{-(i,j)}X^{(i,j)}}$. The regression operator in this form was formally defined in Lee et al. (2016b), but earlier forms existed in Fukumizu et al. (2004); see also Li (2018a). About the regression operator, we make the following assumption.

Assumption 9. *$R_{X^{-(i,j)}X^{(i,j)}}$ is a finite-rank operator. Let its rank be denoted by d_{ij}.*

Again, this assumption imposes collective smoothness between functions of $X^{(i,j)}$ and functions of $X^{-(i,j)}$. The regression operator plays a crucial role in nonlinear sufficient dimension reduction. Let $P_{X^{-(i,j)}} = P\circ(X^{-(i,j)})^{-1}$ be the distribution of $X^{-(i,j)}$, and $L_2(P_{X^{-(i,j)}})$ the set of all square-integrable functions with respect to $P_{X^{-(i,j)}}$. As shown in Lee et al. (2013) (see also Li (2018c), Chapter 13), the range of the regression operator is equal to the central subspace; that is,

$$\mathrm{ran}(R_{X^{-(i,j)}X^{(i,j)}}) = \mathfrak{S}_{X^{(i,j)}|X^{-(i,j)}}. \tag{4.3.2}$$

under the following assumption.

Assumption 10.

1. *$\mathscr{H}_X^{-(i,j)}$ is dense in $L_2(P_{X^{-(i,j)}})$ modulo constants; that is, for any $f \in L_2(P_{X^{-(i,j)}})$ and any $\epsilon > 0$, there is a $g \in \mathscr{H}_X^{-(i,j)}$ such that* $\mathrm{var}[f(X^{-(i,j)}) - g(X^{-(i,j)})] < \epsilon$*;*
2. *$\mathcal{G}_{-(i,j)}$ is a sufficient and complete.*

A basis of the central class $\mathfrak{S}_{X^{(i,j)}|X^{-(i,j)}}$ can be found by solving the generalized eigenvalue problem: for $k = 1, \ldots, d_{ij}$,

$$\begin{aligned} &\text{maximize} && \langle f, \Sigma_{X^{-(i,j)}X^{(i,j)}} A \Sigma_{X^{(i,j)}X^{-(i,j)}} f\rangle_{-(i,j)} \\ &\text{subject to} && \begin{cases} \langle f_k, \Sigma_{X^{-(i,j)}X^{-(i,j)}} f_k\rangle_{-(i,j)} = 1 \\ \langle f_k, \Sigma_{X^{-(i,j)}X^{-(i,j)}} f_i\rangle_{-(i,j)}, \text{ for } i = 1, \ldots, k-1 \end{cases} \end{aligned} \tag{4.3.3}$$

where $A : \mathscr{H}_X^{(i,j)} \to \mathscr{H}_X^{(i,j)}$ is any nonsingular and self adjoint operator, and $\langle \cdot, \cdot \rangle_{-(i,j)}$ is the inner product in $\mathscr{H}_X^{-(i,j)}$. That is, if $f_1^{ij}, \ldots f_{d_{ij}}^{ij}$ are the first d_{ij} eigenfunctions of this eigenvalue problem, then they span the central class. This type of estimates of the central class is called GSIR. Convenient choices of A are the identity mapping I or the operator $\Sigma^{-1}_{X^{(i,j)}X^{(i,j)}}$. If we use the latter, then we need the following assumption.

Assumption 11. $\text{ran}(\Sigma_{X^{(i,j)}X^{-(i,j)}}) \subseteq \text{ran}(\Sigma_{X^{(i,j)}X^{(i,j)}})$.

In this project we use the version of GSIR with $A = \Sigma^{-1}_{X^{(i,j)}X^{(i,j)}}$. Let U^{ij} denote the random vector

$$(f_1^{ij}(X^{-(i,j)}), \ldots f_{d_{ij}}^{ij}(X^{-(i,j)})).$$

The set of random vectors $\{U^{ij} : (i,j) \in \Gamma \times \Gamma, i > j\}$ is the output for the nonlinear sufficient dimension reduction step.

4.3.3 Step 2: estimation of Sufficient Graphical Model

To estimate the edge set of the SGM we need to find a way to determine whether $X^i \perp\!\!\!\perp X^j | U^{ij}$ is true. We use a linear operator introduced by Fukumizu et al. (2008) to perform this task. We next give a brief description of that operator using our notation.

Let U, V, W be random vectors defined on $(\Omega, \mathcal{F}, P)$, taking values in measurable spaces $(\Omega_U, \mathcal{F}_U)$, $(\Omega_V, \mathcal{F}_V)$, and $(\Omega_W, \mathcal{F}_W)$, respectively. Let $P_U = P \circ U^{-1}$, $P_V = P \circ V^{-1}$, and $P_W = P \circ W^{-1}$ be the distributions of U, V, and W. Let $\Omega_{UW} = \Omega_U \times \Omega_W$, $\Omega_{VW} = \Omega_V \times \Omega_W$, $\mathcal{F}_{UW} = \mathcal{F}_U \times \mathcal{F}_V$, and $\mathcal{F}_{VW} = \mathcal{F}_V \times \mathcal{F}_W$. Let $P_{UW} = P \circ (U, W)^{-1}$ and $P_{VW} = P \circ (V, W)^{-1}$ be the distributions of (U, W) and (V, W). Let

$$\kappa_{UW} : \Omega_{UW} \times \Omega_{UW} \to \mathbb{R}, \quad \kappa_{VW} : \Omega_{VW} \times \Omega_{VW} \to \mathbb{R}, \quad \kappa_W : \Omega_W \times \Omega_W \to \mathbb{R}$$

be positive kernel functions. For example, for $(u_1, w_1), (u_2, w_2) \in \Omega_{UW} \times \Omega_{UW}$, κ_{UW} returns a real number denoted by $\kappa_{UW}[(u_1, w_1), (u_2, w_2)]$. Let $\mathscr{H}_{UW}$, $\mathscr{H}_{VW}$, and $\mathscr{H}_W$ be the centered RKHS's generated by the kernels κ_{UW}, κ_{VW}, and κ_W. Define the

covariance operators

$$\begin{aligned} &\Sigma_{(UW)(VW)} : \mathscr{H}_{VW} \to \mathscr{H}_{UW}, \quad \Sigma_{(UW)W} : \mathscr{H}_W \to \mathscr{H}_{UW}, \\ &\Sigma_{(VW)W} : \mathscr{H}_W \to \mathscr{H}_{VW}, \quad \Sigma_{WW} : \mathscr{H}_W \to \mathscr{H}_W \end{aligned} \tag{4.3.4}$$

as before. The following definition is due to Fukumizu et al. (2008). Since it plays a special role in this project, we give it a name – "conjoined conditional covariance operator" (CCCO) that figuratively depicts its form.

Definition 2. *Suppose*

1. *If S is W, or (U, W), or (V, W), then $E[\kappa_S(S, S)] < \infty$;*
2. $\text{ran}(\Sigma_{W(VW)}) \subseteq \text{ran}(\Sigma_{WW}), \quad \text{ran}(\Sigma_{W(UW)}) \subseteq \text{ran}(\Sigma_{WW})$.

Then the linear operator

$$\Sigma_{\ddot{U}\ddot{V}|W} = \Sigma_{(UW)(VW)} - \Sigma_{(UW)W}\Sigma_{WW}^{-1}\Sigma_{W(VW)}$$

is called the conjoined conditional covariance operator between U and V given W.

The word "conjoined" describes the peculiar way in which W appears in $\Sigma_{(UW)W}$ and $\Sigma_{W(VW)}$, which differs from an ordinary conditional covariance operator, where these operators are replaced by Σ_{UW} and Σ_{WV}. The following proposition is due to Fukumizu et al. (2008), a proof of a special case of which is given in Fukumizu et al. (2004).

Proposition 5. *Suppose*

1. *$\mathscr{H}_{UW} \otimes \mathscr{H}_{VW}$ is probability determining;*
2. *for each $f \in \mathscr{H}_{UW}$, the function $w \mapsto E[f(U, W)|W = w]$ is a member of $\mathscr{H}_W$;*
3. *for each $g \in \mathscr{H}_{VW}$, the function $w \mapsto E[g(V, W)|W = w]$ is a member of $\mathscr{H}_W$;*

Then $\Sigma_{\ddot{U}\ddot{V}|W} = 0$ if and only if $U \perp\!\!\!\perp V|W$.

We apply the above proposition to X^i, X^j, U^{ij} for each $(i,j) \in \Gamma \times \Gamma$, $i > j$. Let

$$\kappa_{XU}^{i,ij} : (\Omega_{X^i} \times \Omega_{U^{ij}}) \times (\Omega_{X^i} \times \Omega_{U^{ij}}) \to \mathbb{R}$$

be a positive definite kernel, and $\mathscr{H}_{XU}^{i,ij}$ the centered RKHS generated by $\kappa_{XU}^{i,ij}$. Similarly, let

$$\kappa_{U}^{ij} : \Omega_{U^{ij}} \times \Omega_{U^{ij}} \to \mathbb{R}$$

be a positive definite kernel, and $\mathscr{H}_{U}^{ij}$ the centered RKHS generated by κ_{U}^{ij}. We make the following assumption.

Assumption 12. *Conditions (1) and (2) of Definition 2 and conditions (1), (2), and (3) of Proposition 5 are satisfied with U, V, and W therein replaced by X^i, X^j, and U^{ij}, respectively, for each $(i,j) \in \Gamma \times \Gamma$ and $i > j$.*

Under this assumption, the operators

$$\Sigma_{(X^iU^{ij})(X^iU^{ij})}, \quad \Sigma_{(X^iU^{ij})U^{ij}}, \quad \Sigma_{U^{ij}U^{ij}}, \quad \Sigma_{U^{ij}(X^iU^{ij})}, \quad \Sigma_{\ddot{X}^i\ddot{X}^j|U^{ij}}$$

are well defined, and we have the following result.

Corollary 2. *Under Assumption 12,*

$$(i,j) \notin E \Leftrightarrow \Sigma_{\ddot{X}^i\ddot{X}^j|U^{ij}} = 0.$$

This corollary motivates us to estimate the conjoined conditional covariance operator $\Sigma_{\ddot{X}^i\ddot{X}^j|U^{ij}}$ and estimate the graph by thresholding the norm of the estimated operator.

4.4 Estimation: sample-level implementation

4.4.1 Coordinating mapping

To make the method proposed in the last section into executable procedures we need to represent various linear operators involved as matrices via coordinate map-

ping. A full description of coordinate mapping can be found in Sections 12.3 and 12.4 of Li (2018c). Here, we give an outline of some key notations and results important for our exposition. Let $\mathscr{H}_1$ and $\mathscr{H}_2$ be finite-dimensional Hilbert spaces with spanning sets $\mathscr{B}_1 = \{h_{11}, \ldots, h_{1m_1}\}$ and $\mathscr{B}_2 = \{h_{21}, \ldots h_{2m_2}\}$. Here, we allow the vectors in the spanning sets to be linearly dependent. Any function $f \in \mathscr{H}_1$ can be represented as a linear combination of vectors in $\mathscr{B}_1$; we call the $\mathbb{R}^{m_1}$-vector of linear coefficients the coordinate of f with respect to $\mathscr{B}_1$ and denote the coordinate by $[f]_{\mathscr{B}_1}$. If $A : \mathscr{H}_1 \to \mathscr{H}_2$ is a linear operator, then Af is a member of $\mathscr{H}_2$ and has a coordinate $[Af]_{\mathscr{B}_2}$ with respect to the spanning set $\mathscr{B}_2$ of $\mathscr{H}_2$. There is always a matrix $M \in \mathbb{R}^{m_2 \times m_1}$ such that $[Af]_{\mathscr{B}_2} = M[f]_{\mathscr{B}_1}$, and we call this matrix the coordinate of A with respect to $\mathscr{B}_1$-$\mathscr{B}_2$, and denote it by ${}_{\mathscr{B}_2}[A]_{\mathscr{B}_1}$.

The following proposition will be used in the subsequent discussions. It concerns a single finite-dimensional Hilbert space and its spanning set $\mathscr{B} = \{h_1, \ldots, h_m\}$. Let $A : \mathscr{H} \to \mathscr{H}$ be a self-adjoint operator.

Proposition 6. *Let $G_{\mathscr{B}} = \{\langle h_a, h_b \rangle_{\mathscr{H}}\}_{a,b=1}^n$ be the Gram matrix of the set $\mathscr{B}$, $I : \mathscr{H} \to \mathscr{H}$ the identity mapping, and $\epsilon > 0$ a constant. Then*

1. *$\|A\|_{\text{HS}} = \|G_{\mathscr{B}}^{1/2}({}_{\mathscr{B}}[A]_{\mathscr{B}})G_{\mathscr{B}}^{\dagger 1/2}\|_{\text{F}}$, where $\|\cdot\|_{\text{HS}}$ is the Hilbert-Schmidt norm of a linear operator, $\|\cdot\|_{\text{F}}$ is the Frobenius norm of a matrix, and $G_{\mathscr{B}}^{\dagger 1/2}$ is the Moore-Penrose inverse of the matrix $G_{\mathscr{B}}^{1/2}$.*

2. *${}_{\mathscr{B}}[(A + cI)^{-1}]_{\mathscr{B}} = G_{\mathscr{B}}^{\dagger 1/2}\{G_{\mathscr{B}}^{1/2}({}_{\mathscr{B}}[A]_{\mathscr{B}})G_{\mathscr{B}}^{\dagger 1/2} + cQ_{\mathscr{B}}\}^{\dagger}G_{\mathscr{B}}^{1/2}$, where $Q_{\mathscr{B}}$ is the projection on to* $\text{span}\{[h_1]_{\mathscr{B}}, \ldots, [h_m]_{\mathscr{B}}\}$.

The proof, which is omitted, can be done using Theorem 8 of Li and Solea (2018a). Note that part 2 of the proposition can be equivalently written as

$$ {}_{\mathscr{B}}[(A + \epsilon I)^{-1}]_{\mathscr{B}} = G_{\mathscr{B}}^{\dagger 1/2}\{G_{\mathscr{B}}^{1/2}({}_{\mathscr{B}}[A]_{\mathscr{B}})G_{\mathscr{B}}^{\dagger 1/2} + \epsilon I_n\}^{\dagger}G_{\mathscr{B}}^{1/2}, $$

because $G_{\mathscr{B}}^{\dagger 1/2}G_{\mathscr{B}}^{1/2} = Q_{\mathscr{B}}$.

4.4.2 Implementation of step 1

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an i.i.d. sample of (X, Y). At the sample level, the centered RKHS $\mathscr{H}_X^{-(i,j)}$ is spanned by the functions

$$\{\kappa_X^{-(i,j)}(\cdot, X_a^{-(i,j)}) - E_n[\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)})] : a = 1, \ldots, n\}, \tag{4.4.1}$$

where $\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)})$ stands for the function $u \mapsto \kappa_X^{-(i,j)}(u, X^{-(i,j)})]$ for any $u \in \Omega_{X^{-(i,j)}}$, and $E_n[\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)})]$ the function $u \mapsto E_n[\kappa_X^{-(i,j)}(u, X^{-(i,j)})]$ for any $u \in \Omega_{X^{-(i,j)}}$. Let $G_{X^{-(i,j)}}$ be the Gram matrix of the spanning set (4.4.1) – that is, the $n \times n$ matrix whose (a, b)th entry is

$$\begin{aligned}\langle \kappa_X^{-(i,j)}(\cdot, X_a^{-(i,j)}) - E_n[\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)})], \\ \kappa_X^{-(i,j)}(\cdot, X_b^{-(i,j)}) - E_n[\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)})]\rangle_{-(i,j)}.\end{aligned}$$

It can be easily verified that $G_{X^{-(i,j)}} = QK_{X^{-(i,j)}}Q$, where $K_{X^{-(i,j)}}$ is the $n \times n$ matrix whose (a, b)th entry is $\kappa_X^{-(i,j)}(X_a^{-(i,j)}, X_b^{-(i,j)})$, and $Q = I_n - 1_n 1_n^\intercal / n$ is the projection onto the orthogonal complement of the subspace of $\mathbb{R}^n$ spanned by the n-dimensional vector $1_n = (1, \ldots, 1)$. Similarly, let $\mathscr{H}_X^{(i,j)}$ be the centered RKHS spanned by the functions

$$\{\kappa_X^{(i,j)}(\cdot, X_a^{(i,j)}) - E_n[\kappa_X^{(i,j)}(\cdot, X^{(i,j)})] : a = 1, \ldots, n\},$$

and $G_{X^{(i,j)}} = QK_{X^{(i,j)}}Q$ the Gram matrix of this set, where $K_{X^{(i,j)}}$ is the $n \times n$ matrix whose (a, b)th entry is $\kappa_X^{(i,j)}(x_a^{(i,j)}, x_b^{(i,j)})$.

Mimicking the population-level definition (4.3.1), we estimate the covariance operators $\Sigma_{X^{-(i,j)}X^{(i,j)}}$ and $\Sigma_{X^{-(i,j)}X^{-(i,j)}}$ by

$$\begin{aligned}\hat{\Sigma}_{X^{-(i,j)}X^{(i,j)}} = & E_n\{[\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)}) - E_n\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)})] \\ & \otimes [\kappa_X^{(i,j)}(\cdot, X^{(i,j)}) - E_n\kappa_X^{(i,j)}(\cdot, X^{(i,j)})]\} \\ \hat{\Sigma}_{X^{-(i,j)}X^{-(i,j)}} = & E_n\{[\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)}) - E_n\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)})] \\ & \otimes [\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)}) - E_n\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)})]\},\end{aligned}$$

respectively. We estimate $\Sigma^{-1}_{X^{(i,j)}X^{(i,j)}}$ by the Tychonoff-regularized inverse

$$(\hat{\Sigma}_{X^{(i,j)}X^{(i,j)}} + \epsilon_X^{(i,j)} I)^{-1},$$

where $I : \mathscr{H}_X^{(i,j)} \to \mathscr{H}_X^{(i,j)}$ is the identity operator. So, at the sample level, the generalized eigenvalue problem to be solved is, at the kth step,

$$\begin{aligned} &\text{maximize} \quad \langle f, \hat{\Sigma}_{X^{-(i,j)}X^{(i,j)}} (\hat{\Sigma}_{X^{(i,j)}X^{(i,j)}} + \epsilon_X^{(i,j)} I)^{-1} \hat{\Sigma}_{X^{(i,j)}X^{-(i,j)}} f \rangle_{-(i,j)} \\ &\text{subject to} \quad \begin{cases} \langle f, \hat{\Sigma}_{X^{-(i,j)}X^{-(i,j)}} f \rangle_{-(i,j)} = 1, \\ \langle f, \hat{\Sigma}_{X^{-(i,j)}X^{-(i,j)}} f_i \rangle_{-(i,j)} = 0, \quad i = 1, \ldots, k-1, \end{cases} \end{aligned}$$

where $f_1, \ldots, f_{k-1}$ are the maximizers in the previous steps. The first d_{ij} solutions to the above iterative maximization problem are an estimate of a basis in the central class $\mathfrak{S}_{X^{(i,j)}|X^{-(i,j)}}$.

By representing the operators involved in the above:

$$\hat{\Sigma}_{X^{-(i,j)}X^{(i,j)}}, \quad (\hat{\Sigma}_{X^{(i,j)}X^{(i,j)}} + \epsilon_X^{(i,j)} I)^{-1}, \quad \hat{\Sigma}_{X^{(i,j)}X^{-(i,j)}}, \quad \hat{\Sigma}_{X^{-(i,j)}X^{-(i,j)}}$$

with respect to the spanning systems of their domain and range spaces, we can re-express the above maximization problem in terms of matrices as

$$\begin{aligned} &\text{maximize} \quad b^\intercal G_{X^{-(i,j)}} G_{X^{(i,j)}} (G_{X^{(i,j)}} + \epsilon_X^{(i,j)} I_n)^{-1} G_{X^{-(i,j)}} b \\ &\text{subject to} \quad b^\intercal G^2_{X^{-(i,j)}} b = 1. \end{aligned}$$

To solve the above generalized eigenvalue problem, we set $G_{X^{-(i,j)}} b = a$. Solving this equation for a with Tychonoff regularization, we have $b = (G_{X^{-(i,j)}} + \epsilon_X^{-(i,j)} I_n)^{-1} a$. Thus, at the kth step, a is simply the kth eigenvector of the matrix

$$\begin{aligned} &(G_{X^{-(i,j)}} + \epsilon_X^{-(i,j)} I_n)^{-1} G_{X^{-(i,j)}} \\ &\qquad G_{X^{(i,j)}} (G_{X^{(i,j)}} + \epsilon_X^{(i,j)} I_n)^{-1} G_{X^{-(i,j)}} (G_{X^{-(i,j)}} + \epsilon_X^{-(i,j)} I_n)^{-1}. \end{aligned}$$

Let $a^1, \ldots, a^{d_{ij}}$ be the first d_{ij} eigenvectors for this standard eigenvalue problem,

and

$$b^r = (G_{X^{-(i,j)}} + \epsilon_X^{-(i,j)} I_n)^{-1} a^r, \quad r = 1, \ldots, d_{ij}.$$

The vectors $b^1, \ldots, b^{d_{ij}}$ are the coordinates of $f_1^{ij}, \ldots, f_{d_{ij}}^{ij}$ with respect to the spanning system (4.4.1). That is,

$$f_r^{ij} = \sum_{a=1}^{n} b_a^r \{\kappa_X^{-(i,j)}(\cdot, X_a^{-(i,j)}) - E_n[\kappa_X^{-(i,j)}(\cdot, X^{-(i,j)})]\}.$$

The statistics $\hat{U}_a^{ij} = (f_1^{ij}(X_a^{-(i,j)}), \ldots, f_{d_{ij}}^{ij}(X_a^{-(i,j)}))$, $a = 1, \ldots, n$, will be used as the input for the second step.

4.4.3 Implementation of step 2

This step consists of estimating the CCCO's for each (i, j) and thresholding their norms to obtain the estimated edge set. At the sample level, for each $(i, j) \in \Gamma \times \Gamma$, the centered RKHS's generated by the kernels $\kappa_{XU}^{i,ij}$, $\kappa_{XU}^{j,ij}$, and κ_U^{ij} are, respectively,

$$\begin{aligned}
\mathscr{H}_{XU}^{i,ij} &= \text{span}\{\kappa_{XU}^{i,ij}(\cdot, (X_a^i, U_a^{ij})) - E_n[\kappa_{XU}^{i,ij}(\cdot, (X^i, U^{ij}))] : a = 1, \ldots, n\}, \\
\mathscr{H}_{XU}^{j,ij} &= \text{span}\{\kappa_{XU}^{j,ij}(\cdot, (X_a^j, U_a^{ij})) - E_n[\kappa_{XU}^{j,ij}(\cdot, (X^j, U^{ij}))] : a = 1, \ldots, n\}, \\
\mathscr{H}_U^{ij} &= \text{span}\{\kappa_U^{ij}(\cdot, U_a^{ij}) - E_n[\kappa_U^{ij}(\cdot, U^{ij})] : a = 1, \ldots, n\},
\end{aligned}$$

where, for example, $\kappa_{XU}^{i,ij}(\cdot, (X_a^i, U_a^{ij}))$ denotes the function

$$\Omega_{X^i} \times \Omega_{U^{ij}} \to \mathbb{R}, \quad (x^i, u^{ij}) \mapsto \kappa_{XU}^{i,ij}((x^i, u^{ij}), (X_a^i, U_a^{ij}))$$

and $E_n[\kappa_{XU}^{i,ij}(\cdot, (X^i, U^{ij}))]$ denotes the function

$$\Omega_{X^i} \times \Omega_{U^{ij}} \to \mathbb{R}, \quad (x^i, u^{ij}) \mapsto E_n[\kappa_{XU}^{i,ij}((x^i, u^{ij}), (X^i, U^{ij}))].$$

We now develop the sample estimate of the CCCO. Mimicking the population-level kernel representations of the covariance operators in (4.3.1), we estimate the

covariance operators $\Sigma_{(X^{i}U^{ij})(X^{i}U^{ij})}$, $\Sigma_{(X^{i}U^{ij})U^{ij}}$, $\Sigma_{X^{j}(X^{j}U^{ij})}$, and $\Sigma_{U^{ij}U^{ij}}$ by

$$\begin{aligned}
\hat{\Sigma}_{(X^{i}U^{ij})(X^{j}U^{ij})} =& E_n\{[\kappa_{XU}^{i,ij}(\cdot,(X^{i},U^{ij})) - E_n\kappa_{XU}^{i,ij}(\cdot,(X^{i},U^{ij}))] \\
&\otimes [\kappa_{XU}^{j,ij}(\cdot,(X^{j},U^{ij})) - E_n\kappa_{XU}^{j,ij}(\cdot,(X^{j},U^{ij}))]\} \\
\hat{\Sigma}_{(X^{i}U^{ij})U^{ij}} =& E_n\{[\kappa_{XU}^{i,ij}(\cdot,(X^{i},U^{ij})) - E_n\kappa_{XU}^{i,ij}(\cdot,(X^{i},U^{ij}))] \\
&\otimes [\kappa_{U}^{ij}(\cdot,U^{ij}) - E_n\kappa_{U}^{ij}(\cdot,U^{ij})]\} \\
\hat{\Sigma}_{U^{ij}(X^{j}U^{ij})} =& E_n\{[\kappa_{U}^{ij}(\cdot,U^{ij}) - E_n\kappa_{U}^{ij}(\cdot,U^{ij})] \\
&\otimes [\kappa_{XU}^{j,ij}(\cdot,(X^{j},U^{ij})) - E_n\kappa_{XU}^{j,ij}(\cdot,(X^{j},U^{ij}))]\} \\
\hat{\Sigma}_{U^{ij}U^{ij}} =& E_n\{[\kappa_{U}^{ij}(\cdot,U^{ij}) - E_n\kappa_{U}^{ij}(\cdot,U^{ij})] \\
&\otimes [\kappa_{U}^{ij}(\cdot,U^{ij}) - E_n\kappa_{U}^{ij}(\cdot,U^{ij})]\},
\end{aligned} \tag{4.4.2}$$

respectively. Mimicking the population-level Definition 2 of the CCCO, we estimate this operator by

$$\hat{\Sigma}_{\ddot{X}^{i}\ddot{X}^{j}|U^{ij}} = \hat{\Sigma}_{(X^{i}U^{ij})(X^{j}U^{ij})} - \hat{\Sigma}_{(X^{i}U^{ij})U^{ij}}(\hat{\Sigma}_{U^{ij}U^{ij}} + \epsilon_{U}^{(i,j)}I)^{-1}\hat{\Sigma}_{U^{ij}(X^{j}U^{ij})},$$

where, again, we have used Tychonoff regularization to estimate the inverted covariance operator $\Sigma_{U^{ij}U^{ij}}$.

Next, we derive the coordinate representation of the CCCO. Let $K_{U^{ij}}$, $K_{X^{i}U^{ij}}$, and $K_{X^{j}U^{ij}}$ be the Gram matrices

$$\begin{aligned}
K_{U^{ij}} =& \{\kappa_{U}^{ij}(U_{a}^{ij},U_{b}^{ij})\}_{a,b=1}^{n}, \\
K_{X^{i}U^{ij}} =& \{\kappa_{XU}^{i,ij}((X_{a}^{i},U_{a}^{ij}),(X_{b}^{i},U_{b}^{ij}))\}_{a,b=1}^{n}, \\
K_{X^{j}U^{ij}} =& \{\kappa_{XU}^{j,ij}((X_{a}^{j},U_{a}^{ij}),(X_{b}^{j},U_{b}^{ij}))\}_{a,b=1}^{n},
\end{aligned}$$

and $G_{X^{i}U^{ij}}$, $G_{X^{j}U^{ij}}$, and $G_{U^{ij}}$ their centered versions

$$G_{X^{i}U^{ij}} = QK_{X^{i}U^{ij}}Q, \quad G_{X^{j}U^{ij}} = QK_{X^{j}U^{ij}}Q, \quad G_{U^{ij}} = QK_{U^{ij}}Q.$$

By Theorem 12.1 of Li (2018c), the coordinate representations of the estimated covariance operators in (4.4.2) are

$$\mathscr{B}_{XU}^{i,ij}[\hat{\Sigma}_{(X^{i}U^{ij})(X^{j}U^{ij})}]_{\mathscr{B}_{XU}^{j,ij}} = G_{X^{j}U^{ij}}, \quad \mathscr{B}_{XU}^{i,ij}[\hat{\Sigma}_{(X^{i}U^{ij})U^{ij}}]_{\mathscr{B}_{U}^{ij}} = G_{U^{ij}},$$

$$\mathscr{B}_U^{ij}[\hat{\Sigma}_{U^{ij}(X^jU^{ij})}]_{\mathscr{B}_{XU}^{j,ij}} = G_{X^jU^{ij}}, \quad \mathscr{B}_U^{ij}[\hat{\Sigma}_{U^{ij}U^{ij}}]_{\mathscr{B}_U^{ij}} = G_{U^{ij}}.$$

Applying the above relations and part 2 of Proposition 6, we obtain the coordinate representation of the CCCO for each (i, j) as

$$\mathscr{B}_{XU}^{i,ij}[\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|U^{ij}}]_{\mathscr{B}_{XU}^{j,ij}} = G_{X^jU^{ij}} - G_{U^{ij}}(G_{U^{ij}} + \epsilon_U^{(i,j)}Q)^\dagger G_{X^jU^{ij}}.$$

By part 1 of Proposition 6, the Hilbert Schmidt norm of the above operator is

$$\begin{aligned}\|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\text{HS}} &= \left\| G_{X^iU^{ij}}^{1/2}\left(\mathscr{B}_{XU}^{i,ij}[\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|U^{ij}}]_{\mathscr{B}_{XU}^{j,ij}}\right) G_{X^jU^{ij}}^{\dagger 1/2}\right\|_{\text{F}} \\ &= \left\| G_{X^iU^{ij}}^{1/2}G_{X^jU^{ij}}^{1/2} - G_{X^iU^{ij}}^{1/2}G_{U^{ij}}(G_{U^{ij}} + \epsilon_U^{(i,j)}Q)^\dagger G_{X^jU^{ij}}^{1/2}\right\|_{\text{F}}.\end{aligned}$$

Estimation of the edge set is then based on thresholding this norm; that is,

$$\hat{\mathcal{E}} = \{(i, j) \in \Gamma \times \Gamma : i > j,\ \|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\text{HS}} > \rho_n\}$$

for some chosen $\rho_n > 0$.

4.5 Tuning

We now develop methods to determine the tuning parameters in our 2-step procedure. We have two types of tuning constants: those for the kernels and those for Tychonoff regularization. For the Tychonoff regularization, we have $\epsilon_X^{(i,j)}$ and $\epsilon_X^{-(i,j)}$ for step 1, and $\epsilon_U^{(i,j)}$ for step 2. In this project we use the Gaussian radial basis function (RBF) as the kernel, which takes the form

$$\kappa(u, v) = \exp(-\gamma\|u - v\|^2)$$

with u, v being members of a metric space. For each (i, j), we have five γ's to determine: $\gamma_X^{(i,j)}$ for the kernel $\kappa_X^{(i,j)}$, $\gamma_X^{-(i,j)}$ for $\kappa_X^{-(i,j)}$, $\gamma_{XU}^{i,ij}$ for $\kappa_{XU}^{i,ij}$, $\gamma_{XU}^{j,ij}$ for $\kappa_{XU}^{j,ij}$, and γ_U^{ij} for κ_U^{ij}.

We choose the γ's by the following general formula (see, for example, Li (2018c))

$$1/\sqrt{\gamma} = \binom{n}{2}^{-1}\sum_{a<b}\|s_a - s_b\|, \tag{4.5.1}$$

where, according to which kernel is concerned, $s_1, \ldots, s_n$ can be any of the samples:

i. $\{X_a^{(i,j)} : a = 1, \ldots, n\}$ for $\kappa_X^{(i,j)}$,

ii. $\{X_a^{-(i,j)} : a = 1, \ldots, n\}$ for $\kappa_X^{-(i,j)}$,

iii. $\{(X_a^i, U_a^{ij}) : a = 1, \ldots, n\}$ for $\kappa_{XU}^{i,ij}$,

iv. $\{(X_a^j, U_a^{ij}) : a = 1, \ldots, n\}$ for $\kappa_{XU}^{j,ij}$,

v. $\{U_a^i : a = 1, \ldots, n\}$ for κ_U^{ij}.

For the tuning parameters in Tychonoff regularization, we use the following generalized cross validation scheme (Golub et al., 1979):

$$\mathrm{GCV}(\epsilon) = \mathrm{argmin}_\epsilon \sum_{i<j} \frac{\|G_1 - G_2^\intercal [G_2 + \epsilon\, \lambda_{\max}(G_2)]^{-1} G_1\|_{\mathrm{F}}}{\frac{1}{n}\mathrm{tr}\{I_n - G_2^\intercal [G_2 + \epsilon\, \lambda_{\max}(G_2)]^{-1}\}},$$

where $G_1, G_2 \in \mathbb{R}^{n\times n}$ are positive semidefinite matrices, and $\lambda_{\max}(G_2)$ is the largest eigenvalue of G_2. The matrices G_1 and G_2 are the following matrices for the three tuning parameters:

1. $G_1 = G_{X^{-(i,j)}}$, $G_2 = G_{X^{(i,j)}}$ for $\epsilon_X^{(i,j)}$,

2. $G_1 = G_{X^{(i,j)}}$, $G_2 = G_{X^{-(i,j)}}$ for $\epsilon_X^{-(i,j)}$,

3. $G_1 = G_{X^{(i,j)}}$, $G_2 = G_{U^{ij}}$ for $\epsilon_U^{(i,j)}$,

We minimize this criterion over a grid to determine optimal ϵ, as detailed in Section 6.3.

4.6 Asymptotic theory

4.6.1 Overview

Our ultimate goal is to derive, for each (i, j), the convergence rate of

$$\left| \|\hat{\Sigma}_{\ddot{X}^i \ddot{X}^j | \hat{U}^{ij}}\|_{\mathrm{HS}} - \|\Sigma_{\ddot{X}^i \ddot{X}^j | U^{ij}}\|_{\mathrm{HS}} \right|,$$

because it is $\|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|\hat{U}^{ij}}\|_{\text{HS}}$ that we threshold to estimate the sufficient graphical model. This is by no means an easy task because our estimation procedure involves two steps: in the first step we use GSIR to extract the estimated sufficient predictors $\hat{U}^{ij}$ using a first set of kernels; in the second step we substitute the estimated predictors into a second set of kernels to get the final result. The main challenge is to understand how the error propagates from the first step to the second, and this is the most novel aspect of our asymptotic theory.

Specifically, by the triangular inequality,

$$\begin{aligned}
&\left|\|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|\hat{U}^{ij}}\|_{\text{HS}} - \|\Sigma_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\text{HS}}\right| \\
&\leq \|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|\hat{U}^{ij}} - \Sigma_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\text{HS}} \\
&\leq \|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|\hat{U}^{ij}} - \hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\text{HS}} + \|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|U^{ij}} - \Sigma_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\text{HS}}.
\end{aligned}$$

Thus we need to derive the convergence rates of the following three quantities:

$$\begin{aligned}
&\text{(i)} \quad \|\hat{U}^{ij} - U^{ij}\|_{[\mathscr{H}^{-(i,j)}(X)]^{d_{ij}}}; \\
&\text{(ii)} \quad \|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|\hat{U}^{ij}} - \hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\text{HS}}; \\
&\text{(iii)} \quad \|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|U^{ij}} - \Sigma_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\text{HS}},
\end{aligned} \tag{4.6.1}$$

where, to avoid overly crowded notation, we have used $\mathscr{H}^{-(i,j)}(X)$ to replace $\mathscr{H}_X^{-(i,j)}$ when it occurs as a subscript. The first and third convergence rates can be derived using the tools for asymptotic analysis of linear operators developed in Fukumizu et al. (2007), Li and Song (2017b), Lee et al. (2016b), and Solea and Li (2020). The second convergence rate is, however, a new problem, and it will also be useful in similar settings that require constructing estimators based on predictors extracted by sufficient dimension reduction. In some sense, this is akin to the post dimension reduction problem considered in Kim et al. (2020).

To make the basic logic line underlying the development discernable, we will present not only the final results but also the key intermediate results and the intuitions about them, even though some of the intermediate results are used only in the proofs in the Supplementary Material.

A few words about the notation: if $\{a_n\}$ and $\{b_n\}$ are sequences of positive numbers, then we write $a_n \prec b_n$ if $a_n/b_n \to 0$. We write $a_n \asymp b_n$ if

$0 < \liminf_n(b_n/a_n) \leq \limsup_n(b_n/a_n) < \infty$. We write $b_n \preceq a_n$ if either $b_n \prec a_n$ or $b_n \asymp a_n$. Because (i, j) is fixed in the asymptotic development, and also to emphasize the dependence on n, in the rest of this section we denote $\epsilon_X^{(i,j)}$, $\epsilon_X^{-(i,j)}$, and $\epsilon_U^{(i,j)}$ by ϵ_n, η_n, and δ_n, respectively.

4.6.2 Transparent kernel

In this subsection we first develop what we call the "transparent kernel" that allows information to transmit from step 1 to step 2 efficiently. Let Ω be a nonempty set, and $\kappa : \Omega \times \Omega \to \mathbb{R}$ a positive definite kernel.

Definition 3. *We say that κ is a transparent kernel if, for each $t \in \Omega$, the function $s \mapsto \kappa(s, t)$ is twice differentiable and*

1. *$\partial\kappa(s, t)/\partial s|_{s=t} = 0$;*
2. *the matrix $H(s, t) = \partial^2\kappa(s, t)/\partial s\partial s^\intercal$ has a bounded operator norm; that is, there exist $-\infty < C_1 \leq C_2 < \infty$ such that*

$$C_1 \leq \lambda_{\min}(H(s, t)) \leq \lambda_{\max}(H(s, t)) < C_2$$

for all $(s, t) \in \Omega \times \Omega$, where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ indicate the largest and smallest eigenvalues.

This condition implies a type of Lipschitz continuity in a setting that involves two reproducing kernels κ_0 and κ_1, where the argument of κ_1 is the evaluation of a member of the RKHS generated by κ_0.

Theorem 5. *Suppose Ω is a nonempty set, $\kappa_0 : \Omega \times \Omega \to \mathbb{R}$ is a positive definite kernel, $\mathscr{H}_0$ is the RKHS generated by κ_0, and $\mathscr{H}_0^d$ is the d-fold Cartesian product of $\mathscr{H}_0$ with inner product defined by*

$$\langle U, V\rangle_{\mathscr{H}_0^d} = \langle u_1, v_1\rangle_{\mathscr{H}_0} + \cdots + \langle u_d, v_d\rangle_{\mathscr{H}_0}$$

where $U = (u_1, \ldots, u_d)$ and $V = (v_1, \ldots, v_d)$ are members of $\mathscr{H}_0^d$. Furthermore, suppose $\kappa_1 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a positive definite kernel, and $\mathscr{H}_1$ is the RKHS generated by κ_1.

(i) For any $U, V \in \mathscr{H}_0^d$, $a \in \Omega$, *we have*

$$\|U(a) - V(a)\|_{\mathbb{R}^d} \leq [\kappa_0(a,a)]^{1/2} \, \|U - V\|_{\mathscr{H}_0{}^d}.$$

(ii) If $\kappa_1(s,t)$ *is a transparent kernel, then there exists a* $C > 0$ *such that, for each* $U, V \in \mathscr{H}_0^d$ *and* $a \in \Omega$,

$$\|\kappa_1(\cdot, U(a)) - \kappa_1(\cdot, V(a))\|_{\mathscr{H}_1} \leq C \, [\kappa_0(a,a)]^{1/2} \, \|U - V\|_{\mathscr{H}_0^d}.$$

A direct consequence of this theorem is that, if $\hat{U}$ is an estimate of some U, a member of $\mathscr{H}_0^d$, with $\|\hat{U} - U\|_{\mathscr{H}_0{}^d} = O_P(b_n)$ for some $0 < b_n \to 0$, $\hat{\Sigma}(\hat{U})$ is a linear operator estimated from the sample $\hat{U}_1, \ldots, \hat{U}_n$ (and perhaps some other random vectors), and $\hat{\Sigma}(U)$ is a linear operator estimated from the sample $U_1, \ldots, U_n$, then,

$$\|\hat{\Sigma}(\hat{U}) - \hat{\Sigma}(U)\|_{\text{HS}} = O_P(b_n). \tag{4.6.2}$$

This result is somewhat surprising, because sample estimates such as $\hat{\Sigma}(\hat{U})$ can be viewed as $E_n \mathbb{G}(X, \hat{U})$, where, in our case, $\hat{U}$ is an estimate of a function U in a functional space with norm $\|\cdot\|$ and $\mathbb{G}$ is an operator-valued function. If $\|\hat{U} - U\| = O_P(b_n)$ for some $b_n \to 0$, then it is not necessarily true that

$$\|E_n \mathbb{G}(X, \hat{U}) - E_n \mathbb{G}(X, U)\| = O_P(b_n),$$

particularly when U is an infinite dimensional object. Yet relation (4.6.2) states exactly this. The reason behind this somewhat surprising result is that the reproducing kernel property allows us to separate the function $\hat{U}$ and its argument X_a (i.e. $\hat{U}(x) = \langle \hat{U}, \kappa(\cdot, x) \rangle$) and this implies a type of uniformity among $\hat{U}(X_1), \ldots, \hat{U}(X_n)$. This point will be made clear in the proof in the Supplementary Material. Statement (4.6.2) is made precise by the next theorem, tailored to our context.

Theorem 6. *Suppose conditions (1) and (2) of Definition 2 are satisfied with* U, V, W *therein replaced by* X^i, X^j, *and* U^{ij}. *Suppose, furthermore:*

(a) κ_U^{ij}, $\kappa_{XU}^{i,ij}$, *and* $\kappa_{XU}^{j,ij}$ *are transparent kernels;*

(b) $\|\hat{U}^{ij} - U^{ij}\|_{[\mathscr{H}^{-(i,j)}(X)]^{d_{ij}}} = O_P(b_n)$ *for some* $0 < b_n \to 0$.

Then

(i) $\|\hat{\Sigma}_{\hat{U}^{ij}\hat{U}^{ij}} - \hat{\Sigma}_{U^{ij}U^{ij}}\|_{\text{HS}} = O_P(b_n)$;

(ii) $\|\hat{\Sigma}_{(X^i\hat{U}^{ij})\hat{U}^{ij}} - \hat{\Sigma}_{(X^iU^{ij})U^{ij}}\|_{\text{HS}} = O_P(b_n)$;

(iii) $\|\hat{\Sigma}_{(X^i\hat{U}^{ij})(X^j\hat{U}^{ij})} - \hat{\Sigma}_{(X^iU^{ij})(X^jU^{ij})}\|_{\text{HS}} = O_P(b_n)$.

Using Theorem 6 we can derive the convergence rate of $\|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|\hat{U}^{ij}} - \hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\text{HS}}$.

Theorem 7. *Suppose conditions in Theorem 6 are satisfied and, furthermore,*

(a) $\Sigma^{-1}_{U^{ij}U^{ij}}\Sigma_{U^{ij}(X^iU^{ij})}$ *and* $\Sigma^{-1}_{U^{ij}U^{ij}}\Sigma_{U^{ij}(X^jU^{ij})}$ *are bounded linear operators;*

(b) $b_n \preceq \delta_n \prec 1$.

Then $\|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|\hat{U}^{ij}} - \hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\text{HS}} = O_P(b_n)$.

Note that, unlike in Theorem 6, where our assumptions imply

$$\Sigma^{-1}_{X^{-(i,j)}X^{-(i,j)}}\Sigma_{X^{-(i,j)}X^{(i,j)}}$$

is a finite-rank operator, here, we do not assume $\Sigma^{-1}_{U^{ij}(U^{ij})}\Sigma_{U^{ij}(X^jU^{ij})}$ to be a finite-rank (or even Hilbert-Schmidt) operator; instead, we assume it to be a bounded operator. This is because the random vector (X^j, U^{ij}) contains U^{ij}, and this makes it unreasonable to assume $\Sigma^{-1}_{U^{ij}(U^{ij})}\Sigma_{U^{ij}(X^jU^{ij})}$ to be finite-rank or Hilbert Schmidt. For example, when X^j is a constant, $\Sigma_{U^{ij}(X^jU^{ij})}$ is the same as $\Sigma_{U^{ij}U^{ij}}$ and $\Sigma^{-1}_{U^{ij}U^{ij}}\Sigma_{U^{ij}U^{ij}}$ is not a Hilbert Schmidt operator, though it is bounded.

Theorem 7 shows that convergence rate of (ii) in (4.6.1) is the same as the convergence rate of (i) in (4.6.1). Thus, it now remains to derive the convergence rate of (i) and (iii).

4.6.3 Convergence rates of (i) and (iii) in (4.6.1)

We first present the convergence rate of $\hat{U}^{ij}$ to U^{ij}. The proof is in the same spirit as that of Theorem 5 of Li and Song (2017b), but there are two differences. First, Li and Song (2017b) took A in (4.3.3) to be I, whereas we take it to be Σ_{YY}. In particular, the GSIR in Li and Song (2017b) only has one tuning parameter η_n, but

we have two tuning parameters η_n and ϵ_n. Second, Li and Song (2017b) defined (in the current notation) f_r^{ij} to be the eigenfunctions of

$$\Sigma^{-1}_{X^{-(i,j)}X^{-(i,j)}}\Sigma_{X^{-(i,j)}X^{(i,j)}}\Sigma^{-1}_{X^{(i,j)}X^{(i,j)}}\Sigma_{X^{(i,j)}X^{-(i,j)}}\Sigma^{-1}_{X^{-(i,j)}X^{-(i,j)}},$$

which is different from the generalized eigenvalue problem (4.3.3). For these reasons we need to re-derive the convergence rate of $\hat{U}^{ij}$.

Theorem 8. *Suppose*

(a) Assumption 7 is satisfied;

(b) $\Sigma_{X^{-(i,j)}X^{(i,j)}}$ *is a finite-rank operator with*

$$\begin{aligned}\mathrm{ran}(\Sigma_{X^{-(i,j)}X^{(i,j)}}) &\subseteq \mathrm{ran}(\Sigma^{3/2}_{X^{-(i,j)}X^{-(i,j)}}),\\ \mathrm{ran}(\Sigma_{X^{(i,j)}X^{-(i,j)}}) &\subseteq \mathrm{ran}(\Sigma_{X^{(i,j)}X^{(i,j)}});\end{aligned}$$

(c) $n^{-1/2} \prec \eta_n \prec 1$, $n^{-1/2} \prec \epsilon_n \prec 1$;

(d) for each $r = 1, \ldots, d_{ij}$, $\lambda_1^{ij} > \cdots > \lambda_{d_{ij}}^{ij}$.

Then

$$\|\hat{U}^{ij} - U^{ij}\|_{[\mathscr{H}^{-(i,j)}(X)]^{d_{ij}}} = O_P(\eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n^{1/2} + \epsilon_n).$$

An immediate consequence is that, under the transparent kernel assumption, the b_n in Theorem 7 is the same as this rate. We next derive the convergence rate in (iii) of (4.6.1). This rate depends on the tuning parameter δ_n in the estimate of CCCO, and it reaches b_n for the optimal choice of δ_n.

Theorem 9. *Suppose conditions (1) and (2) of Definition 2 are satisfied with* U, V, W *therein replaced by* X^i, X^j, *and* U^{ij}. *Suppose, furthermore,*

(a) $\Sigma^{-1}_{U^{ij}U^{ij}}\Sigma_{U^{ij}(X^iU^{ij})}$ *and* $\Sigma^{-1}_{U^{ij}U^{ij}}\Sigma_{U^{ij}(X^jU^{ij})}$ *are bounded linear operators;*

(b) $b_n \preceq \delta_n \prec 1$.

Then $\|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|U^{ij}} - \Sigma_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\text{HS}} = O_P(\delta_n)$. *Consequently, if* $\delta_n \asymp b_n$, *then*

$$\|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|U^{ij}} - \Sigma_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\text{HS}} = O_P(b_n).$$

Finally, we combine Theorem 7 through Theorem 9 to come up with the convergence rate of $\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|\hat{U}^{ij}}$. Since there are numerous cross references among the conditions in these theorems, to make a clear presentation we list all the original conditions in the next theorem, even if they already appeared. These conditions are of two categories: those for the step 1 that involves sufficient dimension reduction of $X^{(i,j)}$ versus $X^{-(i,j)}$, and those for the step 2 that involves the estimation of the CCCO. We refer to them as the first-level and second-level conditions, respectively. Each of the two categories then contains three sub-categories: conditions for the kernels, conditions for the linear operators, and conditions for the tuning parameters.

Theorem 10. *Suppose the following conditions hold:*

(a) (First-level kernel) $E[\kappa(S,S)] < \infty$ *is satisfied for* $\kappa = \kappa_X^{(i,j)}$ *and* $\kappa = \kappa_X^{-(i,j)}$;

(b) (First-level operator) $\Sigma_{X^{-(i,j)}X^{(i,j)}}$ *is a finite-rank operator with rank* d_{ij} *and*

$$\begin{aligned}\text{ran}(\Sigma_{X^{-(i,j)}X^{(i,j)}}) &\subseteq \text{ran}(\Sigma^{3/2}_{X^{-(i,j)}X^{-(i,j)}}),\\ \text{ran}(\Sigma_{X^{(i,j)}X^{-(i,j)}}) &\subseteq \text{ran}(\Sigma_{X^{(i,j)}X^{(i,j)}});\end{aligned}$$

all the nonzero eigenvalues of $\Sigma_{X^{(i,j)}X^{-(i,j)}}\Sigma^{-1}_{X^{-(i,j)}X^{-(i,j)}}\Sigma_{X^{-(i,j)}X^{(i,j)}}$ *are distinct;*

(c) (First-level tuning parameters) $n^{-1/2} \prec \eta_n \prec 1$, $n^{-1/2} \prec \epsilon_n \prec 1$, $\eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n^{1/2} + \epsilon_n \prec 1$;

(d) (Second-level kernel) $E[\kappa(S,S)] < \infty$ *is satisfied for* $\kappa = \kappa_U^{ij}$, $\kappa_{XU}^{i,ij}$, *and* $\kappa_{XU}^{j,ij}$; *furthermore, they are transparent kernels;*

(e) (Second-level operators) $\Sigma^{-1}_{U^{ij}U^{ij}}\Sigma_{U^{ij}(X^iU^{ij})}$ *and* $\Sigma^{-1}_{U^{ij}U^{ij}}\Sigma_{U^{ij}(X^jU^{ij})}$ *are bounded linear operators;*

(f) (Second-level tuning parameter) $\delta_n \asymp \eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n^{1/2} + \epsilon_n$.

Then

$$\|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|\hat{U}^{ij}} - \Sigma_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\text{HS}} = O_P(\eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n^{1/2} + \epsilon_n). \tag{4.6.3}$$

Using this result we immediately arrive at the variable selection consistency of the Sufficient Graphical Model.

Corollary 3. *Under the conditions in Theorem 10, if*

$$\eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n^{1/2} + \epsilon_n \prec \rho_n \prec 1, \quad \textit{and}$$
$$\hat{\mathcal{E}} = \{(i,j) \in \Gamma \times \Gamma : \ i > j, \ \|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|\hat{U}^{ij}}\|_{\text{HS}} < \rho_n\}$$

then $\lim_{n\to\infty} P(\hat{\mathcal{E}} = \mathcal{E}) \to 1$.

4.6.4 Optimal rates of tuning parameters

The convergence rate in Theorem 10 depends on ϵ_n and η_n explicitly, and δ_n implicitly (in the sense that $\delta_n \asymp \eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n^{1/2} + \epsilon_n$ is optimal for fixed ϵ_n and η_n). Intuitively, when ϵ_n, η_n, and δ_n increase, the biases increase and variances decrease, and when these tuning parameters decrease, the biases decrease and the variances increase. Thus there exist optimal rates for these tuning parameters that balance the bias and variance to achieve optimal convergence rate for the estimated CCCO.

Theorem 11. *Under the conditions in Theorem 10, if ϵ_n, η_n, and δ_n are of the form n^a, n^b, and n^c for some $a > 0$, $b > 0$, and $c > 0$, then*

(i) the optimal rates the tuning parameters are

$$n^{-1/3} \preceq \epsilon_n \preceq n^{-1/6}, \quad \eta_n \asymp n^{-1/3}, \quad \delta_n \asymp n^{-1/6};$$

(ii) the optimal convergence rate of the estimated CCCO is

$$\|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|\hat{U}^{ij}} - \Sigma_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\text{HS}} = O_P(n^{-1/6}).$$

Note that there is a range of ϵ_n are optimal, this is because the convergence rate does not have a unique minimizer. This also means the result is not very sensitive to this tuning parameter.

We should also mention that the above optimal rate is built upon the convergence rate in Theorem 10, which we suspect can be further improved. To avoid extraordinary lengthy proofs we have taken some (legitimate) shortcuts. As a consequence some upper bounds may not be the tightest possible. In particular, the optimal rate derived by Li and Song (2017b) is $O_P(n^{-1/4})$. It is plausible that such a rate can also be achieved by the current estimator. We leave this point to future research

4.7 Discussion

This project is a first attempt to take advantage of the recently developed nonlinear sufficient dimension reduction methods to construct a nonparametric estimate of the statistical graphical model that does not rely on high-dimensional kernels. We use nonlinear SDR as a module and apply it repeatedly when evaluating $\binom{p}{2}$ conditional independence conditions, which leads to a substantial gain in accuracy in the high-dimensional setting.

Compared with parametric or semiparametric models such as GGM and CGGM, the performance of SGM is not affected by the violation of the Gaussian and copula Gaussian assumptions. Compared with the additive methods such as APCO, SGM retains the interpretation of conditional independence as the criterion to determine the absence of an edge, which is a commonly accepted criterion for constructing graphical models; moreover, it does not require the additive structure imposed by APCO. Compared with fully nonparametric methods such as NP1 and NP2, the nonlinear dimension reduction in the first step of SGM avoids the curse of dimensionality and significantly enhances the performance.

On the theoretical side, under a reasonably mild set of conditions on the kernel (i.e. "transparent kernel"), we discovered that, the relevant linear operators based on the estimated sufficient predictor $\hat{U}^{ij}$ and those based on the true sufficient predictor U^{ij} has the same convergence rate as $\|\hat{U}^{ij} - U^{ij}\|$. This is a powerful insight that can have a much broader impact than it does in the current context. In fact,

this general relation will be useful whenever we need to derive the convergence rate of an estimator based on an estimated sufficient predictor obtained by nonlinear SDR as the initial step, such as nonparametric regression.

The present framework also opens up several potential directions for further research. First, the current model assumes that the central class $\mathfrak{S}_{X^{(i,j)}|X^{-(i,j)}}$ is complete, so that GSIR is the exhaustive nonlinear SDR estimate. When this condition is violated, GSIR is no longer exhaustive and we can employ other nonlinear SDR methods such as the GSAVE (Lee et al., 2013; Li, 2018c) to recover the part of the central class that GSIR misses. It would be useful to work out the numerical algorithms and convergence rates in this circumstance. Second, though we have assumed that there is a proper sub-σ-field $\mathcal{G}^{-(i,j)}$ of $\sigma(X^{-(i,j)})$ that is sufficient, the estimation procedure we propose here is still justifiable even if no such sub-σ-field exists. In this case, U^{ij} is still the most important set of functions that characterize the statistical dependence of $X^{(i,j)}$ on $X^{-(i,j)}$ – even though it does not completely capture the conditional dependence. However, in this case, our method may be more appropriately called the Principal Graphical Model rather than the Sufficient Graphical Model. Third, it is plausible that the current method can be extended to the situations where the observation on each vertex is a random function rather than a random variable. This type of networks are common in medical applications such as EEG and fMRI, and several functional graphical models have been proposed recently. See, for example, Zhu et al. (2016), Qiao et al. (2019), Li and Solea (2018b), and Solea and Li (2020). We could presumably use functional GSIR developed in Li and Song (2017b) to carry out functional sufficient dimension reduction in a functional sufficient graphical model. Fourth, in this project we have used the norm of the conjoined conditional covariance operator as the numerical measure of conditional independence. However, it is plausible that more sensitive measures exist. This speculation is inspired by two existing parallel situations. The first is that, under the Gaussian distribution, the partial correlation often works better than conditional covariance as the former removes the effect of the marginal variance. The second is that, as shown in Lee et al. (2016a), the additive partial correlation operator works better than the additive conditional covariance operator for evaluating additive conditional independence. Likewise, a parallel construct of a *Conjoined Partial Correlation Operator* might

work better than the CCCO in evaluating conditional independence. The asymptotic properties of such an operator, as well as its performance in constructing a Sufficient Graphical Model, deserve to be careful studied.

Chapter 5

Proofs of Asymptotic Results for Sufficient Graphical Models

This appendix is devoted to the proof of all the theoretical results in the main manuscript. We also created a Preliminaries section that gathers together some facts that will be used exclusively in this appendix. All the references to equation numbers refer to the equations created within this appendix.

5.1 Preliminaries

5.1.1 Hilbert-Schmidt norm and operator norm

Lemma 2. *If $\mathscr{H}_1$ and $\mathscr{H}_2$ are Hilbert spaces and f and g are members of $\mathscr{H}_1$ and $\mathscr{H}_2$, respectively, then $\|f \otimes g\|_{\text{HS}} = \|f\|_{\mathscr{H}_1}\|g\|_{\mathscr{H}_2}$.*

Proof. Because

$$(f \otimes g)(f \otimes g)^* f = (f \otimes g)(g \otimes f)f = f\langle g, g\rangle_{\mathscr{H}_2}\langle f, f\rangle_{\mathscr{H}_1},$$

$\|g\|^2_{\mathscr{H}_2}\|f\|^2_{\mathscr{H}_1}$ is the eigenvalue of the rank-1 operator $(f \otimes g)(f \otimes g)^*$, which is by definition the Hilbert-Schmidt norm $\|f \otimes g\|^2_{\text{HS}}$. □

Lemma 3. *If A and B are linear operators, then*

$$\|AB\|_{\text{HS}} \leq \|A\|_{\text{OP}} \|B\|_{\text{HS}}$$

PROOF. Recall that $\|AB\|_{\text{HS}}^2 = \text{tr}(B^* A^* AB)$. Because

$$A^* A \leq \lambda_{\max}(A^* A) I,$$

we have

$$\text{tr}(B^* A^* AB) \leq \lambda_{\max}(A^* A)\text{tr}(B^* B) = \|A\|_{\text{OP}}^2 \|B\|_{\text{HS}}^2. \qquad \square$$

Corollary 4. *If $A_1, \ldots, A_m$ are bounded linear operators with at least one of them, say A_i being a Hilbert-Schmidt operator, then*

$$\|A_1 \cdots A_i \cdots A_m\|_{\text{HS}} \leq \|A_1\|_{\text{OP}} \cdots \|A_i\|_{\text{HS}} \cdots \|A_m\|_{\text{OP}}.$$

Lemma 4. *If A and B are self adjoint Hilbert Schmidt operators, then*

$$\|AB\|_{\text{HS}} \leq \|A\|_{\text{HS}} \|B\|_{\text{HS}}.$$

PROOF. This follows from Lemma 3 and the fact that, for any self adjoint operator A, $\|A\|_{\text{OP}} \leq \|A\|_{\text{HS}}$. $\square$

Corollary 5. *If $A_1, \ldots, A_m$ are self adjoint Hilbert Schmidt operators, then*

$$\|A_1 \cdots A_m\|_{\text{HS}} \leq \|A_1\|_{\text{HS}} \cdots \|A_m\|_{\text{HS}}.$$

5.1.2 Sample mean of operators

The following lemma is taken from Fukumizu et al. (2007).

Lemma 5. *Suppose*

(a) U_1 and U_2 are random vectors taking values in $\Omega_{U_1} \subseteq \mathbb{R}^{p_1}$ and $\Omega_{U_2} \subseteq \mathbb{R}^{p_2}$, respectively;

(b) $\kappa_1 : \Omega_{U_1} \times \Omega_{U_1} \to \mathbb{R}$, $\kappa_2 : \Omega_{U_2} \times \Omega_{U_2} \to \mathbb{R}$ *are positive kernel functions such that* $E[\kappa_1(U_1, U_1)] < \infty$ *and* $E[\kappa_2(U_2, U_2)] < \infty$*;*

(c) $(U_{11}, U_{21}), \ldots, (U_{1n}, U_{2n})$ *are an i.i.d. sample of* (U_1, U_2)*.*

Then $\|\hat{\Sigma}_{U_1U_2} - \Sigma_{U_1U_2}\|_{\text{HS}} = O_P(n^{-1/2})$*.*

5.1.3 Tychonoff regularized inverse

Henceforth, we say that a linear operator A is a CSP operator if it is compact, self-adjoint, and positive semidefinite. Note that A being positive semidefinite implies that A is injective; that is, $\ker(A) = \{0\}$. If $A : \mathscr{H} \to \mathscr{H}$ is an injective linear operator, we define A^{-1} to be the linear operator from $\text{ran}(A)$ to $\mathscr{H}$ such that, for any $g \in \text{ran}(A)$, $A^{-1}g$ is the unique element $f \in \mathscr{H}$ such that $Af = g$. For any $\alpha > 0$, we denote the operator $(A^\alpha)^{-1}$ by $A^{-\alpha}$. The conditions for the following lemma are not the weakest possible, but they make the proof simple and they are all we will need.

Lemma 6. *Suppose* $\mathscr{H}$ *and* $\mathscr{K}$ *are Hilbert spaces and*

(a) $A_1 : \mathscr{K} \to \mathscr{K}$ *is a CSP operator;*

(b) $A_2 : \mathscr{H} \to \mathscr{K}$ *is a finite-rank linear operator;*

(c) $\alpha > 0$*, and* $\text{ran}(A_2) \subseteq \text{ran}(A_1^\alpha)$*.*

Then, for any $\eta > 0$*,* $(A_1 + \eta I)^{-\alpha} A_2$ *is a finite-rank operator with*

$$\|(A_1 + \eta I)^{-\alpha} A_2\|_{\text{HS}} \leq \|A_1^{-\alpha} A_2\|_{\text{HS}}. \tag{5.1.1}$$

Condition (c) is equivalent to $\text{ran}(A_2) \subseteq \text{dom}(A_1^{-\alpha})$, so that the operator $A_1^{-\alpha}A_2$ is a well defined finite-rank operator.

Proof. First, since $A_1^{-\alpha}A_2$ is a finite-rank operator, it is a Hilbert-Schmidt operator. Because

$$(A_1 + \eta I)^{-\alpha} A_2 = (A_1 + \eta I)^{-\alpha} A_1^\alpha A_1^{-\alpha} A_2$$

and $(A_1+\eta I)^{-\alpha}A_1^{\alpha} \le (A_1+\eta I)^{-\alpha}(A_1+\eta I)^{\alpha} = I$, we have

$$A_2^*[(A_1+\eta I)^{-\alpha}]^2 A_2 = A_2^* A_1^{-\alpha}[(A_1+\eta I)^{-\alpha}A_1^{\alpha}]^2 A_1^{-\alpha}A_2 \le A_2^*(A_1^{-\alpha})^2 A_2,$$

where the first equality holds because $(A_1+\eta I)^{-\alpha}$ and A_1^{α} commute. Hence the trace norm of the left is no greater than the trace norm of the right, which is equivalent to (5.1.4). □

Corollary 6. *Suppose*

1. $A_1 : \mathscr{K} \to \mathscr{K}$ *and* $A_3 : \mathscr{H} \to \mathscr{H}$ *are CSP operators;*
2. $A_2 : \mathscr{H} \to \mathscr{K}$ *is a finite rank linear operator;*
3. $\alpha > 0$, $\beta > 0$, $\mathrm{ran}(A_2) \subseteq \mathrm{ran}(A_1^{\alpha})$.

Then $(A_1+\eta I)^{-\alpha}A_2(A_3+\epsilon I)^{-\beta}$ *is a finite-rank operator and*

$$\|(A_1+\eta I)^{-\alpha}A_2(A_3+\epsilon I)^{-\beta}\|_{\mathrm{HS}} \le \|A_1^{-\alpha}A_2A_3^{-\beta}\|_{\mathrm{HS}}. \tag{5.1.2}$$

Proof. Again, it is obvious that $A_1^{-\alpha}A_2A_3^{-\beta}$ is a finite-rank operator, so it has a finite Hilbert-Schmidt norm. Since the conditions in Lemma 6 are satisfied for A_1 and A_2 therein replaced by A_1 and $A_2A_3^{-\beta}$ in this corollary, the operator $(A_1+\eta I)^{-\alpha}A_2A_3^{-\beta}$ is Hilbert Schmidt with

$$\|(A_1+\eta I)^{-\alpha}A_2A_3^{-\beta}\|_{\mathrm{HS}} \le \|A_1^{-\alpha}A_2A_3^{-\beta}\|_{\mathrm{HS}}. \tag{5.1.3}$$

Similarly, since the conditions in Lemma 6 are satisfied for A_1 and A_2 therein replaced by A_3 and $A_2^*(A_1+\eta I)^{-\alpha}$ in this corollary, the operator $(A_3+\epsilon I)^{-\beta}A_2^*(A_1+\eta I)^{-\alpha}$ is Hilbert Schmidt with

$$\begin{aligned}\|(A_3+\epsilon I)^{-\beta}A_2^*(A_1+\eta I)^{-\alpha}\| &\le \|A_3^{-\beta}A_2^*(A_1+\eta I)^{-\alpha}\|_{\mathrm{HS}} \\ &= \|(A_1+\eta I)^{-\alpha}A_2^*A^{-\beta}\|_{\mathrm{HS}},\end{aligned}$$

where the right-hand side, by (5.1.3), is no greater than $\|A_1^{-\alpha}A_2A^{-\beta}\|_{\mathrm{HS}}$. □

Lemma 7. *Suppose $\mathscr{H}$ and $\mathscr{K}$ are Hilbert spaces and*

(a) $A_1 : \mathscr{K} \to \mathscr{K}$ is a CSP operator;

(b) $\alpha > 0$, and $\mathrm{ran}(A_2) \subseteq \mathrm{ran}(A_1^\alpha)$; $A_2^{-\alpha} A_1$ is a bounded linear operator.

Then, for any $\eta > 0$, $(A_1 + \eta I)^{-\alpha} A_2$ is a finite-rank operator with

$$\|(A_1 + \eta I)^{-\alpha} A_2\|_{\mathrm{OP}} \leq \|A_1^{-\alpha} A_2\|_{\mathrm{OP}}. \tag{5.1.4}$$

The proof is similar to that of Lemma 6 and is omitted.

5.1.4 Negative square root

The next lemma can be verified by simple computation (see, for example, Fukumizu, Bach, and Gretton (2007)).

Lemma 8. *If A and B are self adjoint and invertible linear operators, then*

$$\begin{aligned} A^{-1/2} - B^{-1/2} &= A^{-3/2}(B^{3/2} - A^{3/2})B^{-1/2} + A^{-3/2}(A - B) \\ &= A^{-1/2}(B^{3/2} - A^{3/2})B^{-3/2} + (A - B)B^{-3/2}. \end{aligned}$$

5.1.5 Notations for order of magnitude

If $\{A_n\}$ is a sequence of random operators and $\{a_n\}$ is a sequence of positive numbers such that $\|A_n\|_{\mathrm{OP}} = O_P(a_n)$, then we write $A_n = \dot{O}_P(a_n)$. If $\|A_n\|_{\mathrm{HS}} = O_P(a_n)$, then we write $A_n = \ddot{O}_P(a_n)$. Note that $A_n = \ddot{O}_P(a_n)$ implies $A_n = \dot{O}_P(a_n)$, and $\ddot{O}_P(a_n)\dot{O}_P(b_n) = \ddot{O}_P(a_n b_n)$. Similarly, if $\|A_n\|_{\mathrm{OP}} = o_P(a_n)$, then we write $A_n = \dot{o}_P(a_n)$. If $\|A_n\|_{\mathrm{HS}} = o_P(a_n)$, then we write $A_n = \ddot{o}_P(a_n)$.

Also, as already mentioned in the main manuscript, if $\{a_n\}$ and $\{b_n\}$ are sequences of positive numbers, then we write $a_n \prec b_n$ if $a_n/b_n \to 0$. We write $a_n \asymp b_n$ if $0 < \liminf_n(b_n/a_n) \leq \limsup_n(b_n/a_n) < \infty$. We write $b_n \preceq a_n$ if either $b_n \prec a_n$ or $b_n \asymp a_n$.

5.2 Proof of Theorem 4

$1 \Rightarrow 2$. Since $\mathcal{G}^{-(i,j)} \subseteq \sigma(X^{-(i,j)})$, we have

$$X^i \perp\!\!\!\perp X^j | X^{-(i,j)} \Leftrightarrow X^i \perp\!\!\!\perp X^j | (X^{-(i,j)}, \mathcal{G}^{-(i,j)}).$$

Hence

$$\begin{aligned}
&\begin{cases} X^i \perp\!\!\!\perp X^j | X^{-(i,j)} \\ (X^i, X^j) \perp\!\!\!\perp X^{-(i,j)} | \mathcal{G}^{-(i,j)} \end{cases} \\
&\qquad \Rightarrow \begin{cases} X^i \perp\!\!\!\perp X^j | X^{-(i,j)}, \mathcal{G}^{-(i,j)} \\ (X^i, X^j) \perp\!\!\!\perp X^{-(i,j)} | \mathcal{G}^{-(i,j)} \end{cases} \\
&\qquad \Rightarrow \begin{cases} X^i \perp\!\!\!\perp X^j | X^{-(i,j)}, \mathcal{G}^{-(i,j)} \\ X^i \perp\!\!\!\perp X^{-(i,j)} | \mathcal{G}^{-(i,j)} \end{cases} \\
&\qquad \Rightarrow X^i \perp\!\!\!\perp (X^j, X^{-(i,j)}) | \mathcal{G}^{-(i,j)} \\
&\qquad \Rightarrow X^i \perp\!\!\!\perp X^j | \mathcal{G}^{-(i,j)},
\end{aligned}$$

where the first implication follows from statement 2 of Theorem 2.1 of Li (2018c); the second from statement 4; the third from statement 2 again.

$2 \Rightarrow 1$. Let $A \in \sigma(X^i)$, $B \in \sigma(X^j)$. It suffices to show that

$$P(X^i \in A, X^j \in B | X^{-(i,j)}) = P(X^i \in A | X^{-(i,j)}) P(X^j \in B | X^{-(i,j)}).$$

Because $(X^i, X^j) \perp\!\!\!\perp X^{-(i,j)} | \mathcal{G}^{-(i,j)}$, the left hand side is

$$P(X^i \in A, X^j \in B | \mathcal{G}^{-(i,j)}),$$

which, by condition 2, is equal to $P(X^i \in A | \mathcal{G}^{-(i,j)}) P(X^j \in B | \mathcal{G}^{-(i,j)})$. However, by statement 2 of Theorem 2.1 of Li (2018c), we have $X^i \perp\!\!\!\perp X^{-(i,j)} | \mathcal{G}^{-(i,j)}$ and $X^j \perp\!\!\!\perp X^{-(i,j)} | \mathcal{G}^{-(i,j)}$. Hence

$$P(X^i \in A | \mathcal{G}^{-(i,j)}) P(X^j \in B | \mathcal{G}^{-(i,j)}) = P(X^i \in A | X^{-(i,j)}) P(X^j \in B | X^{-(i,j)}),$$

as desired.

5.3 Proof of Theorem 5

(i) Because, by the reproducing property of an RKHS,

$$u_r(a) = \langle u_r,\, \kappa_0(\cdot, a)\rangle_{\mathscr{H}}, \quad v_r(a) = \langle v_r,\, \kappa_0(\cdot, a)\rangle_{\mathscr{H}}, \quad r = 1, \ldots, d,$$

we have

$$\begin{aligned}\|U(a) - V(a)\|^2_{\mathbb{R}^d} &= \sum_{r=1}^{d} \langle u_r - v_r, \kappa_0(\cdot, a)\rangle^2_{\mathscr{H}} \leq \sum_{r=1}^{d} \|u_r - v_r\|^2_{\mathscr{H}} \|\kappa_0(\cdot, a)\|^2_{\mathscr{H}} \\ &= \kappa_0(a, a) \sum_{r=1}^{d} \|u_r - v_r\|^2_{\mathscr{H}} = \kappa_0(a, a)\|U - V\|^2_{\mathscr{H}^d}.\end{aligned}$$

Now take square root on both sides to complete the proof of (a).
(ii) By the definition of the inner product in an RKHS,

$$\begin{aligned}&\|\kappa_1(\cdot, U(a)) - \kappa_1(\cdot, V(a))\|^2_{\mathscr{H}_1} \\ &= \langle \kappa_1(\cdot, U(a)) - \kappa_1(\cdot, V(a)), \kappa_1(\cdot, U(a)) - \kappa_1(\cdot, V(a))\rangle_{\mathscr{H}_1} \\ &= \kappa_1(U(a), U(a)) - 2\kappa_1(U(a), V(a)) + \kappa_1(V(a), V(a)) \\ &\leq |\kappa_1(U(a), U(a)) - \kappa_1(V(a), U(a))| \\ &\quad + |\kappa_1(U(a), V(a)) - \kappa_1(V(a), V(a))|.\end{aligned} \tag{5.3.1}$$

By Taylor's mean value theorem

$$\begin{aligned}&\kappa_1(V(a), U(a)) - \kappa_1(U(a), U(a)) \\ &= \left[\frac{\partial \kappa_1(s, U(a))}{\partial s}\right]_{s=U(a)} [V(a) - U(a)] \\ &\quad + \frac{1}{2}[V(a) - U(a)]^{\intercal} \left[\frac{\partial^2 \kappa_1(s, U(a))}{\partial s \partial s^{\intercal}}\right]_{s=\xi} [V(a) - U(a)]\end{aligned}$$

for some ξ in the line joining $U(a)$ and $V(a)$. Since, by assumption, the first derivative above is 0, and the second derivative has bounded eigenvalues, there is

a constant $C_1 > 0$ such that

$$\begin{aligned} |\kappa_1(V(a), U(a)) - \kappa_1(U(a), U(a))| &\leq C_1 \|V(a) - U(a)\|^2_{\mathbb{R}^d} \\ &\leq C_1 \|V - U\|^2_{\mathscr{H}^d}\ \kappa_0(a, a), \end{aligned} \tag{5.3.2}$$

where the second inequality follows from part (i). By similar computation, we have, for a constant $C_1 > 0$ (which can be taken as the same constant above),

$$|\kappa_1(U(a), V(a)) - \kappa_1(V(a), V(a))| \leq 2^{-1} C_1 \|V - U\|^2_{\mathscr{H}^d}\ \kappa_0(a, a). \tag{5.3.3}$$

Substitute (5.3.2) and (5.3.3) into the right-hand side of (5.3.1) to prove (ii). □

5.4 Proof of Theorem 6

(i) Tentatively abbreviating $U^{ij}(X^{-(i,j)})$ and $\hat{U}^{ij}(X^{-(i,j)})$ by U^{ij} and $\hat{U}^{ij}$, we have

$$\begin{aligned} &\|\hat{\Sigma}_{\hat{U}^{ij}\hat{U}^{ij}} - \hat{\Sigma}_{U^{ij}U^{ij}}\|_{\text{HS}} \\ &= \|E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij}) \otimes \kappa_U^{ij}(\cdot, \hat{U}^{ij})) - E_n(\kappa_U^{ij}(\cdot, U^{ij}) \otimes \kappa_U^{ij}(\cdot, U^{ij})) \\ &\quad - [E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij})) \otimes E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij})) \\ &\qquad\qquad - E_n(\kappa_U^{ij}(\cdot, U^{ij})) \otimes E_n(\kappa_U^{ij}(\cdot, U^{ij}))]\|_{\text{HS}} \\ &\leq \|E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij}) \otimes \kappa_U^{ij}(\cdot, \hat{U}^{ij})) - E_n(\kappa_U^{ij}(\cdot, U^{ij}) \otimes \kappa_U^{ij}(\cdot, U^{ij}))\|_{\text{HS}} \\ &\quad + \|[E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij})) \otimes E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij})) \\ &\qquad\qquad - E_n(\kappa_U^{ij}(\cdot, U^{ij})) \otimes E_n(\kappa_U^{ij}(\cdot, U^{ij}))]\|_{\text{HS}} \\ &\equiv \|\Delta_n^{(1)}\|_{\text{HS}} + \|\Delta_n^{(2)}\|_{\text{HS}}, \end{aligned} \tag{5.4.1}$$

where, for example, $E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij}))$ is the abbreviation of

$$n^{-1} \sum_{a=1}^{n} \kappa_U^{ij}(\cdot, \hat{U}_a^{ij})$$

We now derive the order of magnitude of $\|\Delta_n^{(1)}\|_{\text{HS}}$. Because

$$E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij}) \otimes \kappa_U^{ij}(\cdot, \hat{U}^{ij})) - E_n(\kappa_U^{ij}(\cdot, U^{ij}) \otimes \kappa_U^{ij}(\cdot, U^{ij}))$$

$$
\begin{aligned}
&= E_n[(\kappa_U^{ij}(\cdot,\hat{U}^{ij}) - \kappa_U^{ij}(\cdot,U^{ij})) \otimes (\kappa_U^{ij}(\cdot,\hat{U}^{ij}) - \kappa_U^{ij}(\cdot,U^{ij}))] \\
&\quad + E_n[(\kappa_U^{ij}(\cdot,\hat{U}^{ij}) - \kappa_U^{ij}(\cdot,U^{ij})) \otimes \kappa_U^{ij}(\cdot,U^{ij})] \\
&\quad + E_n[\kappa_U^{ij}(\cdot,U^{ij}) \otimes (\kappa_U^{ij}(\cdot,\hat{U}^{ij}) - \kappa_U^{ij}(\cdot,U^{ij}))],
\end{aligned}
$$

we have, by the triangular inequality,

$$
\begin{aligned}
\|\Delta_n^{(1)}\|_{\text{HS}} \leq E_n\|(\kappa_U^{ij}(\cdot,\hat{U}^{ij}) - \kappa_U^{ij}(\cdot,U^{ij})) \otimes (\kappa_U^{ij}(\cdot,\hat{U}^{ij}) - \kappa_U^{ij}(\cdot,U^{ij}))\|_{\text{HS}} \\
+ 2E_n\|(\kappa_U^{ij}(\cdot,\hat{U}^{ij}) - \kappa_U^{ij}(\cdot,U^{ij})) \otimes \kappa_U^{ij}(\cdot,U^{ij})\|_{\text{HS}}.
\end{aligned}
$$

By Lemma 2, the right-hand side can be rewritten as

$$
\begin{aligned}
&E_n\|\kappa_U^{ij}(\cdot,\hat{U}^{ij}) - \kappa_U^{ij}(\cdot,U^{ij})\|^2_{\mathscr{H}^{ij}(U)} \\
&\quad + 2E_n(\|\kappa_U^{ij}(\cdot,\hat{U}^{ij}) - \kappa_U^{ij}(\cdot,U^{ij})\|_{\mathscr{H}^{ij}(U)}\|\kappa_U^{ij}(\cdot,U^{ij})\|_{\mathscr{H}^{ij}(U)}).
\end{aligned} \tag{5.4.2}
$$

Now applying Theorem 5, part (ii), with

$$
\begin{aligned}
&\Omega = \Omega^{-(i,j)}, \quad \kappa_0 = \kappa_X^{-(i,j)}, \quad \mathscr{H}_0 = \mathscr{H}^{-(i,j)}(X), \\
&\mathbb{R}^d = \mathbb{R}^{d_{ij}}, \quad \kappa_1 = \kappa_U^{ij}, \quad \mathscr{H}_1 = \mathscr{H}^{ij}(U),
\end{aligned}
$$

we see that, for some $C > 0$, (5.4.2) is bounded from above by

$$
\begin{aligned}
&C^2\|\hat{U}^{ij} - U^{ij}\|^2_{[\mathscr{H}^{-(i,j)}(X)]^{d_{ij}}} \; E_n\kappa_X^{-(i,j)}(X^{-(i,j)}, X^{-(i,j)}) \\
&+ 2C\|\hat{U}^{ij} - U^{ij}\|_{[\mathscr{H}^{-(i,j)}(X)]^{d_{ij}}} \\
&\times E_n\{[\kappa_X^{-(i,j)}(X^{-(i,j)}, X^{-(i,j)})\kappa_U^{ij}(U^{ij}(X^{-(i,j)}), U^{ij}(X^{-(i,j)})]^{1/2}\}.
\end{aligned} \tag{5.4.3}
$$

By the weak law of large numbers,

$$
\begin{aligned}
&E_n\kappa_X^{-(i,j)}(X^{-(i,j)}, X^{-(i,j)}) = O_P(1) \\
&E_n\{[\kappa_X^{-(i,j)}(X^{-(i,j)}, X^{-(i,j)})\kappa_U^{ij}(U^{ij}(X^{-(i,j)}), U^{ij}(X^{-(i,j)})]^{1/2}\} = O_P(1).
\end{aligned}
$$

Hence (5.4.3) is of the order $O_P(b_n^2)O_P(1) + O_P(b_n)O_P(1) = O_P(b_n)$.

Next, we derive the order of magnitude of $\|\Delta_n^{(2)}\|_{\text{HS}}$, which can be rewritten as

$$
\|\{[E_n(\kappa_U^{ij}(\cdot,\hat{U}^{ij})) - E_n(\kappa_U^{ij}(\cdot,U^{ij})) + E_n(\kappa_U^{ij}(\cdot,U^{ij}))]
$$

$$
\begin{aligned}
&\otimes [E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij})) - E_n(\kappa_U^{ij}(\cdot, U^{ij})) + E_n(\kappa_U^{ij}(\cdot, U^{ij}))]\} \\
&\quad - E_n(\kappa_U^{ij}(\cdot, U^{ij})) \otimes E_n(\kappa_U^{ij}(\cdot, U^{ij}))\|_{\text{HS}} \\
\leq\ & \|E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij})) - E_n(\kappa_U^{ij}(\cdot, U^{ij}))\|^2_{\mathscr{H}^{ij}(U)} \\
&+ 2\|E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij})) - E_n(\kappa_U^{ij}(\cdot, U^{ij}))\|_{\mathscr{H}^{ij}(U)}\ \|E_n(\kappa_U^{ij}(\cdot, U^{ij}))\|_{\mathscr{H}^{ij}(U)} \\
\leq\ & E_n\|\kappa_U^{ij}(\cdot, \hat{U}^{ij}) - \kappa_U^{ij}(\cdot, U^{ij})\|^2_{\mathscr{H}^{ij}(U)} \\
&+ 2E_n\|\kappa_U^{ij}(\cdot, \hat{U}^{ij}) - \kappa_U^{ij}(\cdot, U^{ij})\|_{\mathscr{H}^{ij}(U)}\ E_n\|\kappa_U^{ij}(\cdot, U^{ij})\|_{\mathscr{H}^{ij}(U)}.
\end{aligned}
$$

As shown in the proof of $\|\Delta_n^{(1)}\|_{\text{HS}}$, this too is of the order
$O_P(b_n^2)O_P(1) + O_P(b_n)O_P(1) = \text{O}_P(b_n)$.
(ii) and (iii): The proofs of (ii) and (iii) are essentially the same as the proof of (i), so we just highlight the proof of (ii) and omit the proof of (iii). Similar to (5.4.1), we have

$$
\|\hat{\Sigma}_{(X^i\hat{U}^{ij})\hat{U}^{ij}} - \hat{\Sigma}_{(X^iU^{ij})U^{ij}}\|_{\text{HS}} \leq \|\Delta_n^{(1)}\|_{\text{HS}} + \|\Delta_n^{(2)}\|_{\text{HS}},
$$

where

$$
\begin{aligned}
\Delta_n^{(1)} =& E_n(\kappa_{XU}^{i,ij}(\cdot, (X^i, \hat{U}^{ij})) \otimes \kappa_U^{ij}(\cdot, \hat{U}^{ij})) - E_n(\kappa_{XU}^{i,ij}(\cdot, (X^i, U^{ij})) \otimes \kappa_U^{ij}(\cdot, U^{ij})) \\
\Delta_n^{(2)} =& E_n(\kappa_{XU}^{i,ij}(\cdot, (X^i, \hat{U}^{ij}))) \otimes E_n(\kappa_U^{ij}(\cdot, \hat{U}^{ij})) \\
&- E_n(\kappa_{XU}^{i,ij}(\cdot, (X^i, U^{ij}))) \otimes E_n(\kappa_U^{ij}(\cdot, U^{ij})).
\end{aligned}
$$

Because

$$
\begin{aligned}
\Delta_n^{(1)} =& E_n[(\kappa_U^{ij}(\cdot, (X^i, \hat{U}^{ij})) - \kappa_U^{ij}(\cdot, (X^i, U^{ij}))) \otimes (\kappa_U^{ij}(\cdot, \hat{U}^{ij}) - \kappa_U^{ij}(\cdot, U^{ij}))] \\
&+ E_n[(\kappa_U^{ij}(\cdot, (X^i, \hat{U}^{ij})) - \kappa_U^{ij}(\cdot, (X^i, U^{ij}))) \otimes \kappa_U^{ij}(\cdot, U^{ij})] \\
&+ E_n[\kappa_U^{ij}(\cdot, (X^i, U^{ij})) \otimes (\kappa_U^{ij}(\cdot, \hat{U}^{ij}) - \kappa_U^{ij}(\cdot, U^{ij}))],
\end{aligned}
$$

we have

$$
\begin{aligned}
&\|\Delta_n^{(1)}\|_{\text{HS}} \\
&\leq E_n\|(\kappa_{XU}^{i,ij}(\cdot, (X^i, \hat{U}^{ij})) - \kappa_{XU}^{i,ij}(\cdot, (X^i, U^{ij}))\|^2_{\mathscr{H}^{ij}(U)} \\
&\quad + 2E_n(\|(\kappa_{XU}^{i,ij}(\cdot, (X^i, \hat{U}^{ij})) - \kappa_{XU}^{i,ij}(\cdot, (X^i, U^{ij}))\|_{\mathscr{H}^{ij}(U)}\|\kappa_U^{ij}(\cdot, U^{ij})\|_{\mathscr{H}^{ij}(U)})
\end{aligned}
$$

$$\leq C^2\|\hat{U}^{ij} - U^{ij}\|^2_{[\mathscr{H}^{-(i,j)}(X)]^{d_{ij}}} \; E_n\kappa_X^{-(i,j)}(X^{-(i,j)}, X^{-(i,j)})$$
$$+ 2C\|\hat{U}^{ij} - U^{ij}\|_{[\mathscr{H}^{-(i,j)}(X)]^{d_{ij}}}$$
$$\times E_n\{[\kappa_X^{-(i,j)}(X^{-(i,j)}, X^{-(i,j)})\kappa_U^{ij}(U^{ij}(X^{-(i,j)}), U^{ij}(X^{-(i,j)})]^{1/2}\}.$$

The rest of the proof is the same as the corresponding part of the proof of part (i). □

5.5 Proof of Theorem 7

Denote

$$\begin{aligned}
&\hat{\Sigma}_{(X^i\hat{U}^{ij})\hat{U}^{ij}},\ (\hat{\Sigma}_{\hat{U}^{ij}\hat{U}^{ij}} + \delta_n I)^{-1},\ \hat{\Sigma}_{\hat{U}^{ij}(X^j\hat{U}^{ij})},\\
&\hat{\Sigma}_{(X^iU^{ij})U^{ij}},\ (\hat{\Sigma}_{U^{ij}U^{ij}} + \delta_n I)^{-1},\ \hat{\Sigma}_{U^{ij}(X^jU^{ij})},\\
&\Sigma_{(X^iU^{ij})U^{ij}},\ \Sigma^{-1}_{U^{ij}U^{ij}},\ \Sigma_{U^{ij}(X^jU^{ij})}
\end{aligned} \tag{5.5.1}$$

by $\hat{A},\ \hat{B},\ \hat{C},\ \tilde{A},\ \tilde{B},\ \tilde{C},\ A,\ B,\ C$, respectively. Then, by the definition of CCCO and the triangular inequality,

$$\begin{aligned}
&\|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|\hat{U}^{ij}} - \hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\text{HS}}\\
&\leq \|\hat{\Sigma}_{(X^i\hat{U}^{ij})(X^j\hat{U}^{ij})} - \hat{\Sigma}_{(X^iU^{ij})(X^jU^{ij})}\|_{\text{HS}} + \|\hat{A}\hat{B}\hat{C} - \tilde{A}\tilde{B}\tilde{C}\|_{\text{HS}}
\end{aligned} \tag{5.5.2}$$

By Theorem 6, the first term is $O_P(b_n)$. The second term (without the norm) is

$$\begin{aligned}
\hat{A}\hat{B}\hat{C} - \tilde{A}\tilde{B}\tilde{C} &= (\hat{A} - \tilde{A})\hat{B}\hat{C} + \tilde{A}(\hat{B} - \tilde{B})\hat{C} + \tilde{A}\tilde{B}(\hat{C} - \tilde{C})\\
&= (\hat{A} - \tilde{A})\hat{B}\hat{C} + \tilde{A}\hat{B}(\tilde{B}^{-1} - \hat{B}^{-1})\tilde{B}\hat{C} + \tilde{A}\tilde{B}(\hat{C} - \tilde{C}).
\end{aligned} \tag{5.5.3}$$

Since, by Theorem 6,

$$\hat{A} - \tilde{A} = \ddot{O}_P(b_n), \quad \tilde{B}^{-1} - \hat{B}^{-1} = \ddot{O}_P(b_n), \quad \hat{C} - \tilde{C} = \ddot{O}_P(b_n),$$

in order for (5.5.3) to hold it suffices to show that

$$\hat{B}\hat{C} = \dot{O}_P(1), \quad \tilde{A}\hat{B} = \dot{O}_P(1), \quad \tilde{B}\hat{C} = \dot{O}_P(1), \quad \tilde{A}\tilde{B} = \dot{O}_P(1). \tag{5.5.4}$$

To simplify the notation, let

$$\check{B} = (\Sigma_{U^{ij}U^{ij}} + \delta_n I)^{-1}, \quad \hat{D} = \hat{\Sigma}_{\hat{U}^{ij}\hat{U}^{ij}}, \quad \tilde{D} = \hat{\Sigma}_{U^{ij}U^{ij}}, \quad D = \Sigma_{\hat{U}^{ij}\hat{U}^{ij}}.$$

For the first relation (5.5.4), by Theorem 6, $\hat{C} - \tilde{C} = \ddot{O}_P(b_n)$; by Lemma 5, $\tilde{C} - C = \ddot{O}_P(n^{-1/2})$. Hence

$$\begin{aligned}\hat{B}\hat{C} =& \hat{B}(\hat{C} - \tilde{C}) + \hat{B}(\tilde{C} - C) + (\hat{B} - \tilde{B})C + (\tilde{B} - \check{B})C \\ &+ (\check{B} - B)C + BC \\ =& \ddot{O}_P(\delta_n^{-1} b_n) + \ddot{O}_P(\delta_n^{-1} n^{-1/2}) + (\hat{B} - \tilde{B})C + (\tilde{B} - \check{B})C + \check{B}C.\end{aligned} \tag{5.5.5}$$

The third term on the right is

$$\begin{aligned}(\hat{B} - \tilde{B})C =& \hat{B}\ddot{O}_P(b_n)\tilde{B}C \\ =& \ddot{O}_P(\delta_n^{-1} b_n)(\tilde{B} - \check{B})C + \ddot{O}_P(\delta_n^{-1} b_n)\check{B}C \\ =& \ddot{O}_P(\delta_n^{-1} b_n)\tilde{B}\ddot{O}_P(n^{-1/2})\check{B}C + \ddot{O}_P(\delta_n^{-1} b_n)\check{B}C \\ =& \ddot{O}_P(\delta_n^{-1} b_n)\ddot{O}_P(\delta_n^{-1} n^{-1/2})\dot{O}_P(1) + \ddot{O}_P(\delta_n^{-1} b_n)\dot{O}_P(1).\end{aligned}$$

So, by condition (b) and the fact that $n^{-1/2} \prec b_n$, this term is $\ddot{o}_P(1)$. The fourth term on the right-hand side of (5.5.5) is

$$(\tilde{B} - \check{B})C = \tilde{B}\ddot{O}_P(n^{-1/2})\check{B}C = \ddot{O}_P(\delta_n^{-1} b_n) = \ddot{o}_P(1).$$

By Lemma 7, $\check{B}C = \dot{O}_P(1)$. Hence the first relation in (5.5.4) holds. For later use, note that in this process we also proved

$$\tilde{B}C = \dot{O}_P(1), \text{ (and hence also) } A\tilde{B} = \dot{O}_P(1). \tag{5.5.6}$$

For the second relation in (5.5.4):

$$\tilde{A}\hat{B} = (\tilde{A} - A)\hat{B} + A(\hat{B} - \tilde{B}) + A\tilde{B}. \tag{5.5.7}$$

The first term is of the order $\ddot{O}_P(n^{-1/2}\delta_n^{-1})$. The second term is

$$A\tilde{B}\ddot{O}_P(b_n)\hat{B} = A\tilde{B}\ddot{O}_P(b_n)\dot{O}_P(\delta_n^{-1}) = \ddot{O}_P(b_n\delta_n^{-1}) = \ddot{O}_P(1),$$

where the second equality follows from (5.5.6), and the third from condition (b). The third term, again by (5.5.6), is of the order $\dot{O}_P(1)$. Thus the second relation in (5.5.4) holds.

For the third relation in (5.5.4):

$$\tilde{B}\hat{C} = \tilde{B}(\hat{C} - \tilde{C}) + \tilde{B}\tilde{C} = \ddot{O}_P(\delta_n^{-1} b_n) + \tilde{B}\tilde{C}. \tag{5.5.8}$$

Using an argument similar to the next step, we can show that

$$\tilde{B}\tilde{C} = \dot{O}_P(1). \tag{5.5.9}$$

Hence the third relation in (5.5.4) holds.

For the fourth relation in (5.5.4):

$$\tilde{A}\tilde{B} = (\tilde{A} - C)\tilde{B} + A\tilde{B} = \dot{O}_P(\delta_n^{-1})\dot{O}_P(\delta_n + n^{-1/2}) + A\tilde{B}.$$

By (5.5.6), the last term is of the order $\dot{O}_P(1)$. Thus the fourth relation in (5.5.4) holds.

5.6 Proof of Theorem 8

Lemma 9. *Suppose*

(a) the conditions in Lemma 6 are satisfied for $\alpha = 3/2$*;*

(b) $\|\hat{A}_1 - A_1\|_{\text{HS}} = O_P(n^{-1/2})$, $\|\hat{A}_2 - A_2\|_{\text{HS}} = O_P(n^{-1/2})$*;*

(c) $n^{-1} \prec \eta_n \prec 1$, $n^{-1/2} \prec \epsilon_n \prec 1$.

Then

$$\begin{aligned}
&\|(\hat{A}_1 + \eta_n I)^{-1/2} A_2 (\hat{A}_3 + \epsilon_n I)^{-1}\|_{\text{HS}} = O_P(1), \\
&\|A_1^{-1/2} A_2 (\hat{A}_3 + \epsilon_n I)^{-1}\|_{\text{HS}} = O_P(1), \\
&\|[(\hat{A}_1 + \eta_n I)^{-1/2} - A_1^{-1/2}] A_2\|_{\text{HS}} = O_P(\eta_n^{-1/2} n^{-1/2} + \eta_n^{1/2}), \\
&\|A_1^{-1/2} A_2 [(\hat{A}_3 + \epsilon_n I)^{-1} - A_3^{-1}] A_2^* A_1^{-1/2}\|_{\text{HS}} = O_P(n^{-1/2} + \epsilon_n).
\end{aligned} \tag{5.6.1}$$

PROOF. Let

$$\begin{aligned} B_1 &= (\hat{A}_1 + \eta_n I)^{-1/2} - (A_1 + \eta_n I)^{-1/2}, \\ B_2 &= (A_1 + \eta_n I)^{-1/2} - A_1^{-1/2}, \\ B_3 &= A_1^{-1/2}, \\ C_1 &= (\hat{A}_3 + \epsilon_n I)^{-1} - (A_3 + \epsilon_n I)^{-1}, \\ C_2 &= (A_3 + \epsilon_n I)^{-1} - A_3^{-1}, \\ C_3 &= A_3^{-1}. \end{aligned} \tag{5.6.2}$$

Then we can reexpress

$$\begin{aligned} (\hat{A}_1 + \eta_n I)^{-1/2} A_2 (\hat{A}_3 + \epsilon_n I)^{-1} &= (B_1 + B_2 + B_3) A_2 (C_1 + C_2 + C_3) \\ &= \sum_{i=1}^{3} \sum_{j=1}^{3} B_i A_2 C_j. \end{aligned} \tag{5.6.3}$$

We now analyze the nine terms in (5.6.3). By Lemma 8,

$$\begin{aligned} B_1 &= \{(\hat{A}_1 + \eta_n I)^{-1/2} [(A_1 + \eta_n I)^{3/2} - (\hat{A}_1 + \eta_n I)^{3/2}] + (\hat{A}_1 - A_1)\} \\ &\quad \times (A_1 + \eta_n I)^{-3/2} \\ &= \dot{O}_P(\eta_n^{-1/2} n^{-1/2} + n^{-1/2})(A_1 + \eta_n I)^{-3/2} \\ &= \dot{O}_P(\eta_n^{-1/2} n^{-1/2})(A_1 + \eta_n I)^{-3/2}. \end{aligned}$$

Similarly,

$$\begin{aligned} B_2 &= \{(A_1 + \eta_n I)^{-1/2} [A_1^{3/2} - (A_1 + \eta_n I)^{3/2}] + \eta_n I\} A_1^{-3/2} \\ &= \dot{O}_P(\eta_n^{-1/2} \eta_n + \eta_n) A_1^{-3/2} = \dot{O}_P(\eta_n^{1/2}) A_1^{-3/2}. \end{aligned}$$

The terms C_1 and C_2 are

$$\begin{aligned} C_1 &= (A_3 + \epsilon_n I)^{-1} (A_3 - \hat{A}_3)(\hat{A}_3 + \epsilon_n I)^{-1} = (A_3 + \epsilon_n I)^{-1} \dot{O}_P(n^{-1/2} \epsilon_n^{-1}) \\ C_2 &= A_3^{-1} (A_3 + \epsilon_n I)^{-1} (-\epsilon_n I) = A_3^{-1} (A_3 + \epsilon_n I)^{-1} \dot{O}_P(\epsilon_n) \end{aligned}$$

Hence

$$\begin{aligned}
&\textstyle\sum_{i=1}^{2}\sum_{j=1}^{2} B_i A_2 C_j \\
&= \dot{O}_P(\eta_n^{-1/2} n^{-1/2})(A_1 + \eta_n I)^{-3/2} A_2 (A_3 + \epsilon_n I)^{-1} \dot{O}_P(n^{-1/2}\epsilon_n^{-1}) \\
&+ \dot{O}_P(\eta_n^{-1/2} n^{-1/2})(A_1 + \eta_n I)^{-3/2} A_2 A_3^{-1} (A_3 + \epsilon_n I)^{-1} \dot{O}_P(\epsilon_n) \\
&+ \dot{O}_P(\eta_n^{1/2}) A_1^{-3/2} A_2 (A_3 + \epsilon_n I)^{-1} \dot{O}_P(n^{-1/2}\epsilon_n^{-1}) \\
&+ \dot{O}_P(\eta_n^{1/2}) A_1^{-3/2} A_2 A_3^{-1} (A_3 + \epsilon_n I)^{-1} \dot{O}_P(\epsilon_n).
\end{aligned} \tag{5.6.4}$$

Because

$$A_2, \quad A_2 A_3^{-1}, \quad A_1^{-3/2} A_2, \quad A_1^{-3/2} A_2 A_3^{-1}$$

are finite-rank operators, by Lemma 6 and Corollary 6, the four operators in the middle of the four terms in (5.6.5) all have finite Hilbert-Schmidt norms which do not depend on n. Thus

$$\begin{aligned}
&\textstyle\sum_{i=1}^{2}\sum_{j=1}^{2} B_i A_2 C_j \\
&= \ddot{O}_P(\eta_n^{-1/2}\epsilon_n^{-1} n^{-1} + \eta_n^{-1/2} n^{-1/2}\epsilon_n + \eta_n^{1/2} n^{-1/2}\epsilon_n^{-1} + \eta_n^{1/2}\epsilon_n) = \ddot{o}_P(1),
\end{aligned} \tag{5.6.5}$$

where the last equality follows from condition (c). Let R be the indices of the rest of the terms except the last term: $R = \{(1,3),(2,3),(3,1),(3,2)\}$. Then

$$\textstyle\sum_{(i,j)\in R} B_i A_2 C_j = \ddot{O}_P(\eta_n^{-1/2} n^{-1/2} + \eta_n^{1/2} + n^{-1/2}\epsilon_n^{-1} + \epsilon_n) = \ddot{o}_P(1), \tag{5.6.6}$$

where the last equality follows from condition (c). Thus we have

$$(\hat{A}_1 + \eta_n I)^{-1/2} A_2 (\hat{A}_3 + \epsilon_n I)^{-1} = B_3 A_2 C_3 + o_P(1) = A_1^{-1/2} A_2 A_3^{-1} + \ddot{o}_P(1),$$

which implies the first relation in (5.6.1). For the second relation in (5.6.1), we have, by (5.6.6),

$$\begin{aligned}
\|A_1^{-1/2} A_2 (\hat{A}_3 + \epsilon_n I)^{-1}\|_{\text{HS}} &\le \|B_3 A_2 C_1\|_{\text{HS}} + \|B_3 A_2 C_2\|_{\text{HS}} + \|A_1^{-1/2} A_2 A_3^{-1}\|_{\text{HS}} \\
&= \|A_1^{-1/2} A_2 A_3^{-1}\|_{\text{HS}} + O_P(n^{-1/2}\epsilon_n^{-1} + \epsilon_n) = O_P(1).
\end{aligned}$$

For the third relation in (5.6.1), we have

$$\begin{aligned}
&[(\hat{A}_1 + \eta_n I)^{-1/2} - A_1^{-1/2}]A_2 \\
&= B_1 A_2 + B_2 A_2 \\
&= \dot{O}_P(\eta_n^{-1/2} n^{-1/2})(A_1 + \eta_n I)^{-3/2} A_2 + \dot{O}_P(\eta_n^{1/2}) A_1^{-3/2} A_2.
\end{aligned}$$

By Lemma 6, $\|(A_1 + \eta_n I)^{-3/2} A_2\|_{\text{HS}} \leq \|A_1^{-3/2} A_2\|_{\text{HS}}$. Hence

$$[(\hat{A}_1 + \eta_n I)^{-1/2} - A_1^{-1/2}]A_2 = \ddot{O}_P(\eta_n^{-1/2} n^{-1/2} + \eta_n^{1/2}).$$

For the last relation in (5.6.1), we have

$$\begin{aligned}
&A_1^{-1/2} A_2 [(\hat{A}_3 + \epsilon_n I)^{-1} - A_3^{-1}] A_2^* A_1^{-1/2} \\
&= A_1^{-1/2} A_2 A_3^{-1} [A_3 - \hat{A}_3 - \epsilon_n I)](\hat{A}_3 + \epsilon_n I)^{-1} A_2^* A_1^{-1/2} \\
&= A_1^{-1/2} A_2 A_3^{-1} \dot{O}_P(n^{-1/2} + \epsilon_n)(\hat{A}_3 + \epsilon_n I)^{-1} A_2^* A_1^{-1/2}.
\end{aligned} \tag{5.6.7}$$

By the second relation in (5.6.1), $(\hat{A}_3 + \epsilon_n I)^{-1} A_2^* A_1^{-1/2} = \ddot{O}_P(1)$. Thus the last relation in (5.6.1) holds. □

Lemma 10. *Suppose*

(a) the conditions in Lemma 6 are satisfied for $\alpha = 1$*;*

(b) $\|\hat{A}_1 - A_1\|_{\text{OP}} = O_P(n^{-1/2})$, $\|\hat{A}_2 - A_2\|_{\text{OP}} = O_P(n^{-1/2})$*;*

(c) $n^{-1/2} \prec \eta_n \prec 1$, $n^{-1/2} \prec \epsilon_n \prec 1$.

Then

$$\begin{aligned}
&\|(\hat{A}_1 + \eta_n I)^{-1} A_2 (\hat{A}_3 + \epsilon_n I)^{-1}\|_{\text{HS}} = O_P(1), \\
&\|A_1^{-1} A_2 (\hat{A}_3 + \epsilon_n I)^{-1}\|_{\text{HS}} = O_P(1), \\
&\|[(\hat{A}_1 + \eta_n I)^{-1} - A_1^{-1}] A_2\|_{\text{HS}} = O_P(\eta_n^{-1} n^{-1/2} + \eta_n), \\
&\|A_1^{-1} A_2 [(\hat{A}_3 + \epsilon_n I)^{-1} - A_3^{-1}] A_2^* A_1^{-1/2}\|_{\text{HS}} = O_P(n^{-1/2} + \epsilon_n).
\end{aligned} \tag{5.6.8}$$

PROOF. Reset B_1, B_2, and B_3 to

$$B_1 = (\hat{A}_1 + \eta_n I)^{-1} - (A_1 + \eta_n I)^{-1}, \quad B_2 = (A_1 + \eta_n I)^{-1} - A_1^{-1}, \quad B_3 = A_1^{-1},$$

and keep C_1, C_2, C_3 the same as before. Then

$$B_1 = \dot{O}_P(n^{-1/2}\eta_n^{-1})(A_1 + \eta_n I)^{-1}, \quad B_2 = \dot{O}_P(\eta_n)(A_1 + \eta_n I)^{-1}A_1^{-1}.$$

Hence

$$\begin{aligned}
&(\hat{A}_1 + \eta_n I)^{-1} A_2 (\hat{A}_3 + \epsilon_n I)^{-1} \\
&= \textstyle\sum_{i=1}^{3}\sum_{j=1}^{3} B_i A_2 C_j \\
&= \ddot{O}_P(n^{-1/2}\eta_n^{-1}n^{-1/2}\epsilon_n^{-1} + n^{-1/2}\eta_n^{-1}\epsilon_n + n^{-1/2}\eta_n^{-1} \\
&\qquad + \eta_n n^{-1/2}\epsilon_n^{-1} + \eta_n\epsilon_n + \eta_n + n^{-1/2}\epsilon_n^{-1} + \epsilon_n) \\
&= \ddot{O}_P(n^{-1/2}\eta_n^{-1} + \eta_n + n^{-1/2}\epsilon_n^{-1} + \epsilon_n) \\
&= A_1^{-1}A_2A_3 + \ddot{o}_P(1),
\end{aligned}$$

where the last equality follows from condition (c). Hence the first relation in (5.6.8) holds. The second relation in (5.6.8) holds because

$$\begin{aligned}
A_1^{-1}A_2(\hat{A}_3 + \epsilon_n I)^{-1} &= A_1^{-1}A_2C_1 + A_1^{-1}A_2C_2 + A_1^{-1}A_2C_3 \\
&= \ddot{O}_P(n^{-1/2}\epsilon_n^{-1} + \epsilon_n) + A_1^{-1}A_2A_3^{-1} = A_1^{-1}A_2A_3^{-1} + \ddot{o}_P(1).
\end{aligned}$$

The third relation in (5.6.8) holds because

$$[(\hat{A}_1 + \eta_n I)^{-1} - A_1^{-1}]A_2 = B_1A_2 + B_2A_2 = \ddot{O}_P(\eta_n^{-1}n^{-1/2} + \eta_n).$$

The proof of the fourth relation in (5.6.8) is similar to the derivation in (5.6.7). □

PROOF OF THEOREM 8. Note that, because $\mathrm{ran}(\Sigma_{XX}^{3/2}) \subseteq \mathrm{ran}(\Sigma_{XX})$, condition (b) implies that $\mathrm{ran}(\Sigma_{XY}) \subseteq \mathrm{ran}(\Sigma_{XX}^{\alpha})$ is satisfied for both $\alpha = 1$ and $\alpha = 3/2$. Also, condition (d) is made to simplify the proof; it can be relaxed with a lengthier proof.

Denote the operators

$$(\hat{\Sigma}_{X^{(i,j)}X^{(i,j)}} + \epsilon_n I)^{-1}, \quad \hat{\Sigma}_{X^{-(i,j)}X^{(i,j)}}, \quad (\hat{\Sigma}_{X^{-(i,j)}X^{-(i,j)}} + \eta_n I)^{-1/2}$$
$$\Sigma^{-1}_{X^{(i,j)}X^{(i,j)}}, \quad \Sigma_{X^{-(i,j)}X^{(i,j)}}, \quad \Sigma^{-1/2}_{X^{-(i,j)}X^{-(i,j)}}$$

by $\hat{A}$, $\hat{B}$, $\hat{C}$, A, B, C, respectively. In this notation, $\hat{f}_1^{ij}, \ldots, \hat{f}_{d_{ij}}^{ij}$ are the first d_{ij} eigenfunctions of the generalized eigenvalue problem

$$\begin{aligned} \text{maximize} \quad & \langle f, \hat{B}\hat{A}\hat{B}^* f \rangle_{-(i,j)} \\ \text{subject to} \quad & \langle f, \hat{C}^{-2} f \rangle_{-(i,j)} = 1, \quad \langle f, \hat{C}^{-2} f_r \rangle_{-(i,j)} = 0, \quad i = 1, \ldots, k-1. \end{aligned}$$

This means $\hat{f}_r^{ij} = \hat{C}\hat{\phi}_r^{ij}$, where $\hat{\phi}_r^{ij}$ is the rth eigenfunction of $\hat{C}\hat{B}\hat{A}\hat{B}^*\hat{C}$. We first derive the order of magnitude of the operator

$$\hat{C}\hat{B}\hat{A}\hat{B}^*\hat{C} - CBAB^*C \tag{5.6.9}$$

in terms of the Hilbert Schmidt norm (another route is to derive this in terms of the operator norm, which is also sufficient for our purpose). By simple calculation,

$$\begin{aligned} \hat{C}\hat{B}\hat{A}\hat{B}^*\hat{C} - CBAB^*C =& \hat{C}(\hat{B}-B)\hat{A}(\hat{B}-B)^*\hat{C} + \hat{C}(\hat{B}-B)\hat{A}B^*\hat{C} \\ & + \hat{C}B\hat{A}(\hat{B}-B)^*\hat{C} + \hat{C}B\hat{A}B^*\hat{C} - CBAB^*C. \end{aligned} \tag{5.6.10}$$

The reason for choosing this particular form of decomposition is to expose the finite-rank operator B, so that, for example, when combined with the operator C (an unbounded operator), BC is still a finite-rank operator. The Hilbert-Schmidt norm of the first term on the right is

$$\|\hat{C}(\hat{B}-B)\hat{A}(\hat{B}-B)^*\hat{C}\|_{\text{HS}} \le \|\hat{C}\|_{\text{OP}}^2 \|\hat{B}-B\|_{\text{HS}}^2 \|\hat{A}\|_{\text{OP}} = O_P(\eta_n^{-1}\epsilon_n^{-1}n^{-1}). \tag{5.6.11}$$

For the second term in (5.6.10), by Lemma 9, $\|\hat{C}B\hat{A}\|_{\text{HS}} = \|\hat{A}B^*\hat{C}\|_{\text{HS}} = O_P(1)$. Hence, by Lemmas 3 and 4,

$$\begin{aligned} \|\hat{C}(\hat{B}-B)\hat{A}B^*\hat{C}\|_{\text{HS}} =& \|\hat{C}B\hat{A}(\hat{B}-B)^*\hat{C}\|_{\text{HS}} \\ & \le \|\hat{C}\|_{\text{OP}} \, \|\hat{B}-B\|_{\text{HS}} \, \|\hat{A}B^*\hat{C}\|_{\text{HS}} = O_P(\eta_n^{-1/2}n^{-1/2}). \end{aligned} \tag{5.6.12}$$

The third term in (5.6.10) is the adjoint operator of the second term, so it has the same norm. The Hilbert-Schmidt norm of the last two terms in (5.6.10) is

$$\begin{aligned}
&\|\hat{C}B\hat{A}B^*\hat{C} - CBAB^*C\|_{\text{HS}} \\
&\leq \|\hat{C}B\hat{A}B^*(\hat{C}-C)\|_{\text{HS}} + \|(\hat{C}-C)B\hat{A}B^*C\|_{\text{HS}} \\
&\quad + \|CB(\hat{A}-A)B^*C\|_{\text{HS}} \\
&\leq \|\hat{C}B\hat{A}\|_{\text{HS}}\,\|B^*(\hat{C}-C)\|_{\text{HS}} + \|(\hat{C}-C)B\|_{\text{HS}}\,\|\hat{A}B^*C\|_{\text{HS}} \\
&\quad + \|CB(\hat{A}-A)B^*C\|_{\text{HS}}
\end{aligned} \tag{5.6.13}$$

By the first relation in (5.6.1), $\|\hat{C}B\hat{A}\|_{\text{HS}} = O_P(1)$; by the second, $\|\hat{A}B^*C\|_{\text{HS}} = O_P(1)$; by the third, $\|(\hat{C}-C)B\|_{\text{HS}} = O_P(\eta_n^{-1/2}n^{-1/2}+\eta_n^{1/2})$; by the fourth, $\|CB(\hat{A}-A)B^*C\|_{\text{HS}} = O_P(n^{-1/2}+\epsilon_n)$. Therefore,

$$\begin{aligned}
\|\hat{C}B\hat{A}B^*\hat{C} - CBAB^*C\|_{\text{HS}} &= O_P(\eta_n^{-1/2}n^{-1/2} + \eta_n^{1/2} + n^{-1/2} + \epsilon_n) \\
&= O_P(\eta_n^{-1/2}n^{-1/2} + \eta_n^{1/2} + \epsilon_n).
\end{aligned} \tag{5.6.14}$$

Combining (5.6.11), (5.6.12), and (5.6.14), we have

$$\|\hat{C}\hat{B}\hat{A}\hat{B}^*\hat{C} - CBAB^*C\|_{\text{HS}} = O_P(\eta_n^{-1}\epsilon_n^{-1}n^{-1} + \eta_n^{-1/2}n^{-1/2} + \eta_n^{1/2} + \epsilon_n).$$

Next, recall that $(\hat{\lambda}_1^{ij}, \hat{\phi}_1^{ij}), \ldots, (\hat{\lambda}_{d_{ij}}^{ij}, \hat{\phi}_{d_{ij}}^{ij})$ are the first d_{ij} pairs of eigenvalue and eigenfunction of $\hat{C}\hat{B}\hat{A}\hat{B}^*\hat{C}$, and let $(\lambda_1^{ij}, \phi_1^{ij}), \ldots, (\lambda_{d_{ij}}^{ij}, \phi_{d_{ij}}^{ij})$ be the first d_{ij} eigenvalue-eigenfunction pairs of $CBAB^*C$. By perturbation theory of linear operators, $|\hat{\lambda}_r^{ij} - \lambda_r^{ij}|$ is of the same order of magnitude as $\|\hat{C}\hat{B}\hat{A}\hat{B}^*\hat{C} - CBAB^*C\|_{\text{HS}}$, and, if condition (d) holds, then

$$\|\hat{\phi}_r^{ij} - \phi_r^{ij}\|_{\mathscr{H}^{(i,j)}(X)}$$

also has the same order of magnitude. That is, for each $r = 1, \ldots, d_{ij}$,

$$\begin{aligned}
\hat{\lambda}_r^{ij} - \lambda_r^{ij} &= O_P(\eta_n^{-1}\epsilon_n^{-1}n^{-1} + \eta_n^{-1/2}n^{-1/2} + \eta_n^{1/2} + \epsilon_n) \\
\|\hat{\phi}_r^{ij} - \phi_r^{ij}\|_{\mathscr{H}^{(i,j)}(X)} &= O_P(\eta_n^{-1}\epsilon_n^{-1}n^{-1} + \eta_n^{-1/2}n^{-1/2} + \eta_n^{1/2} + \epsilon_n).
\end{aligned} \tag{5.6.15}$$

By construction,

$$\hat{f}_r^{ij} = \hat{C}\hat{\phi}_r = \hat{\lambda}_r^{ij}\hat{C}^2\hat{B}\hat{A}\hat{B}^*\hat{C}\hat{\phi}_r^{ij}. \tag{5.6.16}$$

We now derive the order of magnitude of

$$\|\hat{C}^2\hat{B}\hat{A}\hat{B}^*\hat{C} - C^2BAB^*C\|_{\text{HS}}.$$

Similar to (5.6.10),

$$\begin{aligned}
&\hat{C}^2\hat{B}\hat{A}\hat{B}^*\hat{C} - C^2BAB^*C \\
&= \hat{C}^2(\hat{B}-B)\hat{A}(\hat{B}-B)^*\hat{C} + \hat{C}^2(\hat{B}-B)\hat{A}B^*\hat{C} \\
&\quad + \hat{C}^2B\hat{A}(\hat{B}-B)^*\hat{C} + \hat{C}^2B\hat{A}B^*\hat{C} - CBAB^*C
\end{aligned}$$

where the first term on the right, similar to (5.6.11), is of the order

$$\|\hat{C}^2(\hat{B}-B)\hat{A}(\hat{B}-B)^*\hat{C}\|_{\text{HS}} = O_P(\eta_n^{-3/2}\epsilon_n^{-1}n^{-1}). \tag{5.6.17}$$

By the first relation in (5.6.1), $\|\hat{A}B^*\hat{C}\|_{\text{HS}} = O_P(1)$, and hence

$$\begin{aligned}
\|\hat{C}^2(\hat{B}-B)\hat{A}B^*\hat{C}\|_{\text{HS}} &\le \|\hat{C}^2\|_{\text{OP}}\,\|\hat{B}-B\|_{\text{HS}}\,\|\hat{A}B^*\hat{C}\|_{\text{HS}} \\
&= O_P(\eta_n^{-1}n^{-1/2}).
\end{aligned} \tag{5.6.18}$$

By the first relation in (5.6.8), $\|\hat{C}^2B^*\hat{A}\|_{\text{HS}} = O_P(1)$, and hence

$$\begin{aligned}
\|\hat{C}^2B\hat{A}(\hat{B}-B)^*\hat{C}\|_{\text{HS}} &\le \|\hat{C}^2B\hat{A}\|_{\text{HS}}\,\|\hat{B}-B\|_{\text{HS}}\|\hat{C}\|_{\text{OP}} \\
&= O_P(n^{-1/2}\eta_n^{-1/2}).
\end{aligned} \tag{5.6.19}$$

Similar to (5.6.13), we have

$$\begin{aligned}
&\|\hat{C}^2B\hat{A}B^*\hat{C} - C^2BAB^*C\|_{\text{HS}} \\
&\le \|\hat{C}^2B\hat{A}\|_{\text{HS}}\,\|B^*(\hat{C}-C)\|_{\text{HS}} \\
&\quad + \|(\hat{C}^2-C^2)B\|_{\text{HS}}\,\|\hat{A}B^*C\|_{\text{HS}} + \|C^2B(\hat{A}-A)B^*C\|_{\text{HS}}
\end{aligned}$$

Applying Lemma 9 and Lemma 10 to the right-hand side above, we obtain

$$\begin{aligned}
&\|\hat{C}^2B\hat{A}B^*\hat{C} - C^2BAB^*C\|_{\text{HS}} \\
&= O_P(1)O_P(\eta_n^{-1/2}n^{-1/2} + \eta_n^{1/2}) + O_P(\eta_n^{-1}n^{-1/2} + \eta_n) \\
&\quad + O_P(n^{-1/2} + \epsilon_n) \\
&= O_P(\eta_n^{-1}n^{-1/2} + \eta_n^{1/2} + \epsilon_n)
\end{aligned} \tag{5.6.20}$$

Combining (5.6.17) through (5.6.20), we have

$$\begin{aligned}
&\hat{C}^2\hat{B}\hat{A}\hat{B}^*\hat{C} - C^2BAB^*C \\
&= \ddot{O}_P(\eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n^{-1/2}n^{-1/2} + \eta_n^{-1}n^{-1/2} + \eta_n^{1/2} + \epsilon_n) \\
&= \ddot{O}_P(\eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n^{1/2} + \epsilon_n).
\end{aligned} \tag{5.6.21}$$

Finally, let us derive the convergence rate of $\hat{U}^{ij}$. By (5.6.16), we have

$$\begin{aligned}
\hat{f}_r^{ij} - f_r^{ij} \; &= (\hat{\lambda}_r^{ij} - \lambda_r^{ij})\hat{C}^2\hat{B}\hat{A}\hat{B}^*\hat{C}\hat{\phi}_r^{ij} \\
&\quad + \lambda_r^{ij}(\hat{C}^2\hat{B}\hat{A}\hat{B}^*\hat{C} - C^2BAB^*C)\hat{\phi}_r^{ij} + \lambda_r^{ij}C^2BAB^*C(\hat{\phi}_r^{ij} - \phi_r^{ij}).
\end{aligned}$$

Hence

$$\begin{aligned}
\|\hat{f}_r^{ij} - f_r^{ij}\|_{\mathscr{H}^{-(i,j)}(X)} \le\; &|\hat{\lambda}_r^{ij} - \lambda_r^{ij}|\,\|\hat{C}^2\hat{B}\hat{A}\hat{B}^*\hat{C}\|_{\text{OP}}\,\|\hat{\phi}_r^{ij}\|_{\mathscr{H}^{-(i,j)}(X)} \\
&+ \lambda_r^{ij}\,\|\hat{C}^2\hat{B}\hat{A}\hat{B}^*\hat{C} - C^2BAB^*C\|_{\text{OP}}\,\|\hat{\phi}_r^{ij}\|_{\mathscr{H}^{-(i,j)}(X)} \\
&+ \lambda_r^{ij}\|C^2BAB^*C\|_{\text{OP}}\,\|\hat{\phi}_r^{ij} - \phi_r^{ij}\|_{\mathscr{H}^{-(i,j)}(X)}.
\end{aligned}$$

By (5.6.15) and (5.6.21), the right-hand side is of the order

$$\begin{aligned}
&O_P(\eta_n^{-1/2}n^{-1/2} + \eta_n^{1/2} + \epsilon_n) \\
&+ O_P(\eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n^{1/2} + \epsilon_n) \\
&+ O_P(\eta_n^{-1/2}n^{-1/2} + \eta_n^{1/2} + \epsilon_n) \\
&= O_P(\eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n^{1/2} + \epsilon_n)
\end{aligned} \tag{5.6.22}$$

Because

$$\|\hat{U}_{ij} - U_{ij}\|_{[\mathscr{H}^{-(i,j)}(X)]^{d_{ij}}} = (\textstyle\sum_{r=1}^{d_{ij}}\|\hat{f}_r^{ij} - f_r^{ij}\|_{\mathscr{H}^{-(i,j)}(X)})^{1/2},$$

$\|\hat{U}_{ij} - U_{ij}\|_{[\mathscr{H}^{-(i,j)}(X)]^{d_{ij}}}$ has the same order of magnitude as (5.6.22). □

5.7 Proof of Theorem 9

Using the notation defined in (5.5.1), we have

$$\begin{aligned} &\|\hat{\Sigma}_{\ddot{X}^i\ddot{X}^j|U^{ij}} - \Sigma_{\ddot{X}^i\ddot{X}^j|U^{ij}}\|_{\text{HS}} \\ &= \|\hat{\Sigma}_{(X^iU^{ij})(X^jU^{ij})} - \Sigma_{(X^iU^{ij})(X^jU^{ij})}\|_{\text{HS}} + \|\tilde{A}\tilde{B}\tilde{C} - ABC\|_{\text{HS}} \end{aligned} \tag{5.7.1}$$

By Lemma 5, the first term is of the order $O_P(n^{-1/2})$. Similar to (5.5.3),

$$\tilde{A}\tilde{B}\tilde{C} - ABC = (\tilde{A} - A)\tilde{B}\tilde{C} + A(\tilde{B} - B)\tilde{C} + AB(\tilde{C} - C). \tag{5.7.2}$$

By Lemma 5, $\tilde{A} - A = \ddot{O}_P(n^{-1/2})$; by (5.5.9), $\tilde{B}\tilde{C} = \dot{O}_P(1)$. Hence the first term is of the order $\ddot{O}_P(n^{-1/2})$. The second term is

$$A(\tilde{B} - B)\tilde{C} = AB(B^{-1} - \tilde{B}^{-1})\tilde{B}\tilde{C} = \ddot{O}_P(\delta_n + n^{-1/2}). \tag{5.7.3}$$

It is easy to see that the third term on the right-hand side of (5.7.2) is also of the order $\ddot{O}_P(\delta_n + n^{-1/2})$. Hence

$$\tilde{A}\tilde{B}\tilde{C} - ABC = \ddot{O}_P(n^{-1/2} + \delta_n) = \ddot{O}_P(\delta_n),$$

where the last equality holds because $n^{-1/2} \prec b_n \preceq \delta_n$.

5.8 Proof of Theorem 11

When ϵ_n, η_n, and ϵ_n take the given form, the convergence rate in (4.6.3) becomes

$$b_n \asymp n^{3b/2+a-1} + n^{b-1/2} + n^{-b/2} + n^{-a} \asymp \max(n^{3b/2+a-1}, n^{b-1/2}, n^{-b/2}, n^{-a})$$

We need to minimize b_n over the set

$$C = \{(a, b) : \ a < \tfrac{1}{2},\ b < \tfrac{1}{2},\ \tfrac{3b}{2} + a - 1 < 0\}.$$

Equivalently, we need to minimize

$$f(a,b) = \max(\tfrac{3b}{2} + a - 1, b - \tfrac{1}{2}, -\tfrac{b}{2}, -a)$$

over C. Our strategy is to minimize $f(a,b)$ over $(0,\frac{1}{2}) \times (0,\frac{1}{2})$ and then check the minimizers (there are more than one) belong to C.

Because $b - \frac{1}{2} \geq -\frac{b}{2}$ iff $b \geq \frac{1}{3}$, we have

$$f(a,b) = \begin{cases} \max(\frac{3b}{2} + a - 1, b - \frac{1}{2}, -a) & b \geq \frac{1}{3} \\ \max(\frac{3b}{2} + a - 1, -\frac{b}{2}, -a) & b < \frac{1}{3} \end{cases}$$

Furthermore, for $b \geq \frac{1}{3}$,

$$f(a,b) = \max(\tfrac{3b}{2} + a - 1, b - \tfrac{1}{2}, -a) = \begin{cases} -a & a \in (0, \frac{1}{2} - b) \\ b - \frac{1}{2} & a \in [\frac{1}{2} - b, \frac{1}{2} - \frac{b}{2}] \\ \frac{3}{2}b + a - 1 & a \in (\frac{1}{2} - \frac{b}{2}, \frac{1}{2}) \end{cases}$$

which implies

$$\min_{0<a<\frac{1}{2}} f(a,b) = b - \tfrac{1}{2} \Rightarrow \min_{b \geq \frac{1}{2}} \min_{0<a<\frac{1}{2}} f(a,b) = \tfrac{1}{3} - \tfrac{1}{2} = -\tfrac{1}{6}$$

For $b < \frac{1}{3}$,

$$f(a,b) = \max(\tfrac{3b}{2} + a - 1, b - \tfrac{1}{2}, -a) = \begin{cases} -a & a \in (0, \frac{b}{2}) \\ -\frac{b}{2} & a \in [\frac{b}{2}, \frac{1}{2} - \frac{3b}{4}] \\ \frac{3b}{2} + a - 1 & a \in (\frac{1}{2} - \frac{3b}{4}, \frac{1}{2}) \end{cases}$$

which implies

$$\min_{0<a<\frac{1}{2}} f(a,b) = -\tfrac{1}{2}b \Rightarrow f(a,b) > -\tfrac{1}{6} \text{ for all } b \in (0,\tfrac{1}{3}),\ a \in (0,\tfrac{1}{2}).$$

Thus $f(a,b)$ reaches its minimum $-\frac{1}{6}$ when $b = \frac{1}{3}$, $a \in [\frac{1}{2} - b, \frac{1}{2} - \frac{b}{2}] = [\frac{1}{6}, \frac{1}{3}]$. Finally, it is easy to check that this set is contained in C.

Chapter 6

Simulation Studies and Real Data Analysis

6.1 Simulation Studies for Post Dimension Reduction Inference

We investigate the finite-sample performance of our post dimension reduction inference method, and compare with the naive inference method that pretends $\hat{\eta}^{\mathsf{T}}X$ were the true predictor. As discussed in Section 3.2.3, the asymptotic covariances of the two methods are $\Gamma(\eta_0, \theta_0)$ and $\tilde{\Gamma}(\eta_0, \theta_0)$, respectively. Given the data, $\Gamma(\eta_0, \theta_0)$ and $\tilde{\Gamma}(\eta_0, \theta_0)$ are estimated by $\Gamma(\hat{\eta}, \hat{\theta})$ and $\tilde{\Gamma}(\hat{\eta}, \hat{\theta})$. We consider five dimension reduction methods, SIR, SAVE, DR, y-PHD and r-PHD, and one estimation method, GMM. For GMM, let $m(\eta^{\mathsf{T}}X, \theta)$ denote the mean function, which is the same as the median function in our simulations as a symmetric error distribution is employed, and we set

$$g_1(\theta, \eta^{\mathsf{T}}X, Y) = Y - m(\eta^{\mathsf{T}}X, \theta), \quad g_2(\theta, \eta^{\mathsf{T}}X, Y) = \mathrm{I}(Y \leq m(\eta^{\mathsf{T}}X, \theta)) - 1/2.$$

That is, the GMM combines mean regression and median regression, which strikes a balance between efficiency and robustness. We compare the performance in terms of the coverage probability of confidence interval and the local power in hypothesis testing.

6.1.1 Comparison of confidence interval

For confidence interval comparison, we consider two models. The first model is

$$\text{Model I}: \; Y = \theta_1(\eta^\mathsf{T} X) + \theta_2(\eta^\mathsf{T} X)^2 + \sigma\epsilon,$$

where $X \sim N(0, I_5)$, $\epsilon \sim N(0,1)$, $X \perp\!\!\!\perp \epsilon$, $\theta_1 = \theta_2 = 1$, $\sigma = 0.5, 1$, the predictor dimension $p = 5$, and the sample size $n = 300, 400, 800, 1200$. In this example, $\mathcal{S}_{Y|X} = \text{span}(\eta)$ with $\eta = (1,0,0,0,0)^\mathsf{T}$. For the number of slices for SIR, SAVE, and DR, the general rule of thumb is to choose a larger value for SIR, and a smaller value for SAVE and DR (Li, 2018b). In our simulations, we have chosen $H = 20$ for SIR, $H = 2$ for SAVE, and $H = 8$ for DR. After obtaining $\hat{\eta}$ and $\hat{\theta}(\hat{\eta})$, we calculate the 95% confidence intervals for θ_1 and θ_2. We report the coverage probabilities of the two methods based on 200 data replications in Table 6.1. We see that the coverage probability from the naive method is considerably smaller than the nominal value, whereas the coverage probability from our proposed method is much closer. Table 6.1 also shows that the coverage probability for the naive method becomes closer to the nominal value as the sample size increases, but it does not converge to the nominal value.

Table 6.1. Coverage probability of confidence interval for θ_1 and θ_2 in model I.

n	Θ	σ^2	SIR		SAVE		DR		y-PHD		r-PHD	
			Γ	$\tilde{\Gamma}$	Γ	$\tilde{\Gamma}$	Γ	$\tilde{\Gamma}$	Γ	$\tilde{\Gamma}$	Γ	$\tilde{\Gamma}$
300	θ_1	0.5	0.96	0.82	0.96	0.81	0.96	0.83	0.95	0.81	0.96	0.82
		1	0.95	0.79	0.95	0.78	0.94	0.81	0.96	0.80	0.93	0.79
	θ_2	0.5	0.93	0.80	0.94	0.80	0.94	0.79	0.96	0.81	0.96	0.81
		1	0.94	0.78	0.93	0.80	0.94	0.80	0.94	0.79	0.96	0.79
400	θ_1	0.5	0.95	0.85	0.96	0.85	0.95	0.85	0.96	0.84	0.95	0.85
		1	0.96	0.81	0.94	0.83	0.93	0.82	0.93	0.81	0.93	0.83
	θ_2	0.5	0.95	0.83	0.94	0.85	0.94	0.84	0.96	0.84	0.96	0.84
		1	0.95	0.82	0.94	0.81	0.94	0.81	0.94	0.82	0.94	0.81
800	θ_1	0.5	0.96	0.88	0.96	0.89	0.94	0.89	0.96	0.88	0.95	0.88
		1	0.96	0.88	0.95	0.87	0.93	0.87	0.95	0.86	0.93	0.86
	θ_2	0.5	0.96	0.87	0.94	0.88	0.96	0.87	0.95	0.86	0.95	0.87
		1	0.93	0.86	0.96	0.85	0.94	0.86	0.93	0.85	0.94	0.85
1200	θ_1	0.5	0.95	0.92	0.96	0.91	0.96	0.91	0.96	0.92	0.96	0.91
		1	0.94	0.90	0.94	0.90	0.94	0.88	0.95	0.89	0.96	0.90
	θ_2	0.5	0.94	0.92	0.95	0.91	0.95	0.91	0.96	0.91	0.95	0.91
		1	0.94	0.91	0.93	0.88	0.94	0.90	0.95	0.90	0.96	0.89

The second model is

$$\text{Model II}: \ Y = \theta_1 \frac{\eta^\intercal X}{(\eta^\intercal X + 2)^2 + 0.1} + \sigma\epsilon,$$

where $X \sim N(0, I_{10})$, $X \perp\!\!\!\perp \epsilon$, $\theta_1 = 1$, and $\eta = (1, 0, \ldots, 0)^\intercal$. The rest of the setup is the same as model I. We report the coverage probabilities in Table 6.2. Again, the coverage probability of our method is much closer to 95% than the naive method.

Table 6.2. Coverage probability of confidence interval for θ_1 in model II.

n	Θ	σ^2	SIR		SAVE		DR		y-PHD		r-PHD	
			Γ	$\tilde{\Gamma}$	Γ	$\tilde{\Gamma}$	Γ	$\tilde{\Gamma}$	Γ	$\tilde{\Gamma}$	Γ	$\tilde{\Gamma}$
300	θ_1	0.5	0.94	0.85	0.93	0.83	0.95	0.82	0.93	0.83	0.96	0.83
		1	0.93	0.84	0.94	0.80	0.94	0.81	0.95	0.80	0.93	0.81
400	θ_1	0.5	0.95	0.86	0.95	0.87	0.95	0.86	0.93	0.86	0.95	0.85
		1	0.94	0.84	0.95	0.84	0.94	0.87	0.94	0.84	0.94	0.85
800	θ_1	0.5	0.95	0.90	0.94	0.88	0.96	0.89	0.96	0.90	0.95	0.90
		1	0.94	0.87	0.93	0.86	0.93	0.90	0.95	0.88	0.94	0.89
1200	θ_1	0.5	0.96	0.91	0.95	0.91	0.94	0.92	0.94	0.91	0.96	0.92
		1	0.95	0.92	0.94	0.90	0.95	0.90	0.95	0.89	0.95	0.90

Since the true model is known in the simulation experiments, we can also estimate (η, θ) and make inference about them directly using the maximum likelihood method without going through dimension reduction. It would be informative to compare this "oracle" inference method with the naive and objective inference methods. We have carried out this comparison using Model I, with $n = 400$. We read off the standard errors for $\hat{\theta}_1^{\text{MLE}}, \hat{\theta}_2^{\text{MLE}}$ from the asymptotic variance matrix of $(\hat{\eta}^{\text{MLE}}, \hat{\theta}^{\text{MLE}})$, which is the inverted Fisher information evaluated at the MLE. We also compute the standard errors for the $(\hat{\theta}_1, \hat{\theta}_2)$ obtained by SIR+GMM as described above, using the naive and objective inference methods. We repeat the process 200 times to compute the average standard errors. The results are reported in Table 6.3.

Table 6.3. Standard errors for θ_1 and θ_2 in model I and comparison to the oracle method.

parameter	naive	objective	oracle
θ_1	0.03	0.06	0.09
θ_2	0.02	0.04	0.03

In theory, we would expect the standard errors for $\hat{\theta}_1^{\text{MLE}}$ and $\hat{\theta}_2^{\text{MLE}}$ using the oracle inference method to be smaller than their counterparts for $\hat{\theta}_1$ and $\hat{\theta}_2$ using

the objective method, because MLE is asymptotically efficient. But this is not necessarily true in finite-sample, as indicated by our results. Also, Table 6.3 shows that both the objective and oracle mean standard errors are substantially larger than their counterparts by the naive method, which is not surprising because the naive method claims more information than it actually possesses.

6.1.2 Comparison of local power

For power comparison, we again consider two models. The first model is

$$\text{Model III}: \; Y = \theta_1(\eta_1^\mathsf{T} X)^2 + \theta_2 \exp(\eta_2^\mathsf{T} X) + \sigma\epsilon,$$

where $X \sim N(0, I_{10})$, $\epsilon \sim N(0,1)$, $X \perp\!\!\!\perp \epsilon$, $\theta_1 = \theta_2 = 1$, $\sigma = 0.5$, $p = 10$, and $n = 300, 500, 800, 1200$ with 50 replications. In this example, $\mathcal{S}_{Y|X} = \text{span}(\eta_1, \eta_2)$ with $\eta_1 = (1, 0, \ldots, 0)^\mathsf{T}$ and $\eta_2 = (0, 1, 0, \ldots, 0)^\mathsf{T}$. We consider the pair of hypotheses

$$H_0 : \theta_1 = 0 \quad \text{vs} \quad H_1 : \theta_1 \neq 0$$

which amounts to taking $h(\theta_1, \theta_2) = \theta_1$ in Section 3.5. The asymptotic power is computed as in (3.5.2). Figure 6.1 reports this asymptotic power as a function of the local parameter λ when the sample size is 500, with one panel corresponding to one of the five SDR methods. Figures 6.7, 6.8, 6.9 in the online Supplementary present the results for $n = 300, 800, 1200$, respectively. It is seen that the powers of the naive method, as shown by the red curves, are higher than those by our proposed method, as shown by the blue curves. This reflects that the naive method yields an overly optimistic power, as it does not take into account the estimation error induced by the dimension reduction step. Furthermore, by comparing Figures 6.1, 6.7, 6.8, and 6.9, we observe that the difference between the local powers of the naive and the objective methods tends to be smaller as the sample size increases, which echoes the pattern in the comparison of confidence intervals.

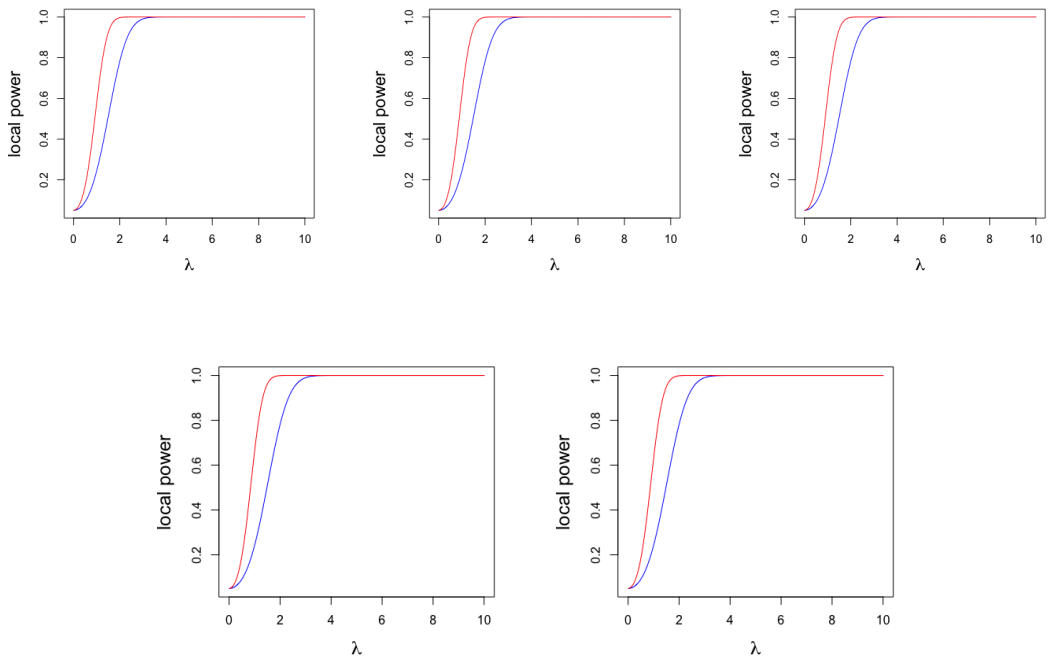

Figure 6.1. Local power of hypothesis testing in model III with sample size $n = 500$. The five panels, left to right, top to bottom, correspond to the results from five SDR methods, SIR, SAVE, DR, y-PHD, and r-PHD. The red curve denotes the naive inference method, and the blue curve denotes our proposed inference method.

Our second model for the local power comparison is

$$\text{Model IV}: \ Y = \theta_1 \frac{\eta_1^\intercal X}{(\eta_2^\intercal X + 1)^2 + 0.5} + \sigma\epsilon,$$

where $\theta_1 = 1$, and the rest of the setup is the same as model III. Figure 6.2 reports the results for $n = 500$. The same pattern is observed as in model III. Figures 6.10, 6.11, and 6.12 in the online Supplementary present the results for $n = 300, 800$, and 1200, respectively.

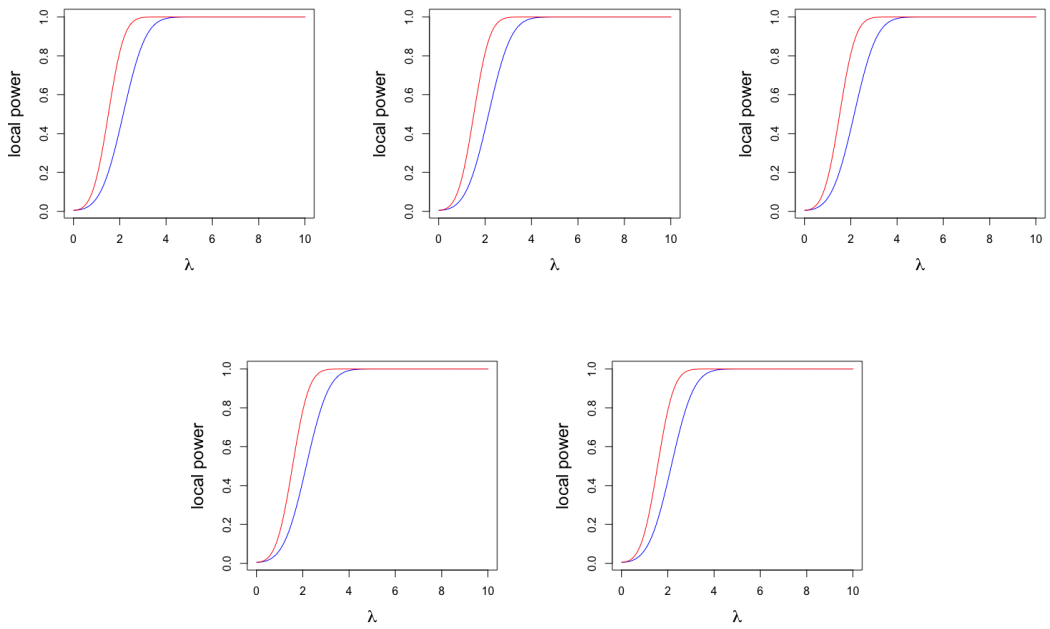

Figure 6.2. Local power of hypothesis testing in model IV with sample size $n = 500$. The five panels, left to right, top to bottom, correspond to the results from five SDR methods, SIR, SAVE, DR, y-PHD, and r-PHD. The red curve denotes the naive inference method, and the blue curve denotes our proposed inference method.

6.2 Application for Post Dimension Reduction Inference

We use the BigMac dataset to illustrate our post dimension reduction inference. The data concerns the relation between the minimum labor to buy a McDonald BigMac and fries, which serves as the response variable, and $p = 9$ economic predictors: minimum labor to buy one kilogram bread, lowest cost of 10k public transit, electrical engineer annual salary, tax rate paid by engineer, annual cost of 19 services, primary teacher salary, tax rate paid by primary teacher, average days of vacation per year, and average hours of work per year. The data is at `http://www.stat.umn.edu/arc/software.html`. Before the dimension reduction analysis, we applied the box-cox transformation to each individual predictor.

The sequential tests based on SIR yielded the p-values, 0.02, 0.20, 0.77, for the hypotheses $q = 0$ versus $q > 0$, $q = 1$ versus $q > 1$, and $q = 2$ versus $q > 2$, respectively, suggesting that the dimension of the central subspace is one and a

single linear combination is sufficient to fully capture the relationship between the response and the nine predictors. Figure 6.3 shows the scatterplot of the response versus the estimated sufficient predictor based on SIR.

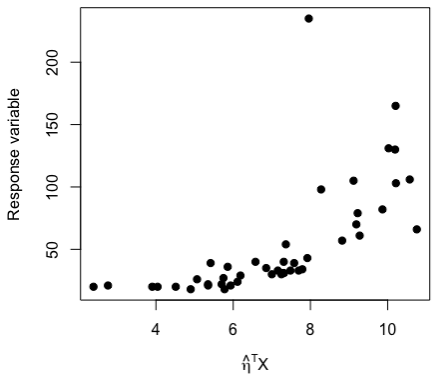

Figure 6.3. Response versus the first SIR predictor in BigMac data.

The scatter plot shows a clear nonlinear trend and a possible heteroscedastic pattern. As such we consider the following model,

$$Y = \theta_0 + \theta_1 \eta^{\mathsf{T}} X + \theta_2 (\eta^{\mathsf{T}} X)^2 + (\theta_3 + \theta_4 \eta^{\mathsf{T}} X)\epsilon,$$

where $\epsilon \sim N(0, 1)$. Based on this model, we aim to address two questions: First, is the nonlinear trend in Figure 6.3 significant? Second, is the heteroscedasticity in Figure 6.3 significant? These lead to the following two pairs of hypotheses,

$$H_0^{(1)} : \theta_2 = 0 \quad \text{vs} \quad H_1^{(1)} : \theta_2 \neq 0,$$
$$H_0^{(2)} : \theta_4 = 0 \quad \text{vs} \quad H_1^{(2)} : \theta_4 \neq 0.$$

To test these hypotheses we applied the naive method and the post dimension reduction method to the five SDR methods combined with the differential estimation equations. We use each method to construct confidence intervals for θ_2 and θ_4. The estimating equations are 5-dimensional $g(\theta, \eta^{\mathsf{T}} X, Y)$ obtained by

differentiating with respect to θ the objective function

$$\left[\frac{Y-\theta_0-\theta_1(\eta^\tau X)-\theta_2(\eta^\tau X)^2}{\theta_3+\theta_4(\eta^\tau X)}\right]^2.$$

Figure 6.4 shows the confidence intervals for θ_2 (the upper panel) and θ_4 (the bottom panel) obtained by different methods. In each plot, the left bar corresponds to the naive inference method, and the right one our proposed inference method. It is seen that, for θ_2, the confidence intervals produced by both inference methods do not cover 0, a clear evidence for the nonlinearity. For θ_4, all confidence intervals do not cover 0, a strong evidence for the heteroscedasticity. Moreover, the confidence intervals produced by the naive method are consistently narrower than those by our objective inference method.

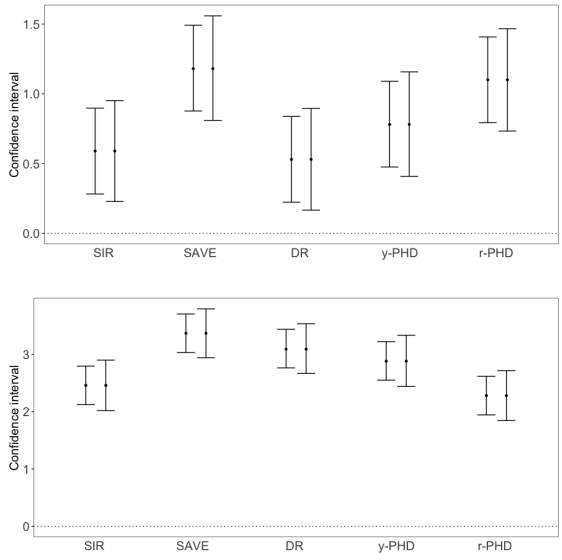

Figure 6.4. Confidence intervals for θ_2 (upper panel) and θ_4 (lower panel) in the BigMac data analysis.

To compare the local powers of the naive method and the post dimension reduction method, we applied them to the five SDR methods combined with the GMM estimation method. The GMM is based on two 5-dimensional estimating equations, with the first one, $g_1(\theta, \eta^\tau X, Y)$, being obtained by differentiating the objective function $[Y-\theta_0-\theta_1(\eta^\tau X)-\theta_2(\eta^\tau X)^2]^2$ with respect to θ, and the second

one, $g_2(\theta, \eta^\tau X, Y)$, being the the function

$$[Y - \theta_0 - \theta_2(\eta^\tau X) - \theta_2(\eta^\tau X)^2]^2 - (\theta_3 + \theta_4 \eta^\tau X)^2,$$

which is derived from the second moment assumption. Figure 6.5 shows the local powers of the five SDR methods based on the GMM. To save space, we only report the results for θ_2; the results for θ_4 exhibit a similar pattern. Again, the naive method yields an overly optimistic power compared with the objective method, which agrees with what we have observed in the simulations.

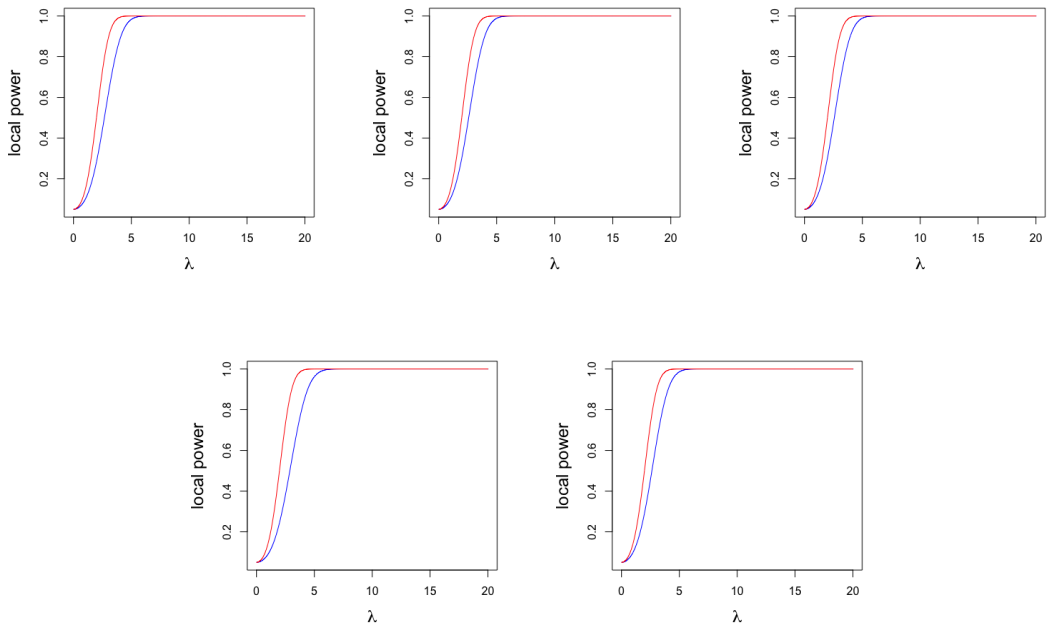

Figure 6.5. Local power for θ_2 in the BigMac data analysis. The five panels, left to right, top to bottom, correspond to the results from five SDR methods, SIR, SAVE, DR, y-PHD, and r-PHD. The red curve denotes the naive inference method, and the blue curve denotes our proposed inference method.

While the above analysis shows substantial differences in the confidence intervals by the naive and the objective inference methods, none of them is large enough to make the parameter statistically significant by one method and insignificant by the other. This turns out to be the case for the intercept parameter θ_1 when DR is used for dimension reduction. Figure 6.6 shows the confidence interval for θ_1

by the five SDR methods and the two inference methods. For DR, the naive inference method produces a confidence interval that does not contain 0, whereas the objective inference method produces a confidence interval that does. Thus θ_1 is statistically significant by the naive method but insignificant by the objective method.

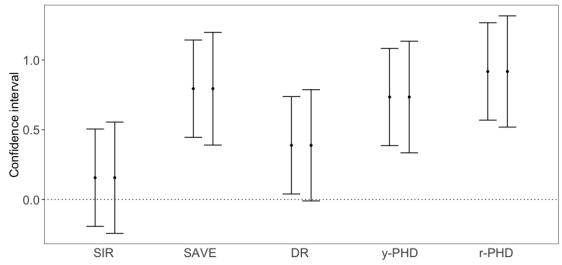

Figure 6.6. Confidence interval for θ_1 in the BigMac data analysis.

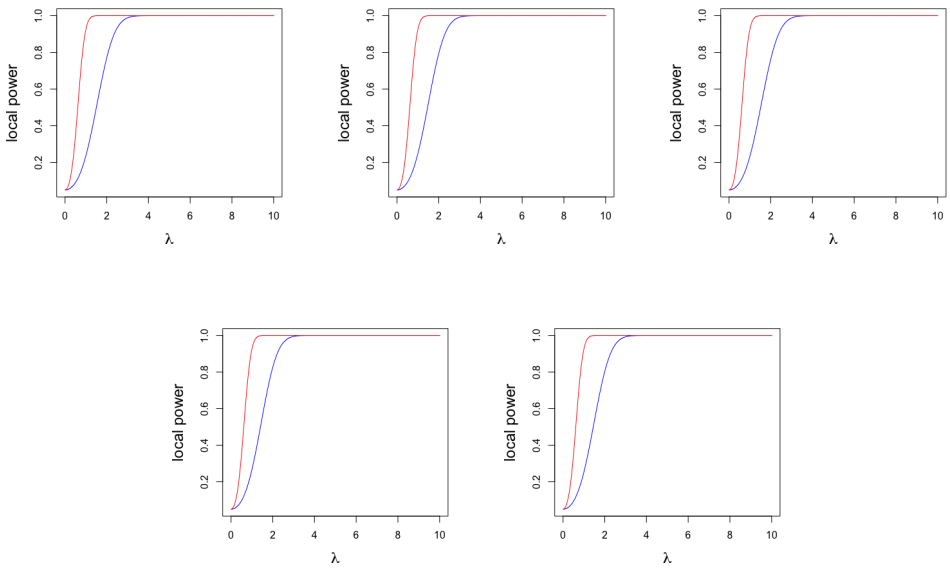

Figure 6.7. Local power of hypothesis testing in model III with sample size $n = 300$. The five panels, left to right, top to bottom, correspond to the results from five SDR methods, SIR, SAVE, DR, y-PHD, and r-PHD. The red curve denotes the naive inference method, and the blue curve denotes our proposed inference method.

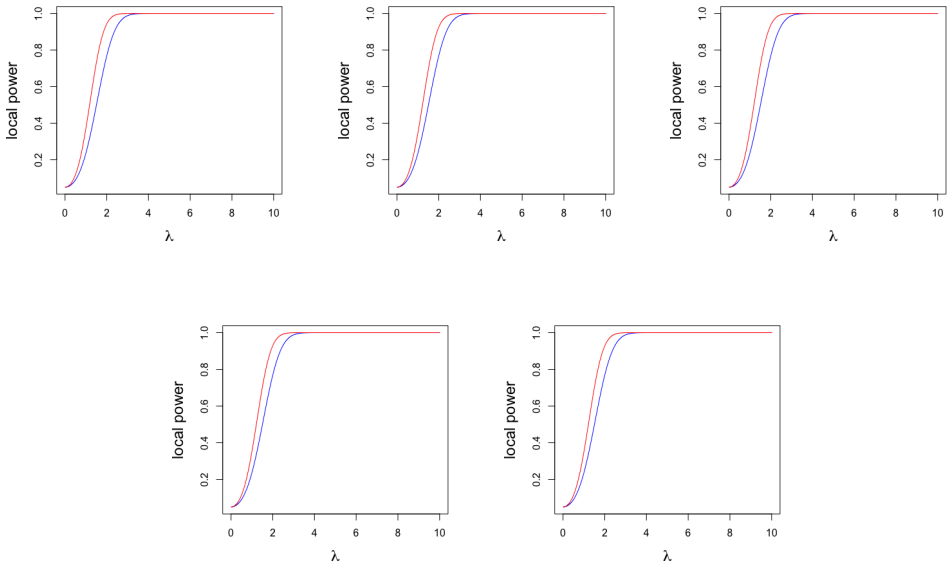

Figure 6.8. Local power of hypothesis testing in model III with sample size $n = 800$. The rest of setup is the same as in Figure 6.7.

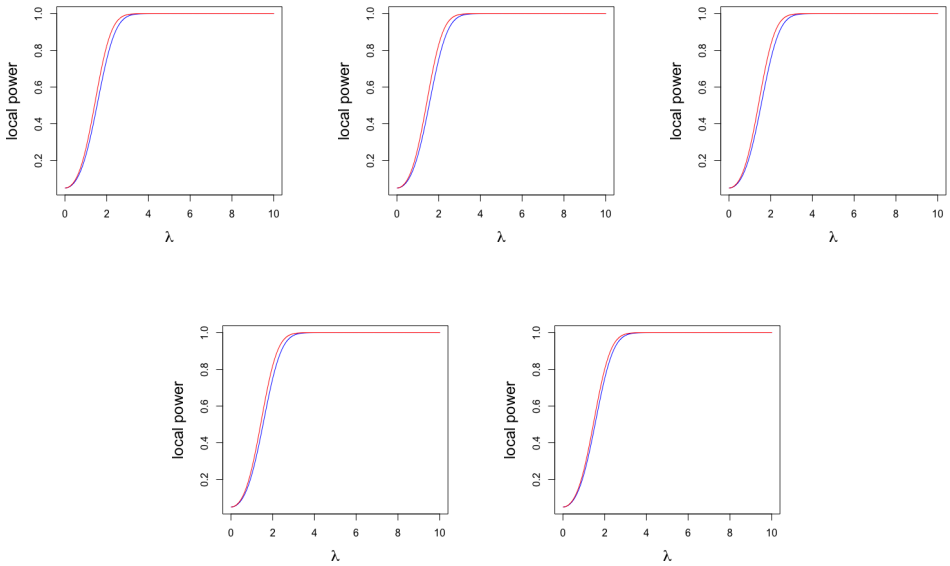

Figure 6.9. Local power of hypothesis testing in model III with sample size $n = 1200$. The rest of setup is the same as in Figure 6.7.

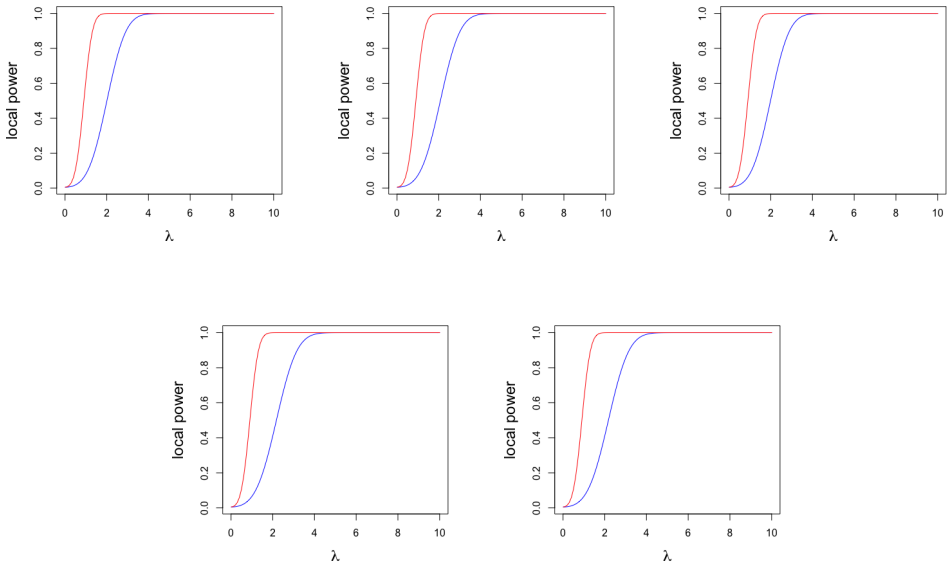

Figure 6.10. Local power of hypothesis testing in model IV with sample size $n = 300$. The five panels, left to right, top to bottom, correspond to the results from five SDR methods, SIR, SAVE, DR, y-PHD, and r-PHD. The red curve denotes the naive inference method, and the blue curve denotes our proposed inference method.

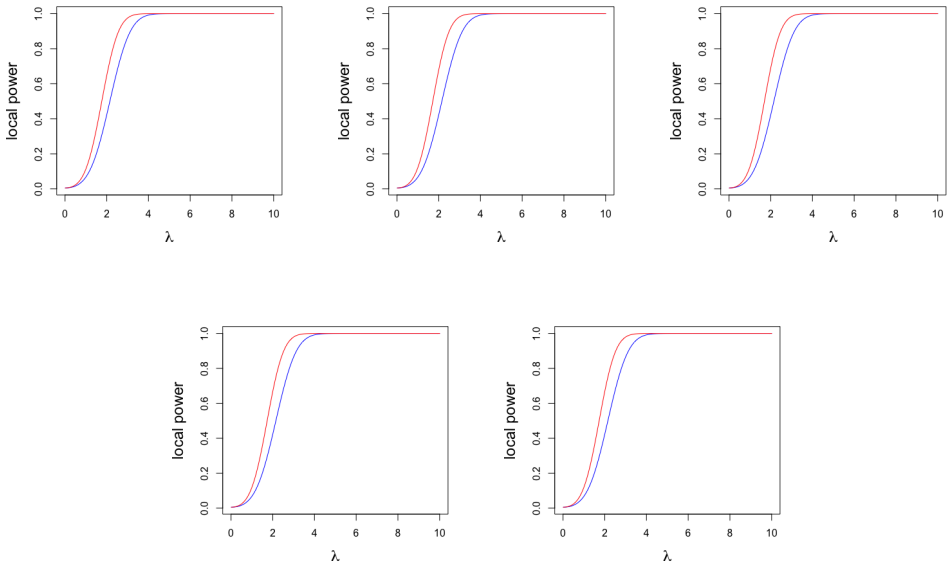

Figure 6.11. Local power of hypothesis testing in model IV with sample size $n = 800$. The rest of setup is the same as in Figure 6.10.

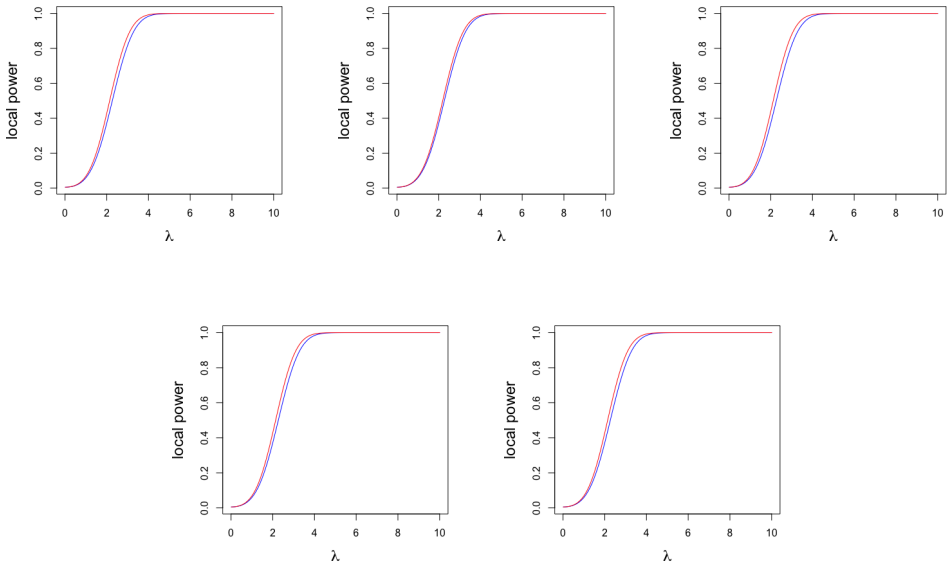

Figure 6.12. Local power of hypothesis testing in model IV with sample size $n = 1200$. The rest of setup is the same as in Figure 6.10.

6.3 Simulation Studies for Sufficient Graphical Models

In this section we compare the performance of our Sufficient Graphical Model with previous methods such as the Gaussian Graphical Model (GGM) in Yuan and Lin (2007), the Copula Gaussian Graphical Model (CGGM) in Liu et al. (2009), the nonparametric method (NP1) in Voorman et al. (2013), the nonparametric model (NP2) in (Fellinghauer et al., 2013), and the additive partial correlation operator (APCO) in Lee et al. (2016a).

By design, the SGM has advantages over these existing methods under the following circumstances. First, since SGM does not make any distributional assumption, it should outperform GGM and copula GGM when the Gaussian or copula Gaussian assumptions are violated; second, due to the sufficient dimension reduction in the first step of SGM, it avoids the curse of dimensionality and should outperform the fully nonparametric methods (NP1 and NP2) in the high-dimensional setting; third, since SGM does not require additive structure, it should outperform APCO when there is severe nonadditivity in the model. Our simulation comparisons will reflect these aspects.

For the CCCO and the APCO, we use the Gaussian radial basis function (RBF) as the kernel. The regularization constants $\epsilon_X^{(i,j)}$, $\epsilon_X^{-(i,j)}$, and $\epsilon_U^{(i,j)}$ are chosen by the GCV criterion described in Section 4.5 with the grid $\{10^{-\ell} : \ell = -1, 0, 1, 2, 3, 4\}$. The kernel parameters $\gamma_X^{(i,j)}$, $\gamma_X^{-(i,j)}$, $\gamma_{XU}^{i,ij}$, $\gamma_{XU}^{j,ij}$, and γ_U^{ij} are chosen according to (4.5.1). Because the outcomes of tuning parameters are stable, for each model, we compute the GCV for the first five samples and use their average value for the rest of the simulation.

The performance of each estimate is assessed using the averaged receiver operating characteristic (ROC) curve. Specifically, let ρ denote the threshold (as in the case of SGM, APCO, and NP2) or some sparsity-inducing tuning parameter (as in the case of GGM, CGGM, and NP1). Let $\mathcal{E}$ represent the true edge set and $\hat{\mathcal{E}}(\rho)$ the estimated edge set using one of the six methods for a given value of ρ. For each ρ, let $\mathrm{TP}(\rho)$ and $\mathrm{FP}(\rho)$ denote the true positive rate and false positive rate of the estimate $\hat{\mathcal{E}}(\rho)$. That is,

$$\mathrm{TP}(\rho) = \frac{\mathrm{card}(\mathcal{E} \setminus \hat{\mathcal{E}}(\rho))}{\mathrm{card}(\mathcal{E})}, \quad \mathrm{FP}(\rho) = \frac{\mathrm{card}(\hat{\mathcal{E}}(\rho) \setminus \mathcal{E})}{\mathrm{card}(\mathcal{E}^c)},$$

where card$(\cdot)$ denotes the cardinality of a set, and $\mathcal{E}^c$ is the set

$$\{(i,j) \in \Gamma \times \Gamma : \ i > j\} \setminus \mathcal{E}.$$

The ROC curve is the set $\{(\mathrm{TP}(\rho), \mathrm{FP}(\rho)) : \rho \in I\}$ for some interval I sufficiently wide to make $\mathrm{TP}(\rho)$ and $\mathrm{FP}(\rho)$ arbitrarily close to 0 and 1. The accuracy of a method across all ρ is measured by the area under the ROC curve.

To isolate the factors that affect accuracy, we first consider two models with relatively small dimensions and large sample sizes, which are

$$\begin{aligned}
\text{Model I}: \quad & X_1 = \epsilon_1, \ X_2 = \epsilon_2, \ X_3 = \sin(2X_1) + \epsilon_3 \\
& X_4 = X_1^2 + X_2^2 + \epsilon_4, \ X_5 = \epsilon_5, \\
\text{Model II}: \quad & X_1 = \epsilon_1, \ X_2 = X_1 + \epsilon_2, \ X_3 = \epsilon_3, \ X_4 = (X_1 + X_3)^2 + \epsilon_4, \\
& X_5 = \cos(2X_2X_3) + \epsilon_5, \ X_6 = X_4 + \epsilon_6.
\end{aligned}$$

Model I is of dimension $p = 5$ with edge set

$$\mathcal{E} = \{(1,3), (1,4), (2,4), (1,2)\};$$

Model II is of dimension $p = 6$ with edge set

$$\mathcal{E} = \{(1,4), (3,4), (1,3), (2,5), (3,5), (2,5), (4,6)\}.$$

We choose two sample sizes $n = 100, 1000$ for each model, and for each n, we generate 50 samples to compute the averaged ROC curves for the six methods. The dimension d_{ij} for the dimension reduction step of SGM is taken to be 2 for all cases (we have also used $d_{ij} = 1$ and the results are very similar to those presented here).

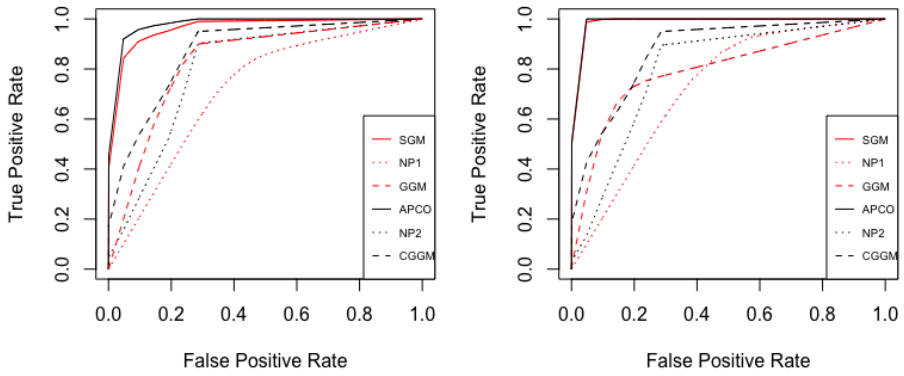

Figure 6.13. Averaged ROC curves of Model I. Left panel: $n = 100$; right panel: $n = 10$

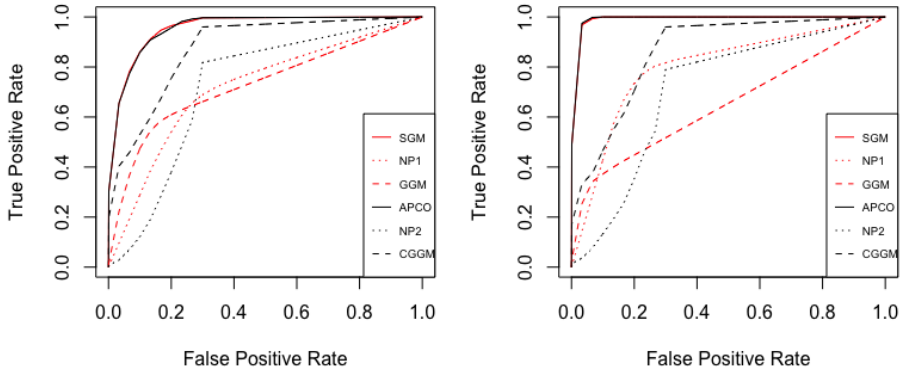

Figure 6.14. Averaged ROC curves of Model II. Left panel: $n = 100$; right panel: $n = 1000$

Figures 6.13 and 6.14 show the averaged ROC curves for the six methods considered, with the following assignment of plotting symbols

SGM:	red solid line
APCO:	black solid line
GGM:	red dashed line
CGGM:	black dashed line
NP1:	red dotted line
NP2:	black dotted line

From the figures we see that the two top performers are clearly SGM and APCO, and their performances are very similar. Note that none of the two models satisfies the Gaussian or copula Gaussian assumption, which explains why SGM and APCO outperform GGM and CGGM. SGM and APCO also outperform NP1 and NP2, indicating that the curse of dimensionality already takes effect on the fully nonparametric methods. Also note that Model I has an additive structure, which explains the slight advantage of APCO over SGM in Figure 6.13; Model II is not additive, and the advantage of APCO disappears in Figure 6.14.

We next consider two models with relatively high dimensions and small sample sizes. A convenient systematic way to generate larger networks is via the hub structure. We choose $p = 200$, and randomly generate ten hubs $h_1, \ldots, h_{10}$ from the set of vertices $\{1, \ldots, 200\}$. For each hub h_k, we randomly select a set H_k of nineteen vertices from $\{1, \ldots, 200\}$ to form the set of neighbors of h_k. As a result, each of the 10 modules has 20 members. With the network structures thus specified, our two probabilistic models are

$$\text{Model III} : X_i = 1 + |X_{h_k}|^2 + \epsilon_i, \quad \text{where} \quad i \in H_k \setminus h_k,$$
$$\text{Model IV} : X_i = \sin(X_{h_k}^3)\epsilon_i, \quad \text{where} \quad i \in H_k \setminus h_k.$$

Note that, in Model III, the dependence of X_i on X_{h_k} is through the conditional mean $E(X_i|X_{h_k})$, whereas in Model IV, the dependence is through the conditional variance $\text{var}(X_i|X_{h_k})$. For each model, we choose two sample sizes $n = 50$ and $n = 100$. For each sample size, we generate 50 samples to compute the averaged ROC curves, which are presented in Figures 6.15 and 6.20.

From the figures we see that, in the high-dimensional setting (in particular, with sample size smaller than dimension), SGM substantially outperforms all the other methods, which clearly indicates the benefit of sufficient dimension reduction in constructing graphical models.

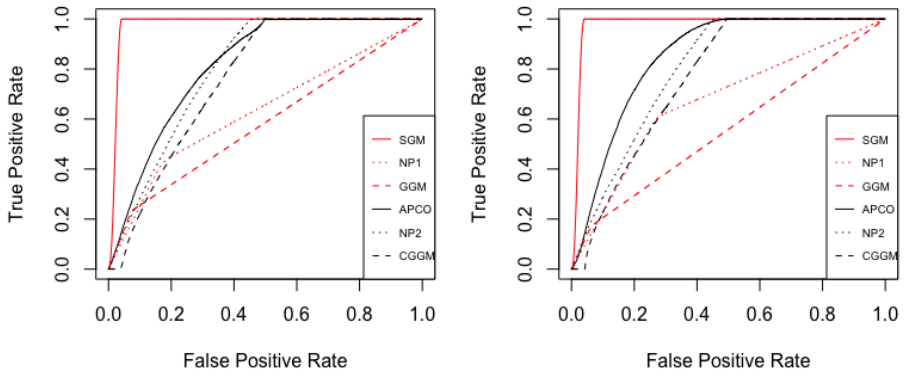

Figure 6.15. Averaged ROC curves for Model III. Left panel: $n = 50$; right panel: $n = 100$.

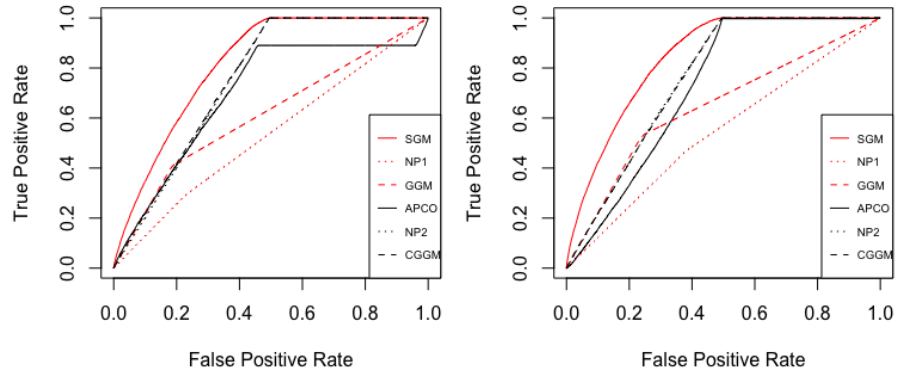

Figure 6.16. Averaged ROC curves for Model IV. Left panel: $n = 50$; right panel: $n = 100$.

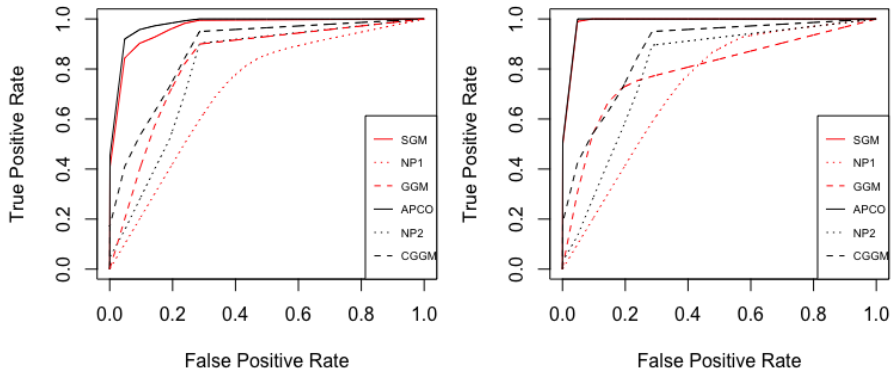

Figure 6.17. Averaged ROC curves for Model I. Left panel: $n = 100$; right panel: $n = 1000$ when we choose one dimension.

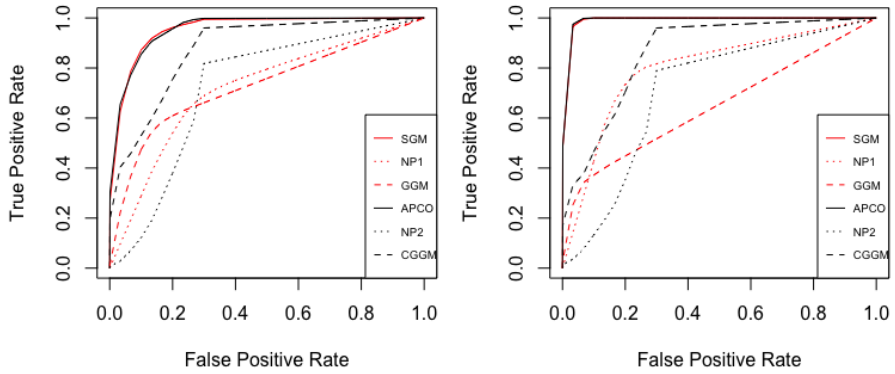

Figure 6.18. Averaged ROC curves for Model II. Left panel: $n = 100$; right panel: $n = 1000$ when we choose one dimension.

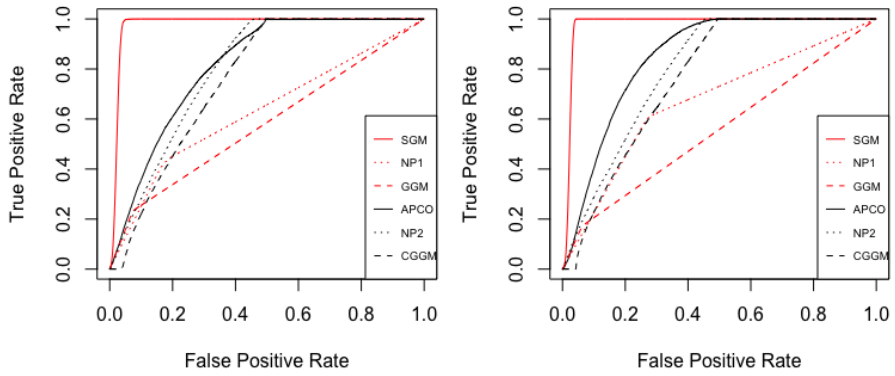

Figure 6.19. Averaged ROC curves for Model III. Left panel: $n = 50$; right panel: $n = 100$ when we choose one dimension.

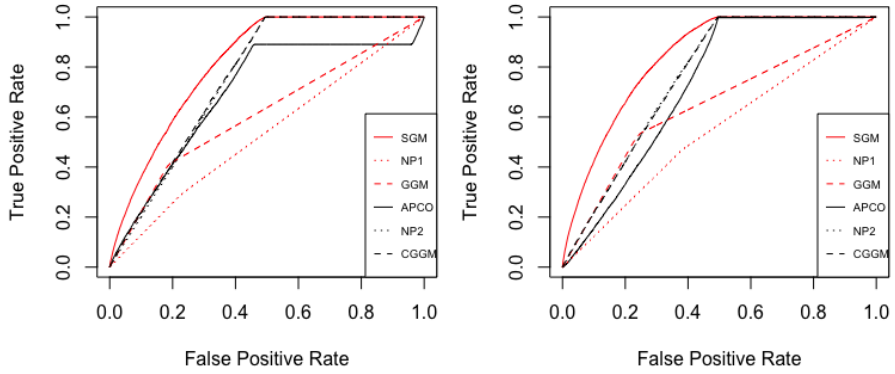

Figure 6.20. Averaged ROC curves for Model IV. Left panel: $n = 50$; right panel: $n = 100$ when we choose one dimension.

6.4 Application for Sufficient Graphical Models

In this section we apply our SGM to a dataset from the DREAM 4 Challenge project (Marbach et al., 2010) and compare it with other methods. The goal of this network challenge is to recover gene regulation networks from simulated steady-state data. The graphs were formed by extracting subgraphs from real biological networks. The gene expression levels are derived from a system of ordinary differential equations controlling the dynamics of the biological interactions

between the genes. For a more detailed description, see Marbach et al. (2010). Because Lee et al. (2016a) already compared APCO with GGM, CGGM, NP1, and NP2 for this dataset and demonstrated the superiority of APCO among these estimators, here we will focus on the comparison of the SGM with APCO and the champion method for the DREAM 4 Challenge.

In the dataset there are five networks each of dimension of 100. For each network, a sample of size 201 is obtained by putting together observations from three different experimental conditions: wild-type, knockdown, and knockout. Because the DREAM 4 Challenge provides the true networks, we are able to produce the ROC curves for the three methods and compare their Areas Under the Curves (AUC). As in the simulation study, we use the Gaussian RBF kernel for SGM and APCO and select tuning parameters by the Generalized Cross Validation described in Section 4.5. For SGM, the dimensions d_{ij} for the dimension reduction step is taken to be 1. We have also repeated the computation for $d_{ij} = 2$ (not presented here) and the results are very similar. Table 6.4 shows the AUC's obtained from applying the three methods to the five networks in the data.

Table 6.4. Comparison of AUC for SGM, APCO, and the champion method for the five networks in the DREAM 4 Challenge dataset.

	Network 1	Network 2	Network 3	Network 4	Network 5
SGM	0.85	0.81	0.83	0.83	0.79
APCO	0.86	0.81	0.83	0.83	0.77
champion	0.91	0.81	0.83	0.83	0.75

As we can see from Table 6.4, SGM has the same AUC values as APCO for Networks 2, 3, and 4, performs better than APCO for Network 5, but trails slightly behind APCO for Network 1. Furthermore, SGM has the same AUC as the champion method, performs better than for Network 5 and worse for Network 1. Overall, SGM and APCO perform similarly in this dataset, and they are on a par with the champion method. Here, we would also like to point out that both the SGM and APCO are purely empirical; they employ no knowledge about the underlying physical mechanism that generates the gene expression data, whereas, according

to Pinna et al. (2010), the champion method did use a differential equation that reflects the underlying physical mechanism.

Chapter 7

Future Research

In the future, I am going to keep doing research with Professor Bing Li. The following projects are some of the future research agendas we have.

7.1 Functional sufficient graphical models with application to f-MRI and EEG dataset

Functional data is one of the most prevalent formats of the dataset in contemporary research, such as anthropology, biology, forensic science, and many others. Functional data have an especially complex and structured format that can be viewed as smooth curves or functions rather than numbers or vectors. Because of a special type of data format, there have been explosive development of a methodology to analyze functional data over the past decade or so. See for example, Ramsay and Silverman (2005), Ramsay and Silverman (2007), Yao, Müller, and Wang (2005), Ferraty and Vieu (2006), Horváth and Kokoszka (2012), and Hsing and Eubank (2015).

Especially, a functional graphical model is popular to infer networks where the observations on the vertices are random functions instead of a random vector. See, Zhu, Strawn, and Dunson (2016), Qiao, Guo, and James (2018), and Li and Solea (2018b). We can easily encounter a dataset of this structure, such as electroencephalography (EEG) or functional magnetic resonance imaging (f-MRI) dataset. The previous research, such as Qiao, Guo, and James (2018), is based on the case that observed random functions at vertices are Gaussian random elements in

Hilbert space. In this project, we propose incorporating ideas and techniques from the most recent advances in functional sufficient dimension reduction in statistics and machine learning into the functional graphical model to enhance accuracy and computational efficiency by reducing the dimension of the mapping kernels.

A graph consists of a finite set of nodes $\Gamma = \{1, \ldots, p\}$ and set of undirected edges

$$E = \{(i, j) \in \Gamma \times \Gamma : i \neq j\}.$$

We assume $i > j$ for the convenience throughout this project. Above equation means that node i and node j are directly linked if and only if $(i, j) \in E$. We denote such a graph by $\mathscr{G} = (\Gamma, E)$. Let T be an interval in $\mathbb{R}$, representing time and $X = (X^1, \ldots, X^p)$ be a random function with each X^i representing random function defined on an interval T in $\mathbb{R}$ and it corresponds to a node. Estimating the set E is a main purpose of a statistical graphical model.

The meaning of statement "A node i is not directly linked to a node j" is that if one wants to go from node i to node j, one must go through the rest of the nodes –that is, nodes in $\Gamma \setminus \{i, j\}$. This corresponds to the following Markovian-type probabilistic statement

$$(i, j) \notin E \Longleftrightarrow X^i \perp\!\!\!\perp X^j | X^{-(i,j)} \tag{7.1.1}$$

where $\perp\!\!\!\perp$ means conditional independence, and $X^{(i,j)}$ represent the random function (X^i, X^j) and $, X^{-(i,j)}$ to denote X with its i-th and j-th components removed. In other words, a statistical graphical model is defined by $(i, j) \notin E$ if and only if right hand side of (7.1.1) holds true. For a comprehensive graphical model, see, Zhu, Strawn, and Dunson (2016), Qiao, Guo, and James (2018), and Li and Solea (2018b).

Our principal motivating example is f-MRI dataset which is collected by ADHD consortium (Milham et al., 2012). In f-MRI dataset, each vertex corresponds to a subregion of a brain and it is described as a collection of voxels. For each voxel, brain oxygen level dependent which is a form of brain activities is recorded over a time point. Then they are aggregated over voxels in each subregion, resulting in a vector of interdependent random functions. One of the goals is making a method-

ology to find inter-dependence of a brain network based on random functions.

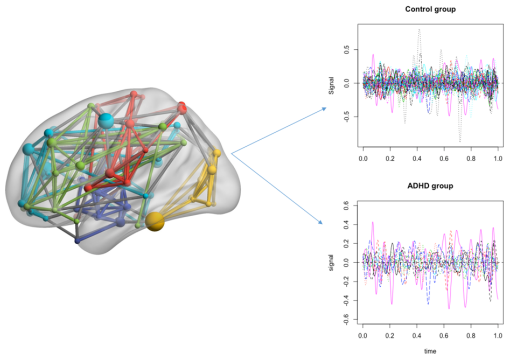

Figure 7.1. Example of human brain network (left) and f-MRI functional data for control group (Upper right) and ADHD group (Lower right).

Left side of figure 7.1 is image of brain network which is plotted by BrainNet viewer (Xia et al., 2013, `http://www.nitrc.org/projects/bnv/`). Right hand side are resting state data of one voxel based on time points between normal children with and without ADHD.

One of the most famous functional graphical models is a functional Gaussian graphical model (FGGM), which is introduced by Qiao, Guo, and James (2018). FGGM assumes $X = (X^1, \ldots, X^p)$ is a multivariate Gaussian random element in a Hilbert space. Under this assumption, the relation (7.1.1) can be modified as

$$(i,j) \notin E \Longleftrightarrow \mathrm{cov}[X^i(s), X^j(t)|X^{-(i,j)}] = 0 \quad \forall s,t \in T. \tag{7.1.2}$$

The fundamental idea of FGGM is to approximate X^i as a Karhunen Loeve expansion (Bosq (2012)) and choose the first m coefficient to form an m dimensional multivariate Gaussian random vector, for each X^i, say $M^i(m)$. Then we can form a pm dimensional Gaussian random vector with p subvectors by $M(m) = (M^1(m)^\intercal, \ldots, M^p(m)^\intercal)$. After that, they applied a group-lasso algorithm with a maximum likelihood method to give blockwise sparsity of pm dimensional precision matrix. It means zero blocks of the precision matrix between the corresponding pairs of subvectors implies conditional independence. They propose if we make m larger with n in appropriate rate, conditional independence between subvectors can correspond to relation (7.1.2). Even though FGGM has a clear

advantage based on intuitive structure, it also has some limitations. First, the Gaussian assumption can be violated in many real life applications such as skewness and kurtosis of the dataset. Second, the Gaussian assumption cannot cover the situation that vertices have a nonlinear or heteroscedasticitic relationship, which is common in the modern functional graphical models.

To overcome this limitation, the copula Gaussian graphical model, which relaxes the Gaussian assumption and retains its simple conditional independence structure, is introduced. See, Liu, Lafferty, and Wasserman (2009), Liu et al. (2012b), Xue and Zou (2012). Solea, Li (2019) expand this approach to the functional graphical model by introducing the functional copula Gaussian graphical model (FCGGM) (Solea, Li, 2019). Furthermore, Li and Solea (2018b) introduced nonparametric approach to functional graphical model. They relaxed Gaussian and copula Gaussian assumption by utilizing RKHS to map the observed random functions to nonlinear spaces. Moreover, instead of using conditional independence, they propose additive conditional independence to determine the absence of edges in the graph. This approach has an advantage in that it can capture nonlinear relationships among vertices with random functions.

We propose incorporating ideas and techniques from the most recent advances in functional nonlinear sufficient dimension reduction in Statistics and machine learning into the graphical models to reduce the dimension of the mapping kernels. For functional sufficient dimension reduction, we use f-GSIR (Li and Song, 2017b). This is an extension of Lee, Li, and Chiaromonte (2013) to the functional data analysis context. Moreover, our approach to the functional graphical model is by using a fully nonparametric method. Therefore our method can relax Gaussian and copula Gaussian assumptions. Besides the relaxation of the Gaussian and copula Gaussian assumptions, an attractive feature of the proposed methods is due to the use of the "kernel trick", which means that the complexity of the proposed methods depends on the sample size rather than the dimension of the networks. This makes their algorithms computationally economic when handling high-dimensional networks. Moreover, by utilizing nonlinear sufficient dimension reduction technique, our new methods can achieve better accuracy by avoiding "curse of dimensionality". Additionally, the asymptotic structures of the kernel estimates of linear operators are mathematically tractable, and some rudimentary

tools have been developed in the recent literature. This gives us a real prospect of developing a reasonably complete asymptotic theory for consistency, convergence rate, and statistical inference for the non-Gaussian functional graphical models.

Similar to Li and Song (2017b), Li and Solea (2018b), we assume random functions in each vertices are in Hilbert space. To develop a fully nonparametric approach, we define second level Hilbert space, which is based on the inner product of first-level Hilbert space to define conditional covariance operator between vertices and reduce dimension based on f-GSIR. We finally define functional sufficient partial correlation operator (f-SUPCO) and can determine independence between nodes based on the smallness of norms from f-SUPCO.

7.2 Extension of Post Dimension Reduction Inference

An immediate project in my plan is to extend Kim, Li, Yu, and Li (2019) to develop objective inference procedures for a much broader class of dimension reduction problems such as independent components analysis (Hyvärinen et al., 2004), nonparametric sufficient dimension reduction methods (Xia et al., 2002; Xia, 2007), and sparse sufficient dimension reduction methods (Bondell and Li, 2009; Chen et al., 2010; Wang and Yin, 2008). In particular, an extension to semiparametric SDR methods (Ma and Zhu, 2012, 2013, 2014) is promising because the influence function can be readily developed from the efficient score. Furthermore, the semiparametric efficiency for sufficient dimension reduction can be inherited, to some degree, by the post dimension reduction inference procedure. It is also my ongoing project (Kim and Li, 2019b) to extend post dimension reduction inference to the functional data analysis based on functional sliced inverse regression (Ferré and Yao, 2003) and functional principal component analysis.

7.3 Learning Causal Networks via Sufficient Faithfulness

Learning causality is crucial in many scientific areas, such as genetics, epidemiology, and business. In the classical setting, a directed acyclic graphs (DAG) can be linked to conditional independence through a *faithfulness condition*. We will borrow this idea and introduce a new statistical graphical model called sufficiently

faithful directed acyclic graphs (SFDAG) based on a new condition called *sufficient faithfulness*, which links a DAG with the sufficient nonparametric graphical model (Li and Kim, 2019; Kim and Li, 2019a). We also plan to develop a computationally efficient PC-type algorithm to implement the estimator of the SFDAG. The complexity of a new algorithm depends on the sparseness of SFDAG instead of a dimension of networks. We expect this feature not only makes our method feasible for high-dimensional networks because the skeleton of a DAG is typically sparse, but can also capture nonlinear interactions.

7.4 Functional Component Selection

Multivariate functional data that are neither one dimensional nor restricted to the same interval are prevalent in real-world applications such as neuroimaging data, which consists of functions and images. Existing methods mostly depend on observations on a common one-dimensional interval. To deal with this problem, we will develop a sparse Principal Component Analysis methodology for multivariate functional data (Song, Kim, and Li, 2020). The applications we have in mind are data with measurement error and sparse functional data. This approach combines functional Principal Component Analysis with sparse estimation. The results from this research could further be applied to *hybrid data*; that is, data consisting of a function and a vector part. This is a joint work with Dr. Jun Song of the University of North Carolina at Charlotte.

Appendix A

Computer Codes for Sufficient Dimension Reduction

```
#########################################################
#           weighted least squares coefficients
#           note that x includes the constant term
#           so its dimension is counted as p+1
#########################################################
wls = function(x,y,w){
  n=dim(x)[1];p=dim(x)[2]-1
  out=apply(x*y*w,2,mean)
  out1=t(x*w)%*%x/n
  out2=c(solve(out1)%*%out)
  return(list(a=out2[1],b=out2[2:(p+1)]))
}


#########################################################
#            slice average
#########################################################
slav = function(x,y,yunit){
  n = nrow(x)
```

```
  p = ncol(x)
  nslice = length(yunit)
  xgy = matrix(0,nslice,p)
  for(i in 1:nslice){
    xgy[i,] = apply(x[y==yunit[i],],2,mean)}
  return(xgy)
}
#####################################################
#          Compute slice proportions
#####################################################
slprob = function(y,yunit){
  n = length(y)
  nslice = length(yunit)
  out = rep(0,nslice)
  for(i in 1:nslice){
    out[i] = length(y[y==yunit[i]])/n}
  return(out)
}
#####################################################
#                 discretize Y
#     note that i added a small perturbation to
#     y; when y is discrete it will create wierd
#     slices if you treat it as continuous
#####################################################
discretize = function(y,yunit){
  n = length(y)
  y = y + .00001*mean(y)*rnorm(n)
  nsli=length(yunit)
  yord = y[order(y)]
  n = length(y)
  nwith = floor(n/nsli)
  divpt = rep(0,nsli-1)
  for(i in 1:(nsli-1)){
```

```
    divpt[i] = yord[i*nwith+1]}
  y1 = rep(0,n)
  y1[y>=divpt[nsli-1]]=nsli
  y1[y<divpt[1]]=1
  for(i in 2:(nsli-1)){
    y1[(y>=divpt[i-1])&(y<divpt[i])]=i}
  return(y1)
}

#########################################################
#           slice covariance
#   y must be discretized according to yunit
#########################################################
slco = function(x1,x2,y,yunit){
  n = nrow(x1)
  p = ncol(x1)
  nslice = length(yunit)
  cx1x2y = array(0,c(p,p,nslice))
  for(i in 1:nslice){
    cx1x2y[,,i] = cov(x1[y==yunit[i],],x2[y==yunit[i],])}
  return(cx1x2y)
}

#########################################################
#           mave
# beta in input is initial value of beta
# nit is number of iteration for mave
#########################################################
mave=function(x,y,h,beta,nit){
  sig=diag(var(x));n=dim(x)[1];p=dim(beta)[1];d=dim(beta)[2]
  x=apply(x,2,standvec)  #standardize X
  kermat=kern(x,h)
  for(iit in 1:nit){
```

```
    b=numeric();a=numeric()
    for(i in 1:n){
      wi=kermat[,i]
      ui=cbind(1,t(t(x)-x[i,])%*%beta)
      out=wls(ui,y,wi);a=c(a,out$a);b=cbind(b,out$b)}
    out=0;out1=0
    for(i in 1:n) {
      xi=kronecker(t(t(x)-x[i,]),t(b[,i]));yi=y-a[i];wi=kermat[,i]
      out=out+apply(xi*yi*wi,2,mean)
      out1=out1+t(xi*wi)%*%xi/n}
    beta=t(matrix(solve(out1)%*%out,d,p))
  }
  return(diag(sig^(-1/2))%*%beta)}


##########################################################
#           opg
##########################################################
opg=function(x,y,h,d){
  n=dim(x)[1];sig=diag(var(x))
  x=apply(x,2,standvec)
  out=kern(x,h)
  b=numeric()
  for(i in 1:n){
    wi=out[,i]
    xi=cbind(1,t(t(x)-x[i,]))
    b=cbind(b,wls(xi,y,wi)$b)}
  beta=eigen(b%*%t(b))$vectors[,1:d]
  return(diag(sig^(-1/2))%*%beta)
}

##########################################################
#           rmave
```

```
# beta in input is initial value of beta
# nit is number of iteration for mave
#########################################################
rmave=function(x,y,h,beta,nit){
  sig=diag(var(x));n=dim(x)[1];p=dim(beta)[1];d=dim(beta)[2]
  x=apply(x,2,standvec)
  for(iit in 1:nit){
    xb=x%*%beta
    kermat=kern(xb,h)
    beta0=beta
    b=numeric();a=numeric()
    for(i in 1:n){
      wi=kermat[,i]
      ui=cbind(1,t(t(x)-x[i,])%*%beta)
      out=wls(ui,y,wi);a=c(a,out$a);b=cbind(b,out$b)}
    out=0;out1=0
    for(i in 1:n) {
      xi=kronecker(t(t(x)-x[i,]),t(b[,i]));yi=y-a[i];wi=kermat[,i]
      out=out+apply(xi*yi*wi,2,mean)
      out1=out1+t(xi*wi)%*%xi/n}
    beta=t(matrix(solve(out1)%*%out,d,p))
    beta1=beta;print(c("distance=",dis(beta0,beta1)),quote=F)
  }
  return(diag(sig^(-1/2))%*%beta)}

#########################################################
#       sliced inverse regression
#  sig = variance of xc; signrt = sig^(-1/2)
#  y must be discretized according to yunit for
#  continuous y; for discrete y, use original y
#  with yunit being the distinct values in
#  discrete y
#########################################################
```

```
index1=which(min(range(y))+(diff(range(y))/5)*(0)<=y
             & y<=min(range(y))+(diff(range(y))/5)*(1))
index2=which(min(range(y))+(diff(range(y))/5)*(1)<=y
             & y<=min(range(y))+(diff(range(y))/5)*(2))
index3=which(min(range(y))+(diff(range(y))/5)*(2)<=y
             & y<=min(range(y))+(diff(range(y))/5)*(3))
index4=which(min(range(y))+(diff(range(y))/5)*(3)<=y
             & y<=min(range(y))+(diff(range(y))/5)*(4))
index5=which(min(range(y))+(diff(range(y))/5)*(4)<=y
             & y<=min(range(y))+(diff(range(y))/5)*(5))

### compute z bar step 3
## first slice
if(length(index1)==0){
  first=c(0,0,0,0,0)
}
else{
  z1bar1=sum(z1[index1])/length(index1)
  z2bar1=sum(z2[index1])/length(index1)
  z3bar1=sum(z3[index1])/length(index1)
  z4bar1=sum(z4[index1])/length(index1)
  z5bar1=sum(z5[index1])/length(index1)


  first=rbind(z1bar1,z2bar1,z3bar1,z4bar1,z5bar1)
}
#second slice
if(length(index2)==0){
  second=c(0,0,0,0,0)
}
else{
  z1bar2=sum(z1[index2])/length(index2)
```

```
  z2bar2=sum(z2[index2])/length(index2)
  z3bar2=sum(z3[index2])/length(index2)
  z4bar2=sum(z4[index2])/length(index2)
  z5bar2=sum(z5[index2])/length(index2)
  second=rbind(z1bar2,z2bar2,z3bar2,z4bar2,z5bar2)
}
## third slice
if(length(index3)==0){
  third=c(0,0,0,0,0)
}
else{
  z1bar3=sum(z1[index3])/length(index3)
  z2bar3=sum(z2[index3])/length(index3)
  z3bar3=sum(z3[index3])/length(index3)
  z4bar3=sum(z4[index3])/length(index3)
  z5bar3=sum(z5[index3])/length(index3)
  third=rbind(z1bar3,z2bar3,z3bar3,z4bar3,z5bar3)
}
## fourth slice
if(length(index4)==0){
  fourth=c(0,0,0,0,0)
}
else{
  z1bar4=sum(z1[index4])/length(index4)
  z2bar4=sum(z2[index4])/length(index4)
  z3bar4=sum(z3[index4])/length(index4)
  z4bar4=sum(z4[index4])/length(index4)
  z5bar4=sum(z5[index4])/length(index4)
  fourth=rbind(z1bar4,z2bar4,z3bar4,z4bar4,z5bar4)
}
## fifth slice

if(length(index5)==0){
```

```
  fifth=c(0,0,0,0,0)
}
else{
  z1bar5=sum(z1[index5])/length(index5)
  z2bar5=sum(z2[index5])/length(index5)
  z3bar5=sum(z3[index5])/length(index5)
  z4bar5=sum(z4[index5])/length(index5)
  z5bar5=sum(z5[index5])/length(index5)
  fifth=rbind(z1bar5,z2bar5,z3bar5,z4bar5,z5bar5)
}



be=matrix(c(beta,sigmahatt%*%beta,sigmahatt%*%sigmahatt%*%beta,sigmahatt%*%sigmah
bee=be%*%t(be)

result[,i]=eigen(bee)$vectors[,1:2]
}

apply(abs(result),1,median)



#########################################################################
#           computing eigendecomposition for SAVE candidate             #
#                        matrix on the Z-scale                          #
#                        y: discretized response                        #
#########################################################################

index1=which(min(range(y))+(diff(range(y))/5)*(0)<=y
             & y<=min(range(y))+(diff(range(y))/5)*(1))
index2=which(min(range(y))+(diff(range(y))/5)*(1)<=y
             & y<=min(range(y))+(diff(range(y))/5)*(2))
```

```
index3=which(min(range(y))+(diff(range(y))/5)*(2)<=y
             & y<=min(range(y))+(diff(range(y))/5)*(3))
index4=which(min(range(y))+(diff(range(y))/5)*(3)<=y
             & y<=min(range(y))+(diff(range(y))/5)*(4))
index5=which(min(range(y))+(diff(range(y))/5)*(4)<=y
             & y<=min(range(y))+(diff(range(y))/5)*(5))

### compute z bar step 3
## first slice

if(length(index1)<=1){
  mat1=martix(c(rep(0,25)),5,5)
}


if(length(index1)>1){
  mat1=matrix(c(z1[index1],z2[index1],z3[index1],
                z4[index1],z5[index1]),
              nrow=length(index1),ncol=5)
}
#second slice
if(length(index2)<=1){
  mat2=matrix(c(rep(0,25)),5,5)
}
if(length(index2)>1){
  mat2=matrix(c(z1[index2],z2[index2],z3[index2],
                z4[index2],z5[index2]),
              nrow=length(index2),ncol=5)
}
## third slice
if(length(index3)<=1){
  mat3=matrix(c(rep(0,25)),5,5)
}
```

```
if(length(index3)>1){
  mat3=matrix(c(z1[index3],z2[index3],z3[index3],
                z4[index3],z5[index3]),
              nrow=length(index3),ncol=5)
}
## fourth slice
if(length(index4)<=1){
  mat4=matrix(c(rep(0,25)),5,5)
}
if(length(index4)>1){
  mat4=matrix(c(z1[index4],z2[index4],z3[index4],
                z4[index4],z5[index4]),
              nrow=length(index4),ncol=5)
}
## fifth slice

if(length(index5)<=1){
  mat5=matrix(c(rep(0,25)),5,5)
}


if(length(index5)>1)
{
  mat5=matrix(c(z1[index5],z2[index5],z3[index5],
                z4[index5],z5[index5]),
              nrow=length(index5),ncol=5)
}



### v hat
dig=diag(c(rep(1,5)),5)
vhat=(dig-cov(mat1))%*%(dig-cov(mat1))*(length(index1)/
```

```
length(y))+(dig-cov(mat2))
%*%(dig-cov(mat2))*(length(index2)/length(y))
+(dig-cov(mat3))%*%(dig-cov(mat3))
*(length(index3)/length(y))+(dig-cov(mat4))%*%
(dig-cov(mat4))*(length(index4)/length(y))+
(dig-cov(mat5))%*%(dig-cov(mat5))*(length(index5)/length(y))
mat[,i]=eigen(vhat)$vectors[,1]
}



newy[!complete.cases(newy),]=0
beta=solve(t(newy)%*%newy)%*%t(newy)%*%cbind(x1,x2,x3,x4,x5)



#####################################################################
#           computing eigendecomposition for phd candidate          #
#                        matrix on the Z-scale                      #
#                        y: discretized response                    #
#####################################################################

j=1
i=1
sigma=(y[i]-mean(y))*(z[,i]-apply(z,1,mean))
%*%t(z[,i]-apply(z,1,mean))/100
sigmahat=sigmahat+sigma
}

result[,j]=eigen(sigmahat)$vectors[,1:2]
}
```

```
#####################################################################
#           computing eigendecomposition for cr candidate           #
#                        matrix on the Z-scale                      #
#                        y: discretized response                    #
#####################################################################

i=1
j=1
matrixx=matrix(c(rep(0,25)),5,5)
for(i in 1:100){
  for(j in i:100){
    if(abs(y[i]-y[j])<=0.1){
      matrixx=(z[,i]-z[,j])%*%t(z[,i]-z[,j])+matrixx
    }
  }
}

comb = function(n, x) {
  return(factorial(n) / (factorial(x) * factorial(n-x)))
}
matt=matrixx/comb(100,3)

eigen(matt)

almost=((solve(sigma))^1/2)%*%matt%*%((solve(sigma))^1/2)
```

Appendix B

Computer Codes for Sufficient Graphical Models

```
#################################################################
######      some useful functions      ########################
#################################################################

matpower = function(a,...){
  a = (a + t(a))/2
  tmp = eigen(a)
  return(tmp[[2]]%*%diag((tmp[[1]])^alpha)%*%
           t(tmp[[2]]))}

CalGam=function(...){
  if(is.vector(A)){
    n = length(A)}
  else n = dim(A)[1]
  tmp=rowSums(as.matrix(A*A))%*%t(rep(1,n))
  K=as.numeric(tmp+t(tmp)-2*A%*%t(A))
  K=K*(K>=0)
```

```
  tou=sum(sqrt(K))/(n*(n-1))
  gam=1/(2*tou^2)
  return(gam)
}

KGaussian=function(gamma,...){
  if(is.vector(A)){
    n = length(A)}
  else n = dim(A)[1]
  if(is.vector(B)){
    m = length(B)}
  else m = dim(B)[1]
  tmp_1=rowSums(as.matrix(A*A))%*%matrix(1,1,m)
  tmp_2=rowSums(as.matrix(B*B))%*%matrix(1,1,n)
  K=tmp_1+t(tmp_2)-2*A%*%t(B)
  K=exp(-K*gamma)
  return(K)
}


#################################################################
#          SUBROUTINE: Moore-Penrose type power                 #
#            Taking power ignoring 0 eigenvalues;               #
#                 ignoring criterion=ignore                     #
#################################################################
mppower = function(matrix,power,...){
  eig = eigen(matrix)
  eval = eig[[1]]
  evec = eig[[2]]
  m = length(eval[abs(eval)>ignore])
  tmp = evec[,1:m]%*%diag(eval[1:m]^power)%*%
    t(evec[,1:m])
  return(tmp)
```

```
}

Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}



#################################################################
#          Conjoined Conditional Covariance  Operator          #
#################################################################


ccco<-function(a,b,x,...){

  gamy = CalGam(res)*sgamy
  gamx = CalGam(pred)*sgamx

  kres = KGaussian(gamy,res,res)
  kpred = KGaussian(gamx,pred,pred)

  gsirsave = generalsir(kpred,kres,...)


  gsirpred = kpred %*% qmat %*% generalsirsave[[1]][,1]


  gamx2 = CalGam(xj)*egamy
  gamgsirpred = CalGam(gsirpred)*egamx

  kx2 = KGaussian(gamx2,xj,xj)
  kgsirpred = KGaussian(gamgsirpred,gsirpred,gsirpred)
```

```
  gramx2=qmat%*%kx2%*%qmat/n
  gramgsirpred=qmat%*%kgsirpred%*%qmat/n

  gamx1 = CalGam(xi)*egamy
  kx1 = KGaussian(gamx1,xi,xi)
  gramx1=qmat%*%kx1%*%qmat/n

  ...
}
```

Bibliography

Bach, F. R. (2008), "Consistency of the group lasso and multiple kernel learning," *Journal of Machine Learning Research*, 9, 1179–1225.

Bach, F. R. and Jordan, M. I. (2002), "Kernel independent component analysis," *Journal of Machine Learning Research*, 3, 1–48.

Bellman, R. (1961), "Curse of dimensionality," *Adaptive control processes: a guided tour. Princeton, NJ.*

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993), *Efficient and adaptive estimation for semiparametric models*, vol. 2, Springer New York.

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

Bondell, H. D. and Li, L. (2009), "Shrinkage inverse regression estimation for model-free variable selection," *Journal of the Royal Statistical Society. Series B.*, 71, 287–299.

Bosq, D. (2012), *Linear processes in function spaces: theory and applications*, vol. 149, Springer Science & Business Media.

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

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

Cook, R. D. (1994), "Using dimension-reduction subspaces to identify important inputs in models of physical systems," in *Proceedings of the section on Physical*

and Engineering Sciences, American Statistical Association Alexandria, VA, pp. 18–25.

— (1998a), "Principal Hessian directions revisited," *Journal of the American Statistical Association*, 93, 84–94.

— (1998b), *Regression graphics: Ideas for studying regressions through graphics*, Wiley, New York.

Cook, R. D. and Li, B. (2002), "Dimension reduction for conditional mean in regression," *Annals of Statistics*, 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, 410–428.

Cook, R. D. and Weisberg, S. (1991), "Comment," *Journal of the American Statistical Association*, 86, 328–332.

Dawid, A. P. (1979), "Conditional independence in statistical theory," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–31.

Duan, N. and Li, K.-C. (1991), "Slicing regression: a link-free regression method," *The Annals of Statistics*, 505–530.

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

Fellinghauer, B., Bühlmann, P., Ryffel, M., Von Rhein, M., and Reinhardt, J. D. (2013), "Stable graphical model estimation with random forests for discrete, continuous, and mixed variables," *Computational Statistics & Data Analysis*, 64, 132–152.

Fernholz, L. T. (2012), *Von Mises calculus for statistical functionals*, vol. 19, Springer Science & Business Media.

Ferraty, F. and Vieu, P. (2006), *Nonparametric functional data analysis: theory and practice*, Springer Science & Business Media.

Ferré, L. and Yao, A.-F. (2003), "Functional sliced inverse regression analysis," *Statistics*, 37, 475–488.

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, 9, 432–441.

Fukumizu, K., Bach, F. R., and Gretton, A. (2007), "Statistical consistency of kernel canonical correlation analysis," *Journal of Machine Learning Research*, 8, 361–383.

Fukumizu, K., Bach, F. R., and Jordan, M. I. (2004), "Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces," *Journal of Machine Learning Research*, 5, 73–99.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008), "Kernel measures of conditional dependence," in *Advances in neural information processing systems*, pp. 489–496.

Golub, G. H., Heath, M., and Wahba, G. (1979), "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter," *Technometrics*.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2010), "Pairwise Variable Selection for High-Dimensional Model-Based Clustering," *Biometrics*, 66, 793–804.

Hansen, L. P. (1982), "Large sample properties of generalized method of moments estimators," *Econometrica: Journal of the Econometric Society*, 1029–1054.

Hansen, L. P., Heaton, J., and Yaron, A. (1996), "Finite-sample properties of some alternative GMM estimators," *Journal of Business & Economic Statistics*, 14, 262–280.

He, X., Fu, B., and Fung, W. (2003), "Median regression for longitudinal data," *Statistics in Medicine*, 22, 3655–3669.

Horváth, L. and Kokoszka, P. (2012), *Inference for functional data with applications*, vol. 200, Springer Science & Business Media.

Hsing, T. and Eubank, R. (2015), *Theoretical foundations of functional data analysis, with an introduction to linear operators*, John Wiley & Sons.

Hyvärinen, A., Karhunen, J., and Oja, E. (2004), *Independent component analysis*, vol. 46, John Wiley & Sons.

Kim, K. and Li, B. (2019a), "Functional sufficient graphical models with application to f-MRI and EEG dataset," *In Progress*.

— (2019b), "On post dimension reduction statistical inference for functional data analysis," *In Progress*.

Kim, K., Li, B., Yu, Z., and Li, L. (2019), "On post dimension reduction statistical inference," *to appear in the Annals of Statistics*.

— (2020), "On post dimension reduction statistical inference," to appear in *The Annals of Statistics*.

Lam, C. and Fan, J. (2009), "Sparsistency and rates of convergence in large covariance matrix estimation," *Annals of statistics*, 37, 4254.

Lauritzen, S. L. (1996), *Graphical models*, vol. 17, Clarendon Press.

Lee, K.-Y., Li, B., and Chiaromonte, F. (2013), "A general theory for nonlinear sufficient dimension reduction: Formulation and estimation," *The Annals of Statistics*, 41, 221–249.

Lee, K.-Y., Li, B., and Zhao, H. (2016a), "On an additive partial correlation operator and nonparametric estimation of graphical models," *Biometrika*, 103, 513–530.

— (2016b), "Variable selection via additive conditional independence," *Journal of the Royal Statistical Society: Series B*, 78, 1037–1055.

Li, B. (1993), "A deviance function for the quasi-likelihood method," *Biometrika*, 80, 741–753.

— (2018a), "Linear operator-based statistical analysis: A useful paradigm for big data," *Canadian Journal of Statistics*, 46, 79–103.

— (2018b), *Sufficient Dimension Reduction; Methods and Applications with R*, CRC Press.

— (2018c), *Sufficient Dimension Reduction: Methods and Applications with R*, CRC Press.

Li, B., Artemiou, A., and Li, L. (2011a), "Principal support vector machines for linear and nonlinear sufficient dimension reduction," *The Annals of Statistics*, 39, 3182–3210.

— (2011b), "Principal support vector machines for linear and nonlinear sufficient dimension reduction," *The Annals of Statistics*, 39, 3182–3210.

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

Li, B. and Kim, K. (2019), "On a sufficient nonparametric graphical models with application to gene network," *In Progress*.

Li, B. and Solea, E. (2018a), "A nonparametric graphical model for functional data with application to brain networks based on fMRI," *Journal of the American Statistical Association*, 113, 1637–1655.

— (2018b), "A nonparametric graphical model for functional data with application to brain networks based on FMRI," *Journal of the American Statistical Association*, 113, 1637–1655.

Li, B. and Song, J. (2017a), "Nonlinear sufficient dimension reduction for functional data," *The Annals of Statistics*, 45, 1059–1095.

— (2017b), "Nonlinear sufficient dimension reduction for functional data," *The Annals of Statistics*, 45, 1059–1095.

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

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

Li, G.-R., Zhu, L.-P., and Zhu, L.-X. (2010), "Adaptive confidence region for the direction in semiparametric regressions," *Journal of Multivariate Analysis*, 101, 1364–1377.

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

— (1991b), "Sliced inverse regression for dimension reduction," *Journal of the American Statistical Association*, 86, 316–327.

— (1992), "On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma," *Journal of the American Statistical Association*, 87, 1025–1039.

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

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

— (2018d), *Sufficient Dimension Reduction*, American Cancer Society, pp. 1–8.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012a), "High-dimensional semiparametric Gaussian copula graphical models," *The Annals of Statistics*, 40, 2293–2326.

— (2012b), "The nonparanormal skeptic," *arXiv preprint arXiv:1206.6488*.

Liu, H., Han, F., and Zhang, C.-H. (2012c), "Transelliptical graphical models," in *Advances in neural information processing systems*, pp. 800–808.

Liu, H., Lafferty, J., and Wasserman, L. (2009), "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs," *Journal of Machine Learning Research*, 10, 2295–2328.

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

Ma, Y. and Zhu, L. (2012), "A semiparametric approach to dimension reduction," *Journal of the American Statistical Association*, 107, 168–179.

— (2013), "Efficient estimation in sufficient dimension reduction," *Annals of statistics*, 41, 250.

— (2014), "On estimation efficiency of the central mean subspace," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 885–901.

Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. (2010), "Revealing strengths and weaknesses of methods for gene network inference," *Proceedings of the national academy of sciences*, 107, 6286–6291.

McCullagh, P. and Nelder, J. (1989), "Nelder. JA (1989), Generalized Linear Models," *CRC Monographs on Statistics & Applied Probability, Springer Verlag, New York*.

Meinshausen, N. and Bühlmann, P. (2006), "High-dimensional graphs and variable selection with the lasso," *The annals of statistics*, 1436–1462.

Milham, M. P., Fair, D., Mennes, M., Mostofsky, S. H., et al. (2012), "The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience," *Frontiers in systems neuroscience*, 6, 62.

Owen, A. (1990), "Empirical likelihood ratio confidence regions," *The Annals of Statistics*, 90–120.

Owen, A. B. (1988), "Empirical likelihood ratio confidence intervals for a single functional," *Biometrika*, 75, 237–249.

Park, C. and Lindsay, B. G. (1999), "Robust estimation and tests based on quadratic inference function," Tech. rep., Technical Report.

Pearl, J. and Verma, T. (1987), *The logic of representing dependencies by directed graphs*, University of California (Los Angeles). Computer Science Department.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009), "Partial correlation estimation by joint sparse regression models," *Journal of the American Statistical Association*, 104, 735–746.

Pinna, A., Soranzo, N., and de la Fuente, A. (2010), "From knockouts to networks: establishing direct cause-effect relationships through graph analysis," *PLoS One*, 5.

Qiao, X., Guo, S., and James, G. M. (2018), "Functional graphical models," *Journal of the American Statistical Association*, 1–12.

— (2019), "Functional graphical models," *Journal of the American Statistical Association*, 114, 211–222.

Qu, A., Lindsay, B. G., and Li, B. (2000), "Improving generalised estimating equations using quadratic inference functions," *Biometrika*, 87, 823–836.

Ramsay, J. O. and Silverman, B. W. (2005), "Functional data analysis," .

— (2007), *Applied functional data analysis: methods and case studies*, Springer.

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

Smola, A. J. and Scholkopt, B. (2004), "A tutorial on support vector regression," *Statistics and Computing*, 3, 199–122.

Solea, E. and Li, B. (2020), "Copula Gaussian graphical models for functional data," submitted to *Journal of the American Statistical Association*.

Song, J., Kim, K., and Li, B. (2020), "Sparse Principal Component Analysis for multivariate functional data," *In Progress*.

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

Van der Vaart, A. W. (1998), *Asymptotic statistics*, vol. 3, Cambridge university press.

Voorman, A., Shojaie, A., and Witten, D. (2013), "Graph estimation with joint additive models," *Biometrika*, 101, 85–101.

Wang, H. J. and Wang, L. (2009), "Locally weighted censored quantile regression," *Journal of the American Statistical Association*, 104, 1117–1128.

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

Wang, Y. (2008), "Nonlinear dimension reduction in feature space," *PhD Thesis, The Pennsylvania State University*.

Wu, H. M. (2008), "Kernel sliced inverse regression with applications to classification," *Journal of Computational and Graphical Statistics*, 17, 590–610.

Xia, M., Wang, J., and He, Y. (2013), "BrainNet Viewer: a network visualization tool for human brain connectomics," *PloS one*, 8, e68910.

Xia, Y. (2007), "A constructive approach to the estimation of dimension reduction directions," *The Annals of Statistics*, 35, 2654–2690.

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

Xue, L. and Zou, H. (2012), "Regularized rank-based estimation of high-dimensional nonparanormal graphical models," *The Annals of Statistics*, 40, 2541–2571.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005), "Functional data analysis for sparse longitudinal data," *Journal of the American Statistical Association*, 100, 577–590.

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, 1733–1757.

Yuan, M. and Lin, Y. (2007), "Model selection and estimation in the Gaussian graphical model," *Biometrika*, 94, 19–35.

Zhu, H., Strawn, N., and Dunson, D. B. (2016), "Bayesian graphical models for multivariate functional data," *The Journal of Machine Learning Research*, 17, 7157–7183.

Zhu, L.-P., Li, L., Li, R., and Zhu, L.-X. (2011), "Model-free feature screening for ultrahigh-dimensional data," *Journal of the American Statistical Association*, 106, 1464–1475.

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

Zhu, L.-X. and Fang, K.-T. (1996), "Asymptotics for kernel estimate of sliced inverse regression," *The Annals of Statistics*, 24, 1053–1068.

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

Vita

Kyongwon Kim

Education

- Ph.D. in Statistics
  The Pennsylvania State University, University Park, PA
- B.S. in Mathematics (Summa Cum Laude)
  Sogang University, Seoul, South Korea

Honors and Awards

- William Harkness Teaching Award, The Pennsylvania State University, Dec 2019
- American Statistical Association (ASA) Best Student Paper Award in the Business and Economic Statistics Section, Aug 2019
- William Harkness Travel Award, The Pennsylvania State University, Aug 2018, 2019
- National Institute of Statistical Sciences (NISS) Travel Award, May 2019
- Conference Board of the Mathematical Sciences Travel Award, Aug 2018
- Graduate Assistantship, The Pennsylvania State University, 2015 – Present
- Dean's List, Sogang University, Aug 2012, 2013
- High Honor Scholarship, Sogang University, 2011 – 2015

Publications

- Kyongwon Kim., Bing Li., Zhou Yu., and Lexin Li. (2019), "On Post Dimension Reduction Statistical Inference" (Accepted for the publication in the *Annals of Statistics*)
- Bing Li., Kyongwon Kim. (2019), "On Sufficient Graphical Models" (Submitted)