MODEL-FREE VARIABLE SELECTION FOR TWO GROUPS OF VARIABLES

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY
IN STATISTICS

---

by
Ahmad Alothman
August, 2018

Examining Committee Members:
Dr. Yuexiao Dong, Advisory Chair, Department of Statistical Science
Dr. Cheng-Yong Tang, Department of Statistical Science
Dr. Pallavi Chitturi, Department of Statistical Science
Dr. Cencheng Shen, Center for Imaging Science, Johns Hopkins University

ABSTRACT

In this dissertation we introduce two variable selection procedures for multivariate responses. Our procedures are based on sufficient dimension reduction concepts and are model-free. In the first procedure we consider the dual marginal coordinate hypotheses, where the role of the predictor and the response is not important. Motivated by canonical correlation analysis (CCA), we propose a CCA-based test for the dual marginal coordinate hypotheses, and devise a joint backward selection algorithm for dual model-free variable selection. The second procedure is based on ordinary least squares (OLS). We derive and study the asymptotic properties of the OLS-based test under the normality assumption of the predictors as well as an asymmetry assumption. When these assumptions are violated, the asymptotic test with elliptical trimming and clustering is still valid with desirable numerical performances. A backward selection algorithm for the predictor is also provided for the OLS-based test. The performances of the proposed tests and the variable selection procedures are evaluated through synthetic examples and a real data analysis.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to Dr. Yuexiao Dong for the constant encouragement and guidance he gave me during my time at Temple. I consider myself lucky to have had Dr. Dong as my academic advisor. Though things might get complicated, he always finds a way to simplify them and make them enjoyable. His advice and patience helped me discover my strengths and improve my weaknesses.

I would also like to express my gratitude to my committee members, Dr. Cheng-Yong Tang, Dr. Pallavi Chitturi, and Dr. Cencheng Shen for their guidance and constructive comments.

Finally I would like to thank my caring, loving and supportive wife, Dalal: thank you. I would also like to thank my daughter, and my family for their tremendous support and motivation. Lastly I would like to express my deepest gratitude to my parents, especially to my father who had always pushed me to work harder and be a better person, father: I will always be indebted to you.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Overview

Variable selection refers to the process of eliminating unnecessary and redundant variables from a statistical model. This problem has become increasingly important, especially with the plethora of information easily available in the time we live in, where dataset could easily contain hundreds and even thousands of variables. Reducing the number of variables in a statistical model may be desired to produce simpler and more interpretable models, reduce computation time, eliminate collinearity, or enhance a model's predictive ability.

To achieve the goal of reducing the number of covariates in a model we have two options. The first option is variable selection, where a subset of the covariates is selected to be included in the final model. Only these covariates are considered to be related to the response. In statistical literature, classical variable selection procedures are mainly stepwise selection and best subset selection. More recently, LASSO regression (Tibshirani, 1996) was introduced, a penalized least squares minimization technique that sets the sum of the absolute value of the regression coefficients to be less than or equal a predetermined constant. Many statistical techniques were inspired by the idea of penalizing the least squares. Such techniques include SCAD (Fan and Li, 2001), elastic net (Zou and Hastie, 2005), adaptive LASSO or ALASSO (Zou, 2006), Dantzig selector (Candes and Tao, 2007), SICA (Li and Fan, 2009), and MCP (Zhang, 2010). An immediate shortcoming of the aforementioned methods is that they assume a linear relation between the response and the predictors.

The second option is sufficient dimension reduction (SDR), an approach that aims to find linear combinations of the covariates that are related to the response variable. Since the introduction of sliced inverse regression (SIR) in Li (1991), many sufficient dimension reduction methods have been proposed in the literature. See, for example, sliced average variance estimation (SAVE) (Cook and Weisberg, 1991), principal Hessian directions (PHD, 1992), and directional regression (DR) (Li and Wang, 2007). In the seminal work of Cook (2004), marginal coordinate test is proposed for testing predictor contributions under the sufficient dimension reduction framework. Without assuming the link function between the response and the predictors, one can test the contribution of one predictor to the univariate response in the presence of all the other predictors. The original test in Cook (2004) is based on SIR, and a similar test based on SAVE is studied in Shao et al. (2007). Yu et al. (2010) and Yu and Dong (2016) further extend the marginal coordinate test through PHD and DR, respectively. All the aforementioned tests deal with univariate response. An important link between sufficient dimension reduction and model-free variable selection is elucidated in Yu et al. (2016), where popular sufficient dimension reduction methods such as SIR, SAVE, and DR are used to construct corresponding model-free variable selection procedures. Unlike model-based variables selection procedures such as stepwise regression, variable selection through SDR is model-free and does not require assuming specific models between the predictors and the response.

In this dissertation, we study marginal coordinate tests and model-free variable selection for multivariate responses. Consider a predictor $\mathbf{x} \in \mathbb{R}^p$ and a multivariate response $\mathbf{y} \in \mathbb{R}^q$. Our contribution is two-fold. First, we propose a dual-model free variable selection procedure for $\mathbf{x}$ and $\mathbf{y}$ simultaneously. The proposed procedure is based on canonical correlation analysis (CCA) (Hotelling, 1936), where the role of the predictor and the response is not important. This viewpoint motivates us to consider dual variable selection,

where the goal is to simultaneously identify the important variables among $\mathbf{x}$ for the prediction of $\mathbf{y}$ and the important variables among $\mathbf{y}$ for the prediction of $\mathbf{x}$.

Second, we propose a marginal coordinate test based on ordinary least squares (OLS). The asymptotic distribution of the sample level test statistic is carefully studied. Instead of assuming the specific forms of the link function between the predictor and the response, the asymptotic test is valid under the normality assumption of $\mathbf{x}$ as well as an asymmetry assumption. We also provide practical solutions when these assumptions are violated. When the normality assumption is violated, we demonstrate that the asymptotic test with elliptical trimming is still applicable. In the case when the asymmetry assumption is violated, clustering and elliptical trimming can be used to preprocess the data before applying the asymptotic test.

Throughout this dissertation, denote $\mathbb{R}^p$ as the set of all real numbers and $\mathbb{R}^p$ as the p-dimensional Euclidean space. For random variables, we use upper case letters such as X or Y, and a random vector by a bold lower case like $\mathbf{x}$ or $\mathbf{y}$. Scalers are denoted by lower case letters (a), vectors by bold lower case letters ($\mathbf{a}$), and matrices by bold upper case letters ($\mathbf{A}$).

1.2 Central Mean Space and Central Space

Consider a univariate response $Y \in \mathbb{R}$ and p-dimensional predictor $\mathbf{x} = (X_1, X_2, \cdots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$, dimension reduction (Li, 1991) aims to find d linear combinations of $\mathbf{x}$, while preserving the regression information of Y on $\mathbf{x}$. Alternatively, we can say that we want to find a $p \times d$ matrix $\mathbf{B}$ such that Y is independent of $\mathbf{x}$ given $\mathbf{B}^{\mathrm{T}}\mathbf{x}$,

$$Y \perp\!\!\!\perp \mathbf{x} | \mathbf{B}^{\mathrm{T}}\mathbf{x}, \tag{1.1}$$

where $\perp\!\!\!\perp$ means statistical independence. Note that for any $d \times d$ non-singular matrix $\mathbf{A}$, $\mathbf{BA}$ also satisfies (1.1), i.e. $Y \perp\!\!\!\perp \mathbf{x} | (\mathbf{BA})^{\mathrm{T}}\mathbf{x}$. Therefore, $\mathbf{B}$ and $\mathbf{A}$ have the same columns space. We define the column space of $\mathbf{B}$ as the dimension reduction subspace (DRS; Cook, 1994, 1996). Denote $\mathcal{S}(\mathbf{B})$ as the subspace spanned by the columns of $\mathbf{B}$.

Definition 1.2.1 *The central space (CS), denoted by $\mathcal{S}_{Y|\mathbf{x}}$, for $(\mathbf{x}, Y)$ is the intersection of all dimension reduction spaces for $(\mathbf{x}, Y)$.*

Note that in the literature the central space $\mathcal{S}_{\mathbf{x}|Y}$ may also be called central subspace, sufficient dimension reduction subspace, or effective dimension reduction (e.d.r.) central space. The dimension d of $\mathcal{S}_{\mathbf{x}|Y}$ is called the structure dimension. In applications we hope that d is much smaller than p.

Depending on the application, sometimes we are interested in the regression conditional mean function instead of the conditional distribution of Y given $\mathbf{x}$. Sufficient dimension reduction can be very useful to lower the dimensionality of the predictor before applying regression analysis to the data. Therefore, we introduce the concept of central mean space (CMS; Cook & Li, 2002). Consider

$$E(Y|\mathbf{x}) \perp\!\!\!\perp Y | \mathbf{C}^{\mathrm{T}}\mathbf{x}, \tag{1.2}$$

For any $p \times d$ matrix $\mathbf{C}$ that satisfies (1.2), we refer to its column space, $\mathcal{S}(\mathbf{C})$, as the mean dimension reduction space for the regression of Y on $\mathbf{x}$.

Definition 1.2.2 *The central mean space (CMS), denoted by $\mathcal{S}_{E(\mathbf{x}|Y)}$, for the regression of Y on $\mathbf{x}$ is the intersection of all mean dimension reduction spaces, given that the intersection itself is a mean dimension reduction space.*

Since $E(Y|\mathbf{x}) \perp\!\!\!\perp \mathbf{x} | \mathbf{B}^{\mathrm{T}}\mathbf{x}$ is implied by $Y \perp\!\!\!\perp \mathbf{x} | \mathbf{B}^{\mathrm{T}}\mathbf{x}$, a dimension reduction subspace is also a mean reduction subspace. Therefore, the central mean space is a subset of the central space, i.e. $\mathcal{S}_{E(Y|\mathbf{x})} \subseteq \mathcal{S}_{Y|\mathbf{x}}$.

A very useful property of the central space is the invariance property (Cook, 1998).

Proposition 1.2.1 *Let $\mathcal{S}_{Y|\mathbf{x}}$ be the central mean space for $(\mathbf{x}, Y)$, $\mathbf{w} = \mathbf{A}\mathbf{x} + \mathbf{b}$, where $\mathbf{A}$ is a $p \times p$ non-singular matrix and $\mathbf{b} \in \mathbb{R}^p$ is a constant vector. Then*

$$\mathcal{S}_{E(Y|\mathbf{w})} = \mathbf{A}^{-1}\mathcal{S}_{E(Y|\mathbf{x})}$$

This property allows us to work with the standardized $\mathbf{z}$ predicator instead of working with $\mathbf{x}$ directly. For example, for the predictor $\mathbf{x}$, let $\boldsymbol{\mu} = E(\mathbf{x})$ be the mean vector and $\boldsymbol{\Sigma} = \text{var}(\mathbf{x})$ be the covariance matrix. We can make the standardization $\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$, obtain the central mean space for the standardized predictor $\mathcal{S}_{E(Y|\mathbf{z})}$, and then transform it back to $\mathbf{x}$-scale using $\mathcal{S}_{E(Y|\mathbf{x})} = \boldsymbol{\Sigma}^{-1/2}\mathcal{S}_{E(Y|\mathbf{z})}$.

1.3 Popular Sufficient Dimension Reduction Methods

Sufficient dimension reduction methods can generally be categorized into either inverse conditional moments (or just moment-based) methods or nonparametric methods. Nonparametric methods usually involve multi-dimensional smoothing to obtain the central mean space (or central space). Moment-based methods are among the most popular in the literature, and they require working directly with the moment functions. Methods in this category deal with both first and second order moments (or inverse conditional moments). Many methods were developed that deal with the first order moments including Ordinary Least Squares (OLS; K. C. Li & Duan, 1989), Sliced Inverse Regression (SIR; K. C. Li, 1991), Canonical Correlation (Fung, He, Liu, & Shi, 2002). These methods deal with

$$E(\mathbf{x}Y^k), E(\mathbf{x}|Y),$$

where k is an integer. Note that in $E(\mathbf{x}|Y)$, the roles are switched for $\mathbf{x}$ and Y, unlike the traditional regression setting. The literature for sufficient dimension reduction methods

that address the second order moment include Principal Hessian Directions (PHD; K. C. Li, 1992 and Cook, 1998), Sliced Average Variance Estimator (SAVE; Cook & Weisberg, 1991), and Directional Regression (DR; B. Li & Wang, 2007). These methods depend on

$$E(\mathbf{x}Y^k), E(\mathbf{x}|Y), E(\mathbf{x}\mathbf{x}^{\mathrm{T}}Y^k), E(\mathbf{x}\mathbf{x}^{\mathrm{T}}|Y),$$

where k is an integer.

Throughout this proposal, for simplicity of derivation, we assume $E(\mathbf{x}) = \mathbf{0}$. Before going into the details of sufficient dimension reduction methods we will introduce some standard assumptions in the literature. First order moment approaches require the linear conditional mean (LCM) assumption, that is

Assumption 1.3.1 *Suppose that* $\mathbf{B}$ *is a basis for the central space* $\mathcal{S}_{Y|\mathbf{x}}$*, i.e. span*$(\mathbf{B}) = \mathcal{S}_{Y|\mathbf{x}}$*. Then, we assume that* $E(\mathbf{x}|\mathbf{B}^T\mathbf{x})$ *is linear function of* $\mathbf{x}$*.*

The second order moment approaches require the linear conditional mean assumption as well as the constant conditional variance (CCV) assumption,

Assumption 1.3.2 *Suppose that* $\mathbf{B}$ *is a basis for the central space* $\mathcal{S}_{Y|\mathbf{x}}$*. Then, we assume that* $Var(\mathbf{x}|\mathbf{B}^T\mathbf{x})$ *is a nonrandom matrix.*

A parallel version of the LCM and CCV assumptions in (1.3.1) and (1.3.2), receptively, can be defined for the central mean space. Without ambiguity, we will refer to assumptions (1.3.1) and (1.3.2) when LCM and CCV, receptively, are required in the central mean space.

1.3.1 Ordinary Least Squares

In statistical applications, Ordinary Least Squares (OLS) is one of most popular techniques to estimate regression coefficients. However, OLS was first used as a tool for sufficient

dimension reduction by Li and Duan (1989). In a traditional linear regression setting we have

$$Y = f(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) + \epsilon,$$

where $f(\cdot)$ is the link function, $\alpha \in \mathbb{R}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}} \in \mathbb{R}^p$, ϵ is independent of $\mathbf{x}$. In applications, if $E(\mathbf{x}) = \boldsymbol{\mu}$ and $Var(\mathbf{x}) = \boldsymbol{\Sigma}$, we can make the following transformation $\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$. For the purpose of simple derivation, assume that $E(\mathbf{x}) = \mathbf{0}$ and $Var(\mathbf{x}) = \mathrm{I}_p$, then the OLS estimator is given by

$$\boldsymbol{\beta}_{\mathrm{OLS}} = E(\mathbf{x}\mathbf{x}^{\mathrm{T}})^{-1}E(\mathbf{x}(Y - E(Y))) = E(\mathbf{x}Y),$$

where $E(\mathbf{x}\mathbf{x}^{\mathrm{T}}) = \mathrm{I}_p$ and $E(Y) = \mathbf{0}$. Li and Duan (1989) showed how OLS is related to the concept of sufficient dimension reduction. Cook and Li (2002) proved that OLS estimates belong to the central mean space, which in SDR literature is called single-index model.

Proposition 1.3.1 *Suppose that $E(\mathbf{x}) = \mathbf{0}$ and $Var(\mathbf{x}) = \mathrm{I}_p$, under the LCM assumption (1.3.1), the OLS estimator is in the central mean space, that is $E(\mathbf{x}Y) \in \mathcal{S}_{E(Y|\mathbf{x})}$*

In a similar fashion we can extend proposition (1.3.1) to the case when $Var(\mathbf{x}) = \boldsymbol{\Sigma}$. Suppose that the projection matrix is $\mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathrm{T}}$, then $E(\mathbf{x}Y) = \mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})E(\mathbf{x}Y)$, and the OLS estimator is in the central mean space $\mathcal{S}_{E(Y|\mathbf{x})}$, that is

$$\boldsymbol{\beta}_{\mathrm{OLS}} = \boldsymbol{\Sigma}^{-1}E(\mathbf{x}Y) \in \mathcal{S}_{E(Y|\mathbf{x})}$$

Let $\{(\mathbf{x}_1, Y_1), \cdots, (\mathbf{x}_n, Y_n)\}$ be an i.i.d sample, where $\mathbf{x}_i \in \mathbb{R}^p$. To estimate the central mean space using OLS, we use the following procedure

1. Let $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i$ be the sample mean of $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^{\mathrm{T}}$. Calculate $\widetilde{\mathbf{x}}_i = \mathbf{x}_i - \hat{\boldsymbol{\mu}}$, $\widetilde{Y}_i = Y_i - \bar{Y}$, and let $\widehat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^{n}\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^{\mathrm{T}}$ be the sample covariance of $\mathbf{X}$.

2. To get and estimator in $\mathcal{S}_{E(\mathbf{x}|Y)}$ we use the OLS estimator $\hat{\boldsymbol{\beta}} = (n\widehat{\boldsymbol{\Sigma}})^{-1} \sum_{i=1}^{n} \widetilde{\mathbf{x}}_i \widetilde{Y}_i$.

The OLS estimator for $\mathcal{S}_{E(Y|\mathbf{x})}$ is $\sqrt{n}$-consistent, i.e. it converges at $\sqrt{n}$-rate to a vector that belongs to the central mean space $\mathcal{S}_{E(Y|\mathbf{x})}$.

An obvious shortcoming of sufficient dimension reduction through OLS is its limitation to allowing us to estimate only one direction in the central mean space, in the case of univariate response variable Y. In the case where we have multivariate response variables (Multi-Response), the number of directions that can be estimated is bounded by the dimension of the response variable. Li and Duan (1989) showed that we are guaranteed that $\boldsymbol{\beta}_{\text{OLS}}$ is proportional to the true parameter $\boldsymbol{\beta}$. However, the proportionality of the OLS estimator to the true parameter could fail when the link function $f(.)$ is symmetric about zero.

1.3.2 Sliced Inverse Regression

Sliced Inverse Regression (SIR), introduced by Li (1991), is one of the most popular and widely used dimension reduction methods. It is based on the idea that instead of regressing Y on $\mathbf{x}$, we switch roles by regressing $\mathbf{x}$ on Y. Inverse regression allows us to regress each predictor in $\mathbf{x}$ on Y, separately. One benefit of switching the roles of $\mathbf{x}$ and Y is that rather than dealing with a high dimensional regression problem (forward regression), we can work with p one-demension to one-dimension inverse regressions, given that $\mathbf{x}$ is p dimensional. Li (1991) showed that $E(\mathbf{x}|Y = y)$ is in the central space $\mathcal{S}_{Y|\mathbf{x}}$.

Proposition 1.3.2 *Suppose that $E(\mathbf{x}) = \mathbf{0}$ and $Var(\mathbf{x}) = I_p$, under the LCM assumption (1.3.1)*

$$E(\mathbf{x}|Y = y) \in \mathcal{S}_{Y|\mathbf{x}}$$

SIR does not work with continuous Y directly, rather, it uses a discretized version of Y. We will divide Y into H equally spaced partitions, $J_1, \ldots, J_h$. We can show that a natural extension to proposition (1.3.2) is that $Var(E(\mathbf{x}|Y))$ is a subspace of the central space

$\mathcal{S}_{Y|\mathbf{x}}$. Let $p_h = Pr(Y \in J_h)$, and $\mathbf{m}_h = E(\mathbf{x}|Y \in J_h)$, then we define $V = Var(E(\mathbf{x}|Y))$ where

$$Var(E(\mathbf{x}|Y)) = \sum_{i=1}^{H} Pr(Y \in J_h) E(\mathbf{x}|Y \in J_h) E^{\mathrm{T}}(\mathbf{x}|Y \in J_h).$$

Let $\{(\mathbf{x}_1, Y_1), \cdots, (\mathbf{x}_n, Y_n)\}$ be an i.i.d sample, where $\mathbf{x}_i \in \mathbb{R}^p$. To estimate the central space using SIR, we use the following procedure

1. Let $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i$ be the sample mean of $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^{\mathrm{T}}$. Calculate $\widetilde{\mathbf{x}}_i = \mathbf{x}_i - \hat{\boldsymbol{\mu}}$, and let $\widehat{\boldsymbol{\Sigma}} = n^{-1} \sum_{i=1}^{n} \widetilde{\mathbf{x}}_i \widetilde{\mathbf{x}}_i^{\mathrm{T}}$ be the sample covariance of $\mathbf{X}$.

2. Standardize $\mathbf{x}_1, \cdots, \mathbf{x}_n$ such that $\widehat{\mathbf{Z}} = \widehat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{X} - \hat{\boldsymbol{\mu}})$, or alternatively $\hat{\mathbf{z}}_i = \widehat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})$.

3. Divide the range of Y into H slices, $J_1, \ldots, J_h$. Calculate $\hat{p}_h = 1/n \sum_{i=1}^{n} \delta_h(Y_i)$, where $\delta_h(Y_i) = 1$ if Y_i is in the hth slice, and zero otherwise.

4. For each slice h, compute the sample mean

$$\hat{\mathbf{m}}_h = \frac{1}{n\hat{p}_h} \sum_{Y_i \in J_h} \hat{\mathbf{z}}_i.$$

5. Construct the SIR matrix

$$\widehat{\mathbf{V}} = \sum_{i=1}^{H} \hat{p}_h \hat{\mathbf{m}}_h \hat{\mathbf{m}}_h^{\mathrm{T}}.$$

6. Perform eigenvalue decomposition on $\widehat{\mathbf{V}}$. Assume that d is known, and let $\mathbf{v}_1, \ldots, \mathbf{v}_d$ be the eigenvectors corresponding to the largest d eigenvalues. Then, $\boldsymbol{\beta}_* = (\mathbf{v}_1, \ldots, \mathbf{v}_d)$ is an estimator of the central space $\mathcal{S}_{\mathbf{z}|Y}$.

7. To get and estimator in $\mathcal{S}_{\mathbf{x}|Y}$, we transform back to the $\mathbf{x}$-scale, $\hat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\Sigma}}^{-1/2} \hat{\boldsymbol{\beta}}_*$.

The above procedure assumes that the structure dimension d is known. However, in applications, it is hard to justify such an assumption. To estimate the central space $\mathcal{S}_{Y|\mathbf{x}}$ accurately, we need a mechanism to determine the number of directions to be included in the final estimate. Li (1991) proposed a test statistic along with its asymptotic distribution to identify d. Let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p$ be the estimated eigenvalues, and let $\hat{\bar{\lambda}}_{(p-k)}$ be the average of the smallest $p-k$ eigenvalues obtained from $\widehat{\mathbf{V}}$ by the above procedure. We want these values to be significantly larger than zero compared to sampling error, that is we want to test the null hypothesis

$$H_0 : \lambda_{p-k} = \lambda_{p-k+1} = \ldots = \lambda_p,$$

Proposition 1.3.3 *Suppose that* $\mathbf{x}$ *is normally distributed with* $E(\mathbf{x}) = \boldsymbol{\mu}$ *and* $Var(\mathbf{x}) = \boldsymbol{\Sigma}$*, then under the LCM assumption (1.3.1)*

$$n(p-k)\hat{\bar{\lambda}}_{(p-k)} = n(p-k)\frac{\sum_{i=p-k+1}^{p}\hat{\lambda}_i}{p-k} = n\sum_{i=p-k+1}^{p}\hat{\lambda}_i$$

follows a χ^2 *distribution with* $(p-k)(H-k-1)$ *degrees of freedom.*

We can use proposition (1.3.3) to determine d using a sequential testing scheme, starting with the smallest eigenvalue.

SIR provides $\sqrt{n}$-consistent estimators. However, it will not detect the direction in the central space when the response is symmetric. Another limitation of SIR is that the number of eigenvalues greater than zero depends on the number of slices H. In the case of categorical variables, say we have three categories, the maximum number of slices is two, i.e. SIR can only produce two directions in the central space. When the response is binary, SIR is limited to only one direction.

1.3.3 Sliced Average Variance Estimator

Sliced Average Variance Estimator (SAVE) introduced by Cook & Weisberg (1991) is another method for dimension reduction that is based on the idea of inverse regression. SAVE was mainly inspired by SIR, however, rather than using the means, it uses the average variance within each slice. Like SIR, we will partition the response variable into H slices. In addition to the linear conditional mean assumption (1.3.1), SAVE also requires the constant conditional variance assumption (1.3.2). SAVE captures the central space by utilizing the second order moments, avoiding problems such as symmetric response variables.

Proposition 1.3.4 *Suppose that* $E(\mathbf{x}) = \mathbf{0}$ *and* $Var(\mathbf{x}) = \mathrm{I}_p$*, under the LCM and CCV assumptions*

$$\mathrm{I}_p - Var(\mathbf{x}|Y = y) \in \mathcal{S}_{Y|\mathbf{x}},$$

where I_p *is the identity matrix with dimension* p*.*

From proposition (1.3.4) we can see that $E\left((\mathrm{I}_p - Var(\mathbf{x}|Y))^2\right)$, the SAVE matrix, is also a subspace of $\mathcal{S}_{Y|\mathbf{x}}$, where for a matrix $\mathbf{A}$, $\mathbf{A}^2 = \mathbf{A}\mathbf{A}^{\mathrm{T}}$. Define $\mathbf{\Sigma}_h$ as the variance of $\mathbf{x}$ given Y within slice h, that is

$$\mathbf{\Sigma}_h = Var(\mathbf{x}|Y \in J_h).$$

The estimate of $SAVE$ matrix is defined as

$$E_n\left((\mathrm{I}_p - Var(\mathbf{x}|Y))^2\right) = \hat{p}_h \sum_{h=1}^{H} (\mathrm{I}_p - \widehat{\mathbf{\Sigma}}_h)(\mathrm{I}_p - \widehat{\mathbf{\Sigma}}_h)^{\mathrm{T}}, \tag{1.3}$$

where $\hat{p}_h = 1/n \sum_{i=1}^{n} \delta_h(Y_i)$, $\delta_h(Y_i) = 1$ if Y_i is in the hth slice, and zero otherwise, and $\widehat{\mathbf{\Sigma}}_h$ is given by

$$\widehat{\mathbf{\Sigma}}_h = \frac{1}{n_h - 1} \sum_{Y_i \in J_h} (\mathbf{x}_i - \bar{\mathbf{x}}_h)(\mathbf{x}_i - \bar{\mathbf{x}}_h)^{\mathrm{T}}.$$

Here $\bar{\mathbf{x}}_h$ and n_h are the mean and size of slice h, respectively. To estimate the central space we perform the same procedure given in section (1.3.2), using the $SAVE$ matrix given in equation (1.3) instead of $\widehat{\mathbf{V}}$. Note that the standardized version of the data should be used to in calculating $SAVE$. It is worth mentioning that SAVE, unlike SIR, is more sensitive to the number of slices H.

1.3.4 Principle Hessian Directions

Principle Hessian Directions (PHD), unlike OLS and SIR, can handle cases where the response is symmetric about its mean. PHD is based on the concept of the Hessian matrix $\mathbf{H}(\mathbf{x})$ of the regression function $f(\mathbf{x}) = E(Y|\mathbf{x})$. Suppose that $\mathbf{x}$ is a predictor vector with dimension p, then the Hessian matrix is defined as

$$\mathbf{H}(\mathbf{x}) = \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^{\mathrm{T}}},$$

where at any point $\mathbf{x}$, the Hessian matrix $\mathbf{H}(\mathbf{x})$ is degenerate along any directions that are orthogonal to the central mean space $\mathcal{S}_{E(Y|\mathbf{x})}$. PHD attempts to recover the CMS, and it requires both LCM and CCV assumptions. Let the average of the Hessian matrix be denoted as $\bar{\mathbf{H}}_{\mathbf{x}} = E(\mathbf{H}(\mathbf{x}))$. One can show using Stein's Lemma that the weighted covariance

$$\boldsymbol{\Sigma}_{Y\mathbf{x}\mathbf{x}} = E((Y - \mu_Y)(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^{\mathrm{T}})$$

is related to the average Hessian matrix through the identity

$$\bar{\mathbf{H}}_{\mathbf{x}} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \boldsymbol{\Sigma}_{Y\mathbf{x}\mathbf{x}} \boldsymbol{\Sigma}_{\mathbf{x}}^{-1},$$

where $E(Y) = \mu_Y$, $E(\mathbf{x}) = \boldsymbol{\mu}_{\mathbf{x}}$, and $Var(\mathbf{x}) = \boldsymbol{\Sigma}_{\mathbf{x}}$.

Proposition 1.3.5 *Suppose that* $\mathbf{x}$ *is normal with* $E(\mathbf{x}) = \mathbf{0}$ *and* $Var(\mathbf{x}) = \mathrm{I}_p$*, under the LCM and CCV assumptions*

$$\bar{\mathbf{H}}_{\mathbf{x}} \in \mathcal{S}_{E(Y|\mathbf{x})},$$

Note that subtracting a linear combination of $\mathbf{x}$ from Y does not change the Hessian matrix. Let r be the residual resulting from regressing Y on $\mathbf{x}$, using the OLS estimator. Then we can define the average Hessian matrix as

$$\bar{\mathbf{H}}_{\mathbf{x}} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\Sigma}_{r\mathbf{x}\mathbf{x}}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1},$$

where $\boldsymbol{\Sigma}_{r\mathbf{x}\mathbf{x}} = E(r(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^{\mathrm{T}}))$. Let $\{(\mathbf{x}_1, Y_1), \cdots, (\mathbf{x}_n, Y_n)\}$ be an i.i.d sample, where $\mathbf{x}_i \in \mathbb{R}^p$. To estimate the central space using PHD, we can use the following procedure

1. Estimate the weighted covariance matrix using

$$\widehat{\boldsymbol{\Sigma}}_{Y\mathbf{x}\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}}.$$

2. Perform an eigenvalue decomposition on $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1}\widehat{\boldsymbol{\Sigma}}_{Y\mathbf{x}\mathbf{x}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1}$.

3. The eigenvectors corresponding to the largest eigenvalues are the most important directions.

Alternatively, we can regress Y on $\mathbf{x}$, and use the least squares estimator to get the estimated residuals $\hat{r}$. Then we can proceed as above to calculate $\widehat{\boldsymbol{\Sigma}}_{r\mathbf{x}\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\hat{r}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}}$, and finally find the eigenvalue decomposition of $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1}\widehat{\boldsymbol{\Sigma}}_{r\mathbf{x}\mathbf{x}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1}$.

1.4 SDR for Multivariate Responses

Traditional sufficient dimension reduction methods, such as SIR and SAVE, focus on reducing the dimensionality of a p-predictor vector $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$ with a univariate response. Consider a multivariate response $\mathbf{y} = (Y_1, \ldots, Y_q)^{\mathrm{T}} \in \mathbb{R}^q$. Then the Central space is defined similar to the univariate case, that is, the central space denoted $\mathcal{S}_{\mathbf{y}|\mathbf{x}}$ is the intersection of all $\mathbf{B}$'s that satisfy

$$\mathbf{y} \perp\!\!\!\perp \mathbf{x} | \mathbf{B}^{\mathrm{T}}\mathbf{x}.$$

The central mean space $\mathcal{S}_{E(\mathbf{y}|\mathbf{x})}$ for the multivariate response is defined in parallel to $\mathcal{S}_{\mathbf{y}|\mathbf{x}}$.

SIR can be used directly for multivariate responses by generalizing the slicing technique, see, for example Aragon (1997), and Hsing (1999). Such generalization is done by slicing $\mathbf{y}$ into hypercubes. However, as the dimension of $\mathbf{y}$ increases, the number of slices increases exponentially, and consequently leading to inaccurate estimates within each slice. In practice, this generalization is only feasible for $\mathbf{y}$ with small dimension. Other methods for estimating $\mathcal{S}_{\mathbf{y}|\mathbf{x}}$ have been suggested to avoid large number of slices, such as, K-means inverse regression (KIR) (Setodji and Cook, 2004). In KIR, $\mathbf{y}$ is sliced using k-means clustering, which means that the number of slices can be controlled reasonably. However, selecting the number of clusters to use is subjective.

Instead of estimating $\mathcal{S}_{\mathbf{y}|\mathbf{x}}$ jointly, one can resort to marginally estimating each central space $\mathcal{S}_{Y_i|\mathbf{x}}$ and then combine all marginal $\mathcal{S}_{Y_i|\mathbf{x}}$'s to estimate the joint $\mathcal{S}_{\mathbf{y}|\mathbf{x}}$, see Cook and Setodji (2003), Saracco (2005), and Yin and Bura (2006)). Such methods are not guaranteed to recover the joint $\mathcal{S}_{\mathbf{y}|\mathbf{x}}$. Another option to estimate $\mathcal{S}_{\mathbf{y}|\mathbf{x}}$ is use CCA to estimate a univariate predictor $\mathbf{a}^{\mathrm{T}}\mathbf{x}$ and a univariate response $\mathbf{b}^{\mathrm{T}}\mathbf{y}$.

Iaci et al. (2016) introduced the concept of dual central subspaces (DCS). They provide a method for estimating DCS for simultaneous sufficient dimension reduction of two multivariate random vectors.

1.5 Popular Variable Selection Methods

Variable selection methods are concerned with recovering a subset of relevant predictors to the response for model building. The theme of variable selection, or sometimes called feature selection, is based on the idea that predictors can be classified into three categories: significant, redundant, and irrelevant. In a variable selection procedure, we usually want to get rid of the redundant and irrelavent variables. Depending on the interpretation, redundant variables could be considered significant since they are correlated with the significant variables, but the information they contain about the response are usually contained within other predictors. Suppose that $\mathbf{x}$ is a p dimensional predictor vector and $Y \in \mathbb{R}$ is a univariate response variable. In statistical literature, most methods are aimed at recovering the subset of variables that are most significant for the regression mean function $E(Y|\mathbf{x})$. Denote $\mathcal{S} = \{1, 2, \ldots, p\}$ as the full predictor index set. Then we can define the active variables set $\mathcal{A}$ as

$$\mathcal{A} = \{k \in \mathcal{S} : E(Y|\mathbf{x}) \text{ functionally depends on } X_k\}.$$

In the next sections we will go over some of the most commonly used variable selection methods.

1.5.1 All Possible Subsets

The all possible subsets (APS) method is very simple and straightforward, but very expensive in terms of computation. APS generates all possible models and uses a criterion, or

combination of more than one, to decide which model is the best one. For example, the linear regression R^2 could be used as a decision rule to select the best model. Other possible measures include adjusted R^2, Mallows' C_p, Akaike's Information Criterion (AIC), Schwarz's Bayesian Information Criterion BIC, or PRESS.

The problem with APS is that it becomes computationally intractable even for a modest number of covariates, that is for a p dimensional predictor, the total number of possible models is given by

$$\sum_{k=1}^{p} \frac{p!}{k!(p-k)!} = 2^p - 1.$$

Consequently, APS becomes impossible to implement for large datasets.

1.5.2 Sequential Search Algorithms

Forward Selection (FS) and Backward Elimination (BE) belong to class of algorithms called greedy. Both FS and BE reach their final solution by taking the optimal decision at each stage of the algorithm. The FS procedure starts with an empty set of variables. Then each predictor is tested whether it is significant or not

$$H_0 : \beta_i = 0 \text{ vs } \; H_a : \beta_i \neq 0, \text{ for } \; i = 1, \ldots, p.$$

From those predictors that turn out to be significant, we choose the one that improves the model the most using, for example, adjusted R^2. This process is repeated until we are left with no significant predictors. On the other hand, BE starts by including all candidate predictors and removing non-significant variables one at a time. At each step, if more than one predictor is not significant, we remove the one with the least contribution to our prespecified criterion, for example adjusted R^2.

Stepwise Regression (SR) is another greedy procedure that combines the ideas of both FS and BE. SR starts by making two forward addition steps just like in the FS algorithm. We then take a backward step removing the least significant predictor (if any). We keep alternating between adding and removing predictors until two consecutive steps yield neither addition nor removal of any predictors.

1.5.3 Variable Selection Using Sufficient Dimension Reduction

All of the afore mentioned variable selection procedures assume linearity between the response and the predictor. Using sufficient dimension reduction, we can build model-free variable selection frameworks. There are many advantages to working in a model-free framework for variable selection. One benefit is for large datasets we can perform the model-free variable selection as an initial step, separate from the model building step. This allows to toss out as much of the irrelevant predictors to make model building in the next stage easier. In the literature, there is a number of methods that deal with model-free variable selection using sufficient dimension reduction concepts. Regularized Sliced Inverse Regression (Zhong et al., 2005) and Sparse Sliced Inverse Regression (Li, 2007; Bondell and Li, 2009) are a variable selection method that combine the concept of variable selection with SIR.

CHAPTER 2

DUAL MODEL-FREE VARIABLE SELECTION WITH TWO GROUPS OF VARIABLES

In this chapter, we consider dual model-free variable selection with two groups of variables $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{y} \in \mathbb{R}^q$. Most existing variable selection methods are model-based, and consider selecting important predictors under a given parametric or semi-parametric model. The motivation behind our method is that in CCA the role of the predictor and the response is not important, potentially allowing us to build dual marginal coordinate hypothesis tests. Unlike model-based procedures in the literature, our proposal is model-free and does not require assuming specific models between $\mathbf{x}$ and $\mathbf{y}$. Existing model-free variable selection methods all focus on selecting important variables among $\mathbf{x}$ for the prediction of $\mathbf{y}$. See, for example, Li et al. (2005), Li (2007), Jiang and Liu (2014), and Yu et al (2016). The aforementioned model-free variable selection methods are closely related to sufficient dimension reduction (Cook, 1998). To achieve dual model-free variable selection, we demonstrate that CCA can be viewed as a valid sufficient dimension reduction procedure under suitable conditions. There is an important difference between CCA and popular sufficient dimension reduction methods such as SIR: CCA maintains the symmetry between $\mathbf{x}$ and $\mathbf{y}$ while SIR does not. We follow Yu et al (2016) and develop CCA-based model-free variable selection procedures. Unlike the procedures proposed in Yu et al (2016) that select important variables among $\mathbf{x}$, the symmetry in CCA provides a unique opportunity to perform dual variable selection among both $\mathbf{x}$ and $\mathbf{y}$ simultaneously. A backward algorithm is provided for the dual variable selection to demonstrate the performance of our test.

2.1 Review of SIR-based Trace Test

Let $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ and $\mathbf{y} = (Y_1, \ldots, Y_q)^{\mathrm{T}}$. Without loss of generality, assume $\mathrm{E}(\mathbf{x}) = \mathbf{0}$ and $\mathrm{E}(\mathbf{y}) = \mathbf{0}$. Denote $\mathbf{x}_{-k} = (X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_p)^{\mathrm{T}}$ for $k = 1, \ldots, p$. To test the importance of the kth predictor X_k, we may consider the following hypotheses

$$H_0^{-k} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{-k} \text{ v.s. } H_a^{-k} : \mathbf{y} \not\perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{-k}, \tag{2.4}$$

where $\perp\!\!\!\perp$ means independence and $\not\perp\!\!\!\perp$ means no independence. $H_0^{-k} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{-k}$ above implies that X_k is not important for the prediction of $\mathbf{y}$ in the presence of all the other predictors. Hypotheses (2.4) are known as the marginal coordinate hypotheses (Cook, 2004). Once we have a valid test for (2.4), sequential procedures such as forward selection, backward selection and stepwise regression can be designed in parallel to the classical procedures in linear regression. For example, Li et al. (2005) considers backward selection through the marginal coordinate test proposed by Cook (2004), while forward selection and stepwise regression through the trace test are discussed in Yu et al. (2016).

Since Cook (2004), different tests for (2.4) have been proposed in the literature. Most tests have the same flavor as the original marginal coordinate test in Cook (2004), such as Shao et al. (2007), Yu et al. (2010), and Yu and Dong (2016). Yu et al. (2016) introduce a novel family of trace tests, which can be combined with various sufficient dimension reduction methods. In the following, we first review SIR as a sufficient dimension reduction method, and then we revisit the SIR-based trace test for (2.4).

Classical sufficient dimension reduction aims to find $\boldsymbol{\beta} \in \mathbb{R}^{p \times d}$ with the smallest possible column space such that $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$. The corresponding column space is known as the central space for the regression of $\mathbf{y}$ on $\mathbf{x}$, and is denoted by $\mathcal{S}_{\mathbf{y}|\mathbf{x}}$. Let $\boldsymbol{\Sigma}_{\mathbf{x}} = \mathrm{var}(\mathbf{x})$

and denote $\{J_1, \ldots, J_H\}$ as a measurable partition of $\boldsymbol{\theta}_{\mathbf{y}}$, the sample space of $\mathbf{y}$. Under the linearity condition that $\mathrm{E}(\mathbf{x}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ is linear in $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$, we know from Li (1991) that

$$\mathrm{E}(\mathbf{z}|\mathbf{y} \in J_h) \in \boldsymbol{\Sigma}_{\mathbf{x}}^{1/2}\mathcal{S}_{\mathbf{y}|\mathbf{x}}, \tag{2.5}$$

where $\mathbf{z} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}\mathbf{x}$ is the standardized version of $\mathbf{x}$. Define $\mathbf{z}$-scale SIR kernel matrix as $\mathbf{M}^{\mathrm{SIR}} = \sum_{h=1}^{H} \pi_h \mathrm{E}(\mathbf{z}|\mathbf{y} \in J_h)\mathrm{E}^{\mathrm{T}}(\mathbf{z}|\mathbf{y} \in J_h)$, where $\pi_h = \Pr(\mathbf{y} \in J_h)$. From (2.5), we know

$$\mathrm{Span}(\mathbf{M}^{\mathrm{SIR}}) \subseteq \boldsymbol{\Sigma}_{\mathbf{x}}^{1/2}\mathcal{S}_{\mathbf{y}|\mathbf{x}}, \tag{2.6}$$

where $\mathrm{Span}(\cdot)$ denotes the column space.

Let $\boldsymbol{\Sigma}_{\mathbf{x}_{-k}} = \mathrm{var}(\mathbf{x}_{-k})$ and $\mathbf{z}_{-k} = \boldsymbol{\Sigma}_{\mathbf{x}_{-k}}^{-1/2}\mathbf{x}_{-k}$. Similar to $\mathbf{M}^{\mathrm{SIR}}$, we define $\mathbf{M}_{-k}^{\mathrm{SIR}} = \sum_{h=1}^{H} \pi_h \mathrm{E}(\mathbf{z}_{-k}|\mathbf{y} \in J_h)\mathrm{E}^{\mathrm{T}}(\mathbf{z}_{-k}|\mathbf{y} \in J_h)$. Yu et al. (2016) considers

$$\delta_k^{\mathrm{SIR}} = \mathrm{tr}(\mathbf{M}^{\mathrm{SIR}}) - \mathrm{tr}(\mathbf{M}_{-k}^{\mathrm{SIR}}), \tag{2.7}$$

where $\mathrm{tr}(\cdot)$ denotes the trace. Yu et al. (2016) provides the asymptotic distribution of $\hat{\delta}_k^{\mathrm{SIR}}$ under H_0^{-k}, where $\hat{\delta}_k^{\mathrm{SIR}}$ is the sample version of δ_k^{SIR}. By noting that $\delta_k^{\mathrm{SIR}} = 0$ under H_0^{-k} in (2.4), H_0^{-k} is rejected if $\hat{\delta}_k^{\mathrm{SIR}}$ is larger than a threshold determined by its asymptotic distribution under null.

2.2 The Principle of Dual Model-Free Variable Selection

Denote $\mathcal{I}_{\mathbf{x}} = \{1, 2, \ldots, p\}$ as the full index set for $\mathbf{x}$. Define the active set $\mathcal{A}$ for the regression of $\mathbf{y}$ on $\mathbf{x}$ as

$$\mathcal{A} = \{k \in \mathcal{I}_{\mathbf{x}} : \mathbf{y} \text{ depends on } \mathbf{x} \text{ through } X_k\}. \tag{2.8}$$

Similarly, let $\mathcal{I}_{\mathbf{y}} = \{1, 2, \ldots, q\}$ denote the full index set for $\mathbf{y}$, and the active set $\mathcal{B}$ for the regression of $\mathbf{x}$ on $\mathbf{y}$ is defined as

$$\mathcal{B} = \{j \in \mathcal{I}_{\mathbf{y}} : \mathbf{x} \text{ depends on } \mathbf{y} \text{ through } Y_j\}. \tag{2.9}$$

Let $\mathbf{x}_{\mathcal{A}} = \{X_k : k \in \mathcal{A}\}$ and $\mathbf{y}_{\mathcal{B}} = \{Y_j : j \in \mathcal{B}\}$. We have the following result.

Proposition 2.2.1 *The following three conditions are equivalent, and all are implied from the definitions of $\mathcal{A}$ in (2.8) and $\mathcal{B}$ in(2.9).*
(i) $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{A}}$ *and* $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{B}}$*;*
(ii) $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{A}}$ *and* $\mathbf{y} \perp\!\!\!\perp \mathbf{x}_{\mathcal{A}} \mid \mathbf{y}_{\mathcal{B}}$*;*
(iii) $\mathbf{y}_{\mathcal{B}} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{A}}$ *and* $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{B}}$.

Denote $\varnothing$ as the empty set. It follows from Proposition 2.2.1 that $\mathcal{A} = \varnothing$ if and only if $\mathcal{B} = \varnothing$. We remark that Proposition 2.2.1 is parallel to Proposition 1 in Iaci et al. (2016), where they study dual central spaces for sufficient dimension reduction.

The goal of dual model-free variable selection is to identify $\mathcal{A}$ and $\mathcal{B}$ without assuming specific models between $\mathbf{x}$ and $\mathbf{y}$. Let $\mathbf{x}_{\mathcal{F}} = \{X_k : k \in \mathcal{F}\}$ and $\mathbf{y}_{\mathcal{G}} = \{Y_j : j \in \mathcal{G}\}$, where $\mathcal{F} \subseteq \mathcal{I}_{\mathbf{x}}$ is the working active set for $\mathbf{x}$ and $\mathcal{G} \subseteq \mathcal{I}_{\mathbf{y}}$ is the working active set for $\mathbf{y}$. Motivated from part (i) in Proposition 2.2.1 and the marginal coordinate hypotheses (2.4) in Section 2, we consider the following dual marginal coordinate hypotheses

$$\begin{aligned} H_0^{\mathcal{F},[\mathcal{G}]} &: \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}} \text{ and } \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}} \\ \text{v.s. } H_a^{\mathcal{F},[\mathcal{G}]} &: \mathbf{y} \not\perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}} \text{ or } \mathbf{y} \not\perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}. \end{aligned} \tag{2.10}$$

If $H_0^{\mathcal{F},[\mathcal{G}]}$ in (2.10) is true, then obviously we have $\mathcal{A} \subseteq \mathcal{F}$ and $\mathcal{B} \subseteq \mathcal{G}$. We can then recover $\mathcal{A}$ and $\mathcal{B}$ by looking for the combination of the smallest possible $\mathcal{F}$ and the smallest possible $\mathcal{G}$ such that $H_0^{\mathcal{F},[\mathcal{G}]}$ is not rejected.

2.3 CCA-based Trace Tests and Dual Model-Free Variable Selection

We have reviewed in Section 2 that SIR-based trace test can be used to test the marginal coordinate hypotheses (2.4). To test the dual marginal coordinate hypotheses (2.10), where the roles of $\mathbf{x}$ and $\mathbf{y}$ are symmetric, we need a dimension reduction method that maintains the symmetry between $\mathbf{x}$ and $\mathbf{y}$. In Section 4.1, we introduce CCA as a dual sufficient dimension reduction method. In Section 4.2, we study CCA-based trace tests for selecting variables among either $\mathbf{x}$ or $\mathbf{y}$. In Section 4.3, CCA-based test for the dual marginal coordinate hypotheses (2.10) is developed. In Section 4.4, we propose a sample level algorithm for dual model-free variable selection.

2.3.1 CCA for Dual Sufficient Dimension Reduction

Recall that $\mathbf{z} = \mathbf{\Sigma}_{\mathbf{x}}^{-1/2}\mathbf{x}$ is the standardized version of $\mathbf{x}$. Let $\mathbf{w} = \mathbf{\Sigma}_{\mathbf{y}}^{-1/2}\mathbf{y}$ be the standardized version of $\mathbf{y}$, where $\mathbf{\Sigma}_{\mathbf{y}} = \text{var}(\mathbf{y})$. Define kernel matrices

$$\mathbf{M} = \text{E}(\mathbf{z}\mathbf{w}^{\text{T}})\text{E}(\mathbf{w}\mathbf{z}^{\text{T}}) \text{ and } \widetilde{\mathbf{M}} = \text{E}(\mathbf{w}\mathbf{z}^{\text{T}})\text{E}(\mathbf{z}\mathbf{w}^{\text{T}}). \tag{2.11}$$

Given $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{y} \in \mathbb{R}^q$, the ℓth pair of canonical covariates (u_ℓ, v_ℓ) is defined as $u_\ell = \mathbf{a}_\ell^{\text{T}}\mathbf{x}$ and $v_\ell = \mathbf{b}_\ell^{\text{T}}\mathbf{y}$, such that $\text{var}(u_\ell) = \text{var}(v_\ell) = 1$ and $\text{cov}(u_\ell, v_\ell)$ is maximized. For $\ell > 1$, u_ℓ and v_ℓ satisfy the additional constraints that $\text{cov}(u_\ell, u_k) = 0$ and $\text{cov}(v_\ell, v_k) = 0$ for all $k < \ell$. It is well-known that the $\mathbf{a}_\ell = \mathbf{\Sigma}_{\mathbf{x}}^{-1/2}\mathbf{c}_\ell$, where $\mathbf{c}_\ell$ is the eigenvector corresponding to the ℓth largest eigenvalue of $\mathbf{M}$. Similarly, $\mathbf{b}_\ell = \mathbf{\Sigma}_{\mathbf{y}}^{-1/2}\mathbf{d}_\ell$, where $\mathbf{d}_\ell$ is the ℓth eigenvector of $\widetilde{\mathbf{M}}$. Denote $\mathcal{S}_{\mathbf{y}|\mathbf{x}}$ and $\mathcal{S}_{\mathbf{x}|\mathbf{y}}$ as the dual central spaces for the regression of $\mathbf{y}$ on $\mathbf{x}$ and the regression of $\mathbf{x}$ on $\mathbf{y}$, respectively. The next result states that matrices $\mathbf{M}$ and $\widetilde{\mathbf{M}}$ are closely related to sufficient dimension reduction.

Proposition 2.3.1 *Suppose* $\mathrm{E}(\mathbf{x}) = \mathbf{0}$ *and* $\mathrm{E}(\mathbf{y}) = \mathbf{0}$. *Assume* $\boldsymbol{\beta}$ *is the basis for* $\mathcal{S}_{\mathbf{y}|\mathbf{x}}$ *and* $\boldsymbol{\eta}$ *is the basis for* $\mathcal{S}_{\mathbf{x}|\mathbf{y}}$.

(i) If $\mathrm{E}(\mathbf{x}|\boldsymbol{\beta}^T\mathbf{x})$ *is linear in* $\boldsymbol{\beta}^T\mathbf{x}$, *then* $\mathrm{Span}(\mathbf{M}) \subseteq \boldsymbol{\Sigma}_{\mathbf{x}}^{1/2}\mathcal{S}_{\mathbf{y}|\mathbf{x}}$;

(ii) If $\mathrm{E}(\mathbf{y}|\boldsymbol{\eta}^T\mathbf{y})$ *is linear in* $\boldsymbol{\eta}^T\mathbf{y}$, *then* $\mathrm{Span}(\widetilde{\mathbf{M}}) \subseteq \boldsymbol{\Sigma}_{\mathbf{y}}^{1/2}\mathcal{S}_{\mathbf{x}|\mathbf{y}}$.

The assumptions made in this proposition are common in the sufficient dimension reduction literature. Proposition 2.3.1 implies that the column space of $\boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}\mathbf{M}$ can recover the central space for the regression of $\mathbf{y}$ on $\mathbf{x}$, while the column space of $\boldsymbol{\Sigma}_{\mathbf{y}}^{-1/2}\widetilde{\mathbf{M}}$ can recover the central space for the regression of $\mathbf{x}$ on $\mathbf{y}$. It follows that $\mathbf{a}_\ell \in \mathcal{S}_{\mathbf{y}|\mathbf{x}}$ and $\mathbf{b}_\ell \in \mathcal{S}_{\mathbf{x}|\mathbf{y}}$. We remark that the conclusions in Proposition 2.3.1 bare close resemblance to (2.6) about the SIR-based kernel matrix $\mathbf{M}^{\mathrm{SIR}}$.

2.3.2 CCA-based Trace Tests for Marginal Coordinate Hypotheses

We consider two sets of marginal coordinate hypotheses in this section, both of which are related to (2.10). The first set is

$$H_0^{\mathcal{F}} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}} \text{ v.s. } H_a^{\mathcal{F}} : \mathbf{y} \not\perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}. \tag{2.12}$$

Hypotheses (2.12) include (2.4) as a special case, as $\mathbf{x}_{\mathcal{F}}$ becomes $\mathbf{x}_{-k}$ when we take $\mathcal{F} = \{1, \ldots, k-1, k+1, \ldots, p\}$. The second set is

$$H_0^{[\mathcal{G}]} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}} \text{ v.s. } H_a^{[\mathcal{G}]} : \mathbf{y} \not\perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}. \tag{2.13}$$

While hypotheses (2.12) can be used for selecting important variables among $\mathbf{x}$, hypotheses (2.13) are useful for selecting important variables among $\mathbf{y}$.

First we focus on the CCA-based trace test for (2.12). Let $\mathbf{\Sigma}_{\mathbf{x}_\mathcal{F}} = \text{var}(\mathbf{x}_\mathcal{F})$ and $\mathbf{z}_\mathcal{F} = \mathbf{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1/2}\mathbf{x}_\mathcal{F}$. Motivated by the SIR-based trace test, we consider

$$\delta_{-\mathcal{F}} = \text{tr}(\mathbf{M}) - \text{tr}(\mathbf{M}_\mathcal{F}), \tag{2.14}$$

where $\mathbf{M}_\mathcal{F} = \text{E}(\mathbf{z}_\mathcal{F}\mathbf{w}^\text{T})\text{E}(\mathbf{w}\mathbf{z}_\mathcal{F}^\text{T})$. We remark that $\delta_{-\mathcal{F}}$ is constructed as the trace difference of two $\mathbf{z}$-scale CCA kernel matrices, which has the same flavor as δ_k^{SIR} in (2.7).

Let $\mathcal{F}^c$ be the complement of $\mathcal{F}$ in $\mathcal{I}_\mathbf{x}$ and denote $\mathbf{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}} = \text{var}(\mathbf{x}_{\mathcal{F}^c})$. Define $\mathbf{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_\mathcal{F}} = \mathbf{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}} - \text{E}(\mathbf{x}_{\mathcal{F}^c}\mathbf{x}_\mathcal{F}^\text{T})\mathbf{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\text{E}(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^\text{T})$ and $\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_\mathcal{F}} = \mathbf{x}_{\mathcal{F}^c} - \text{E}(\mathbf{x}_{\mathcal{F}^c}\mathbf{x}_\mathcal{F}^\text{T})\mathbf{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\mathbf{x}_\mathcal{F}$. Then we have

Proposition 2.3.2 *Suppose* $\text{E}(\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_\mathcal{F})$ *is a linear function of* $\mathbf{x}_\mathcal{F}$. *Then*
(i) $\delta_{-\mathcal{F}} = \text{tr}\{\mathbf{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_\mathcal{F}}^{-1}\text{E}(\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_\mathcal{F}}\mathbf{y}^T)\mathbf{\Sigma}_\mathbf{y}^{-1}\text{E}(\mathbf{y}\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_\mathcal{F}}^T)\}$.
(ii) $\delta_{-\mathcal{F}} = 0$ *under* $H_0^\mathcal{F} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_\mathcal{F}$.

The assumption made in this proposition is common in the model-free variable selection literature, and it is satisfied if $\mathbf{x}$ is normal. The first part of Proposition 2.3.2 provides the explicit formula to calculate $\delta_{-\mathcal{F}}$. The second part states that if $\mathbf{x}_{\mathcal{F}^c}$ is unimportant for the prediction of $\mathbf{y}$ given $\mathbf{x}_\mathcal{F}$, then $\delta_{-\mathcal{F}}$ becomes zero. Yu et al. (2006) have shown that $\delta_k^{\text{SIR}} = 0$ if X_k is unimportant for the prediction of $\mathbf{y}$ given $\mathbf{x}_{-k}$. Our result here is more general as $\mathcal{F}^c$ can contain more than one variable. Denote $\hat{\delta}_{-\mathcal{F}}$ as the sample version of $\delta_{-\mathcal{F}}$. We reject $H_0^\mathcal{F}$ if $\hat{\delta}_{-\mathcal{F}}$ is too large. The asymptotic distribution of $\hat{\delta}_{-\mathcal{F}}$ under $H_0^\mathcal{F}$ is provided in Corollary B.1 in the Appendix.

Next we introduce the CCA-based trace test for (2.13). Let $\mathbf{\Sigma}_{\mathbf{y}_\mathcal{G}} = \text{var}(\mathbf{y}_\mathcal{G})$ and $\mathbf{w}_\mathcal{G} = \mathbf{\Sigma}_{\mathbf{y}_\mathcal{G}}^{-1/2}\mathbf{y}_\mathcal{G}$. Denote $\widetilde{\mathbf{M}}_\mathcal{G} = \text{E}(\mathbf{w}_\mathcal{G}\mathbf{z}^\text{T})\text{E}(\mathbf{z}\mathbf{w}_\mathcal{G}^\text{T})$ and consider

$$\delta^{-\mathcal{G}} = \text{tr}(\widetilde{\mathbf{M}}) - \text{tr}(\widetilde{\mathbf{M}}_\mathcal{G}). \tag{2.15}$$

Let $\mathcal{G}^c$ be the complement of $\mathcal{G}$ in $\mathcal{I}_{\mathbf{y}}$ and denote $\mathbf{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}} = \text{var}(\mathbf{y}_{\mathcal{G}^c})$. Define $\mathbf{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}} = \mathbf{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}} - \text{E}(\mathbf{y}_{\mathcal{G}^c}\mathbf{y}_{\mathcal{G}}^{\text{T}})\mathbf{\Sigma}_{\mathbf{y}_{\mathcal{G}}}^{-1}\text{E}(\mathbf{y}_{\mathcal{G}}\mathbf{y}_{\mathcal{G}^c}^{\text{T}})$. Let $\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}} = \mathbf{y}_{\mathcal{G}^c} - \text{E}(\mathbf{y}_{\mathcal{G}^c}\mathbf{y}_{\mathcal{G}}^{\text{T}})\mathbf{\Sigma}_{\mathbf{y}_{\mathcal{G}}}^{-1}\mathbf{y}_{\mathcal{G}}$. Parallel to Proposition 2.3.2, we have

Proposition 2.3.3 *Suppose* $\text{E}(\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}})$ *is a linear function of* $\mathbf{y}_{\mathcal{G}}$. *Then*
(i) $\delta^{-\mathcal{G}} = \text{tr}\{\mathbf{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{-1}\text{E}(\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}\mathbf{x}^{T})\mathbf{\Sigma}_{\mathbf{x}}^{-1}\text{E}(\mathbf{x}\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{T})\}$.
(ii) $\delta^{-\mathcal{G}} = 0$ *under* $H_0^{[\mathcal{G}]} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}$.

Let $\hat{\delta}^{-\mathcal{G}}$ be the sample version of $\delta^{-\mathcal{G}}$. We reject $H_0^{[\mathcal{G}]}$ if $\hat{\delta}^{-\mathcal{G}}$ is too large. The asymptotic distribution of $\hat{\delta}^{-\mathcal{G}}$ under $H_0^{[\mathcal{G}]}$ is provided in Corollary B.2 in the Appendix.

2.3.3 CCA-based Trace Test for Dual Marginal Coordinate Hypotheses

In this section, we develop a test for $H_0^{\mathcal{F},[\mathcal{G}]} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$ and $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}$ versus the alternative $H_a^{\mathcal{F},[\mathcal{G}]} : \mathbf{y} \not\perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$ or $\mathbf{y} \not\perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}$. From the definition of $\mathbf{M} = \text{E}(\mathbf{z}\mathbf{w}^{\text{T}})\text{E}(\mathbf{w}\mathbf{z}^{\text{T}})$ and $\widetilde{\mathbf{M}} = \text{E}(\mathbf{w}\mathbf{z}^{\text{T}})\text{E}(\mathbf{z}\mathbf{w}^{\text{T}})$ in (2.11), we have $\text{tr}(\mathbf{M}) = \text{tr}(\widetilde{\mathbf{M}})$. Recall $\mathbf{M}_{\mathcal{F}} = \text{E}(\mathbf{z}_{\mathcal{F}}\mathbf{w}^{\text{T}})\text{E}(\mathbf{w}\mathbf{z}_{\mathcal{F}}^{\text{T}})$ and define $\mathbf{M}^{\mathcal{G}} = \text{E}(\mathbf{z}\mathbf{w}_{\mathcal{G}}^{\text{T}})\text{E}(\mathbf{w}_{\mathcal{G}}\mathbf{z}^{\text{T}})$. It is easy to see that $\text{tr}(\mathbf{M}^{\mathcal{G}}) = \text{tr}(\widetilde{\mathbf{M}}_{\mathcal{G}})$. Hence $\delta^{-\mathcal{G}}$ in (2.15) becomes

$$\delta^{-\mathcal{G}} = \text{tr}(\mathbf{M}) - \text{tr}(\mathbf{M}^{\mathcal{G}}). \tag{2.16}$$

We have seen that $\delta_{-\mathcal{F}} = \text{tr}(\mathbf{M}) - \text{tr}(\mathbf{M}_{\mathcal{F}})$ in (2.14) can be used to test $H_0^{\mathcal{F}} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$, and $\delta^{-\mathcal{G}}$ in (2.16) can be used to test $H_0^{[\mathcal{G}]} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}$. This motivates us to consider

$$\delta_{-\mathcal{F}}^{-\mathcal{G}} = \text{tr}(\mathbf{M}) - \text{tr}(\mathbf{M}_{\mathcal{F}}^{\mathcal{G}}), \tag{2.17}$$

where $\mathbf{M}_{\mathcal{F}}^{\mathcal{G}} = \text{E}(\mathbf{z}_{\mathcal{F}}\mathbf{w}_{\mathcal{G}}^{\text{T}})\text{E}(\mathbf{w}_{\mathcal{G}}\mathbf{z}_{\mathcal{F}}^{\text{T}})$. Note that $\delta_{-\mathcal{F}}^{-\mathcal{G}}$ in (2.17) include $\delta_{-\mathcal{F}}$ and $\delta^{-\mathcal{G}}$ as special cases. If we take $\mathcal{F} = \mathcal{I}_{\mathbf{x}}$, then $\delta_{-\mathcal{F}}^{-\mathcal{G}}$ becomes $\delta^{-\mathcal{G}}$. On the other hand, $\delta_{-\mathcal{F}}^{-\mathcal{G}}$ reduces to $\delta_{-\mathcal{F}}$ when we set $\mathcal{G} = \mathcal{I}_{\mathbf{y}}$.

The symmetry between $\mathbf{z}$ and $\mathbf{w}$ in the definition of $\mathbf{M}$ and $\widetilde{\mathbf{M}}$ allows $\mathrm{tr}(\mathbf{M})$ to simultaneously capture the regression information between $\mathbf{y}$ and $\mathbf{x}$ as well as the regression information between $\mathbf{x}$ and $\mathbf{y}$. This is a unique feature of the CCA-based trace test, as $\mathrm{tr}(\mathbf{M}^{\mathrm{SIR}})$ in (2.7) only captures the regression information between $\mathbf{y}$ and $\mathbf{x}$. Parallel to Proposition 2.3.2 and Proposition 2.3.3, we have

Proposition 2.3.4 *Suppose* $\mathrm{E}(\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}})$ *is a linear function of* $\mathbf{x}_{\mathcal{F}}$ *and* $\mathrm{E}(\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}})$ *is a linear function of* $\mathbf{y}_{\mathcal{G}}$. *Then*

(i) $\delta_{-\mathcal{F}}^{-\mathcal{G}} = \mathrm{tr}\{\boldsymbol{\Sigma}^{-1}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}\mathbf{y}^{T})\boldsymbol{\Sigma}^{-1}_{\mathbf{y}}\mathrm{E}(\mathbf{y}\boldsymbol{\gamma}^{T}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}})\} + \mathrm{tr}\{\boldsymbol{\Sigma}^{-1}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}\mathbf{x}^{T}_{\mathcal{F}})$
$\boldsymbol{\Sigma}^{-1}_{\mathbf{x}_{\mathcal{F}}}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\boldsymbol{\gamma}^{T}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}})\}$.

(ii) $\delta_{-\mathcal{F}}^{-\mathcal{G}} = 0$ *under* $H_0^{\mathcal{F},[\mathcal{G}]} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$ *and* $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}$.

Let $\{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \cdots, (\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}$ be an i.i.d. sample. Let $\bar{\mathbf{x}} = n^{-1}\sum_{i=1}^{n}\mathbf{x}^{(i)}$, $\tilde{\mathbf{x}}^{(i)} = \mathbf{x}^{(i)} - \bar{\mathbf{x}}$, and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}} = n^{-1}\sum_{i=1}^{n}\tilde{\mathbf{x}}^{(i)}(\tilde{\mathbf{x}}^{(i)})^{\mathrm{T}}$. Similarly let $\bar{\mathbf{y}} = n^{-1}\sum_{i=1}^{n}\mathbf{y}^{(i)}$, $\tilde{\mathbf{y}}^{(i)} = \mathbf{y}^{(i)} - \bar{\mathbf{y}}$, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{y}} = n^{-1}\sum_{i=1}^{n}\tilde{\mathbf{y}}^{(i)}(\tilde{\mathbf{y}}^{(i)})^{\mathrm{T}}$, $\mathrm{E}_n(\mathbf{x}\mathbf{y}^{\mathrm{T}}) = n^{-1}\sum_{i=1}^{n}\tilde{\mathbf{x}}^{(i)}(\tilde{\mathbf{y}}^{(i)})^{\mathrm{T}}$, and $\mathrm{E}_n(\mathbf{y}\mathbf{x}^{\mathrm{T}}) = n^{-1}\sum_{i=1}^{n}\tilde{\mathbf{y}}^{(i)}(\tilde{\mathbf{x}}^{(i)})^{\mathrm{T}}$. Then $\hat{\mathbf{M}} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1/2}\mathrm{E}_n(\mathbf{x}\mathbf{y}^{\mathrm{T}})\widehat{\boldsymbol{\Sigma}}_{\mathbf{y}}^{-1}\mathrm{E}_n(\mathbf{y}\mathbf{x}^{\mathrm{T}})\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1/2}$. Similarly one can calculate

$$\hat{\mathbf{M}}_{\mathcal{F}}^{\mathcal{G}} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{x}_{\mathcal{F}}}^{-1/2}\mathrm{E}_n(\mathbf{x}_{\mathcal{F}}\mathbf{y}_{\mathcal{G}}^{\mathrm{T}})\widehat{\boldsymbol{\Sigma}}_{\mathbf{y}_{\mathcal{G}}}^{-1}\mathrm{E}_n(\mathbf{y}_{\mathcal{G}}\mathbf{x}_{\mathcal{F}}^{\mathrm{T}})\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}_{\mathcal{F}}}^{-1/2}.$$

Then the sample version of $\delta_{-\mathcal{F}}^{-\mathcal{G}}$ in (2.17) becomes

$$\hat{\delta}_{-\mathcal{F}}^{-\mathcal{G}} = \mathrm{tr}(\hat{\mathbf{M}}) - \mathrm{tr}(\hat{\mathbf{M}}_{\mathcal{F}}^{\mathcal{G}}). \tag{2.18}$$

We conclude this section with the asymptotic distribution of $\hat{\delta}_{-\mathcal{F}}^{-\mathcal{G}}$ under $H_0^{\mathcal{F},[\mathcal{G}]}$. Assume $|\mathcal{F}| = p_1$ and $|\mathcal{G}| = q_1$, where $|\cdot|$ denotes cardinality.

Theorem 2.3.1 *Suppose* $\mathrm{E}(\mathbf{x}) = \mathbf{0}$, $\mathrm{E}(\mathbf{y}) = \mathbf{0}$, $\mathrm{E}(\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}})$ *is a linear function of* $\mathbf{x}_{\mathcal{F}}$ *and* $\mathrm{E}(\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}})$ *is a linear function of* $\mathbf{y}_{\mathcal{G}}$*. Then under* $H_0^{\mathcal{F},[\mathcal{G}]}$*,*

$$n\hat{\delta}_{-\mathcal{F}}^{-\mathcal{G}} \xrightarrow{D} \sum_{\ell=1}^{L} \tau_\ell \chi_\ell^2(1)$$

as $n \to \infty$*, where "*$\xrightarrow{D}$*" means convergence in distribution,* $L = pq - p_1q_1$*,* $\chi_\ell^2(1)$ *are independent chi-square with one degree of freedom for* $\ell = 1, \ldots, L$*,* $\tau_1 \geq \ldots \geq \tau_L$ *are the eigenvalues of* $\boldsymbol{\Omega}$*, and the exact form of* $\boldsymbol{\Omega}$ *is provided in the Appendix.*

The asymptotic distribution in Theorem 2.3.1 needs to be estimated in practice. Specifically, let $\hat{\boldsymbol{\Omega}}$ be the sample estimators of $\boldsymbol{\Omega}$, and let $\hat{\tau}_1 \geq \ldots \geq \hat{\tau}_L$ be the eigenvalues of $\hat{\boldsymbol{\Omega}}$. Denote $\boldsymbol{\zeta} = (\hat{\tau}_1, \ldots, \hat{\tau}_L)^{\mathrm{T}} \in \mathbb{R}^L$ and let $\boldsymbol{\Xi} \in \mathbb{R}^{N \times L}$ consist of i.i.d. $\chi^2(1)$ realizations. Then the N elements of $\boldsymbol{\Xi}\boldsymbol{\zeta}$ become realizations of the approximate asymptotic distribution of $n\hat{\delta}_{-\mathcal{F}}^{-\mathcal{G}}$ under null. The proportion of these N elements greater than $n\hat{\delta}_{-\mathcal{F}}^{-\mathcal{G}}$ become the approximate p-value. For a given significance level α, we reject $H_0^{\mathcal{F},[\mathcal{G}]}$ if this p-value is smaller than α. We use $N = 500$ in our numerical studies.

2.3.4 Algorithm for Dual Model-Free Variable Selection

Let $\{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \cdots, (\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}$ be an i.i.d sample of $\{\mathbf{x} \in \mathbb{R}^p, \mathbf{y} \in \mathbb{R}^q\}$. We devise a sample-level algorithm for dual model-free variable selection in this section. From the development in Section 3, we have seen that the active sets $\mathcal{A}$ and $\mathcal{B}$ can be recovered by the smallest possible $\mathcal{F}$ and the smallest possible $\mathcal{G}$ such that $H_0^{\mathcal{F},[\mathcal{G}]}$ is not rejected. This motivates us to consider the following joint backward selection procedure.

1. Initial step. Set $\mathcal{F}^{(0)} = \{1, 2, \ldots, p\}$ and $\mathcal{G}^{(0)} = \{1, 2, \ldots, q\}$. Let α be the prespecified significance level.

1.1 For $i = 1, \ldots, p$, denote $\mathcal{F}_{-i}^{(0)}$ as the index set where i is removed from $\mathcal{F}^{(0)}$, and let $\varrho_{i,(0)}$ be the approximate p-value from testing $H_0^{\mathcal{F}_{-i}^{(0)},[\mathcal{G}^{(0)}]}$ against its alternative.

1.2 For $j = 1, \ldots, q$, denote $\mathcal{G}_{-j}^{(0)}$ as the index set where j is removed from $\mathcal{G}^{(0)}$, and let $\varrho_{p+j,(0)}$ be the approximate p-value from testing $H_0^{\mathcal{F}^{(0)},[\mathcal{G}_{-j}^{(0)}]}$ against its alternative.

1.3 Let $k_{(0)} = \underset{\iota=1,\ldots,p+q}{\arg\max}\, \varrho_{\iota,(0)}$ and $\varrho_{(0)} = \underset{\iota=1,\ldots,p+q}{\max} \varrho_{\iota,(0)}$. If $\varrho_{(0)} \geq \alpha$ and $k_{(0)} \leq p$, then set $\mathcal{F}^{(1)} = \mathcal{F}_{-k_{(0)}}^{(0)}$, $\mathcal{G}^{(1)} = \mathcal{G}^{(0)}$, and go to Step 2. If $\varrho_{(0)} \geq \alpha$ and $k_{(0)} > p$, then set $\mathcal{F}^{(1)} = \mathcal{F}^{(0)}$, $\mathcal{G}^{(1)} = \mathcal{G}_{-\{k_{(0)}-p\}}^{(0)}$, and go to Step 2. If $\varrho_{(0)} < \alpha$, then stop the algorithm and return $\hat{\mathcal{A}} = \mathcal{F}^{(0)}$, $\hat{\mathcal{B}} = \mathcal{G}^{(0)}$.

2. Iteration step. At the beginning of the ℓth iteration, let $\mathcal{F}^{(\ell)}$ and $\mathcal{G}^{(\ell)}$ be the working index sets. Assume $|\mathcal{F}^{(\ell)}| = p_{(\ell)}$ and $|\mathcal{G}^{(\ell)}| = q_{(\ell)}$.

2.1 For $i = 1, \ldots, p_{(\ell)}$, denote $\mathcal{F}_{-i}^{(\ell)}$ as the index set where the ith element of $\mathcal{F}^{(\ell)}$ is removed from $\mathcal{F}^{(\ell)}$, and let $\varrho_{i,(\ell)}$ be the approximate p-value from testing $H_0^{\mathcal{F}_{-i}^{(\ell)},[\mathcal{G}^{(\ell)}]}$ against its alternative.

2.2 For $j = 1, \ldots, q_{(\ell)}$, denote $\mathcal{G}_{-j}^{(\ell)}$ as the index set where the jth element of $\mathcal{G}^{(\ell)}$ is removed from $\mathcal{G}^{(\ell)}$, and let $\varrho_{p_{(\ell)}+j,(\ell)}$ be the approximate p-value from testing $H_0^{\mathcal{F}^{(\ell)},[\mathcal{G}_{-j}^{(\ell)}]}$ against its alternative.

2.3 Let $k_{(\ell)} = \underset{\iota=1,\ldots,p_{(\ell)}+q_{(\ell)}}{\arg\max}\, \varrho_{\iota,(\ell)}$ and $\varrho_{(\ell)} = \underset{\iota=1,\ldots,p_{(\ell)}+q_{(\ell)}}{\max} \varrho_{\iota,(\ell)}$. If $\varrho_{(\ell)} \geq \alpha$ and $k_{(\ell)} \leq p_{(\ell)}$, then set $\mathcal{F}^{(\ell+1)} = \mathcal{F}_{-k_{(\ell)}}^{(\ell)}$, $\mathcal{G}^{(\ell+1)} = \mathcal{G}^{(\ell)}$, and repeat Step 2. If $\varrho_{(\ell)} \geq \alpha$ and $k_{(\ell)} > p_{(\ell)}$, then set $\mathcal{F}^{(\ell+1)} = \mathcal{F}^{(\ell)}$, $\mathcal{G}^{(\ell+1)} = \mathcal{G}_{-\{k_{(\ell)}-p_{(\ell)}\}}^{(\ell)}$, and repeat Step 2. If $\varrho_{(\ell)} < \alpha$, then stop the iteration and return $\hat{\mathcal{A}} = \mathcal{F}^{(\ell)}$, $\hat{\mathcal{B}} = \mathcal{G}^{(\ell)}$.

In the initial step, we first test $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{-i}$ against its alternative for $i = 1, \ldots, p$. Then we test $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{-j}$ for $j = 1, \ldots, q$. The corresponding p-values are denoted as $\varrho_{\iota,(0)}$ for

$\iota = 1, \ldots, p+q$. The maximum p-value $\varrho_{(0)}$ is then compared to α. If $\varrho_{(0)}$ is smaller than α, then $\varrho_{\iota,(0)} < \alpha$ for any $\iota = 1, \ldots, p+q$. Thus we reject $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{-i}$ for any $i = 1, \ldots, p$, and we reject $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{-j}$ for any $j = 1, \ldots, q$. Hence we estimate the active sets by $\mathcal{F}^{(0)}$ and $\mathcal{G}^{(0)}$. In the case with $\varrho_{(0)} \geq \alpha$, the least significant element, which is indexed by $k_{(0)}$, can be removed from the active sets. For $k_{(0)} \leq p$, we update $\mathcal{F}$ by removing the least significant element, which corresponds to an element in $\mathbf{x}$. For $k_{(0)} > p$, the least significant element corresponds to an element in $\mathbf{y}$ and we only update $\mathcal{G}$.

In the ℓth iteration, we start from working index sets $\mathcal{F}^{(\ell)}$ and $\mathcal{G}^{(\ell)}$. Note that $H_0^{\mathcal{F}^{(\ell)},[\mathcal{G}^{(\ell)}]}$ is not rejected from the last iteration, as we only go to the ℓth iteration if $\varrho_{(\ell-1)} \geq \alpha$. In another word, the ℓth iteration is needed only when $\mathcal{F}^{(\ell-1)}$ or $\mathcal{G}^{(\ell-1)}$ is updated. Parallel to the initial step, after removing one element at a time from either $\mathcal{F}^{(\ell)}$ or $\mathcal{G}^{(\ell)}$, we test the dual marginal coordinate hypotheses (2.10) and get p-value $\varrho_{\iota,(\ell)}$ for $\iota = 1, \ldots, p_{(\ell)} + q_{(\ell)}$. The maximum p-value $\varrho_{(\ell)}$ is then compared to α. If $\varrho_{(\ell)} < \alpha$, then $\mathcal{F}^{(\ell)}$ and $\mathcal{G}^{(\ell)}$ can not be further reduced. We stop the iteration and estimate the active sets by $\mathcal{F}^{(\ell)}$ and $\mathcal{G}^{(\ell)}$. Otherwise we go to the next iteration, where either $\mathcal{F}^{(\ell)}$ or $\mathcal{G}^{(\ell)}$ is updated by removing the least significant element. Note that in each iteration, we are testing the conditional independence between $\mathbf{x}$ and $\mathbf{y}$, and our procedure asymptotically controls the type-I error rate at the significance level α.

2.4 Numerical Results

We use synthetic data in Section 5.1, and a real data analysis is considered in Section 5.2.

2.4.1 Simulation Studies

Let $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ be multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}_{\mathbf{x}} = \{\sigma_{ij}\}$, where $\sigma_{ij} = \sigma^{|i-j|}$ for $1 \leq i, j \leq p$. Similarly, let $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_q)^{\mathrm{T}}$ be multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}_{\boldsymbol{\epsilon}} = \{\theta_{ij}\}$, where $\theta_{ij} = \theta^{|i-j|}$ for $1 \leq i, j \leq q$. The error $\boldsymbol{\epsilon}$ is independent of $\mathbf{x}$. Then we generate $\mathbf{y} = (Y_1, \ldots, Y_q)^{\mathrm{T}}$ from the following two models:

$$\text{Model 1: } Y_1 = \epsilon_1, Y_2 = \epsilon_2, Y_3 = \epsilon_3, Y_4 = X_1 + X_2 + \epsilon_4;$$

$$\text{Model 2: } Y_1 = e^{0.5X_3} + \sin(X_4) + \epsilon_1, Y_2 = X_3^3 + X_4 + \epsilon_2, Y_3 = \epsilon_3, Y_4 = \epsilon_4.$$

In both models, we set $p = 5$ and $q = 4$. In Model 1, we set $\sigma = 0$ and $\theta = 0.5$. The active set for the regression of $\mathbf{y}$ on $\mathbf{x}$ is $\mathcal{A} = \{1, 2\}$. Due to the nonzero correlation among the ϵ's, we cannot determine $\mathcal{B}$ by evaluating the forward regression between $\mathbf{y}$ and $\mathbf{x}$. Instead we calculate $\mathrm{E}(\mathbf{x} \mid \mathbf{y}) = (0, 0, 0, 4Y_4/11 - 2Y_3/11, 4Y_4/11 - 2Y_3/11)^{\mathrm{T}}$, and thus the active set for the regression of $\mathbf{x}$ on $\mathbf{y}$ is $\mathcal{B} = \{3, 4\}$. More details are provided in the Appendix. In Model 2, we set $\sigma = 0.5$ and $\theta = 0$. We have $\mathcal{A} = \{3, 4\}$ and $\mathcal{B} = \{1, 2\}$ as the result of $\theta = 0$.

First we evaluate the performance of the CCA-based trace test for the dual marginal coordinate hypotheses (2.10). For user-specified $\mathcal{F}$ and $\mathcal{G}$, we test $H_0^{\mathcal{F},[\mathcal{G}]} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$ and $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}$ versus the alternative $H_a^{\mathcal{F},[\mathcal{G}]} : \mathbf{y} \not\perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$ or $\mathbf{y} \not\perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}$. Based on 1000 repetitions, the frequencies of $H_0^{\mathcal{F},[\mathcal{G}]}$ being rejected are reported in Table 2.1. Fix sample size $n = 300$ and let $\alpha = 0.01, 0.05, 0.1$. We consider two combinations of $\mathcal{F}$ and $\mathcal{G}$. When $\mathcal{F} = \{3, 4\}$ and $\mathcal{G} = \{1, 2\}$, $H_0^{\mathcal{F},[\mathcal{G}]}$ does not hold for Model 1. We see from Table 2.1 that $H_0^{\mathcal{F},[\mathcal{G}]}$ is always rejected regardless of the α value. When $\mathcal{F} = \{1, 2\}$ and $\mathcal{G} = \{3, 4\}$, $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$ and $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}$ for Model 1. We see that the frequencies of rejecting $H_0^{\mathcal{F},[\mathcal{G}]}$ are close to the corresponding nominal level. The results for Model 2 are reversed. The first

	Model 1			Model 2		
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
$\mathcal{F} = \{3, 4\}, \mathcal{G} = \{1, 2\}$	1	1	1	0.007	0.033	0.085
$\mathcal{F} = \{1, 2\}, \mathcal{G} = \{3, 4\}$	0.007	0.044	0.093	1	1	1

Table 2.1. Frequencies of rejecting $H_0^{\mathcal{F},[\mathcal{G}]}$ based on 1000 repetitions.

combination of $\mathcal{F}$ and $\mathcal{G}$ now corresponds to the Type-I error, and the frequencies are close to the nominal level. The second combination of $\mathcal{F}$ and $\mathcal{G}$ corresponds to the estimated power, which is 1 for all α values.

Next we investigate the performance of the joint backward selection algorithm proposed in Section 2.3.4. For $i = 1, \ldots, 1000$, denote the estimated active sets of the ith repetition as $\hat{\mathcal{A}}_i$ and $\hat{\mathcal{B}}_i$. We define the over-fitted frequency (OF), the correctly-fitted frequency (CF), and the under-fitted frequency (UF) as

$$OF = 1000^{-1} \sum_{i=1}^{1000} \left\{ I(\mathcal{A} \subseteq \hat{\mathcal{A}}_i) I(\mathcal{B} \subseteq \hat{\mathcal{B}}_i) - I(\mathcal{A} = \hat{\mathcal{A}}_i) I(\mathcal{B} = \hat{\mathcal{B}}_i) \right\},$$

$$CF = 1000^{-1} \sum_{i=1}^{1000} I(\mathcal{A} = \hat{\mathcal{A}}_i) I(\mathcal{B} = \hat{\mathcal{B}}_i), \text{ and } UF = 1 - CF - OF,$$

where $I(\cdot)$ denotes the indicator function. The average model size is defined as $MS = 1000^{-1} \sum_{i=1}^{1000} \left(|\hat{\mathcal{A}}_i| + |\hat{\mathcal{B}}_i| \right)$. Based on 1000 repetitions, we report UF, CF, OF, MS together with the frequencies of each variable being selected.

The variable selection results for Model 1 is summarized in Table 2.2. We fix $\alpha = 0.05$ and let $n = 100, 300, 700$. We see that the variable selection performance improves as n increases. For $n = 300$ and $n = 700$, the unimportant variables X_3, X_4, X_5, Y_1 and Y_2 are selected with very low frequencies; the important variables X_1, X_2, Y_3 and Y_4 are selected with frequency 1 or frequency close to 1; and the average model size is close to 4. We also note that the frequency of correctly-fitted model becomes close to $1 - \alpha$ with $n = 700$.

n	X_1	X_2	X_3	X_4	X_5	Y_1	Y_2	Y_3	Y_4	UF	CF	OF	MS
100	1	1	0.01	0.01	0.02	0.02	0.04	0.15	1	0.85	0.13	0.02	3.25
300	1	1	0.01	0.01	0.01	0.01	0.02	0.9	1	0.1	0.86	0.04	3.96
700	1	1	0.02	0.02	0.02	0.02	0.03	1	1	0	0.94	0.06	4.1

Table 2.2. Variable selection results for Model 1 based on 1000 repetitions.

α	X_1	X_2	X_3	X_4	X_5	Y_1	Y_2	Y_3	Y_4	UF	CF	OF	MS
0.01	0	0	1	0.99	0	0.99	0.99	0	0	0.03	0.96	0.01	3.97
0.05	0.01	0.01	1	1	0.02	1	1	0.02	0.02	0	0.95	0.05	4.08
0.1	0.02	0.03	1	1	0.03	1	1	0.05	0.03	0	0.9	0.1	4.17

Table 2.3. Variable selection results for Model 2 based on 1000 repetitions.

Table 2.3 summarizes the variable selection results for Model 2. We fix $n = 300$ and consider $\alpha = 0.01, 0.05, 0.1$. Our backward algorithm works well at all nominal levels. The important variables X_3, X_4, Y_1 and Y_2 are selected with high frequencies, the unimportant variables X_1, X_2, X_5, Y_3 and Y_4 are selected with low frequencies, and the average model size is close to 4. As expected, $\alpha = 0.01$ leads to smaller models and $\alpha = 0.1$ tend to result in larger models. Again, the frequency of correctly-fitted model is close to $1 - \alpha$.

2.4.2 Real Data Analysis

Beta-carotene and retinol are well studied chemical compounds in the human plasma. Several studies suggest that low levels of both compounds in plasma are associated with increased risk of an array of diseases such as cancer, cardiovascular disease, and cataracts. To determine the role of dietary habits and other health related metrics in plasma concentrations of beta-carotene and retinol, Nierenberg et al. (1989) did a cross-sectional study with 12 personal characteristics and dietary metrics for 315 patients with nonmelanoma skin cancer. After removing three categorical variables, we consider $\mathbf{x} = (X_1, \ldots, X_9)^{\mathrm{T}}$. The response variables are $\mathbf{y} = (Y_1, Y_2)^{\mathrm{T}}$, where Y_1 is the plasma concentration of beta-

carotene and Y_2 is the plasma concentration of retinol. After exploratory data analysis, we remove 6 observations with extreme values and get $n = 309$. We apply our proposed dual variable selection procedure from Section 2.3.4 with significance level $\alpha = 0.05$, and end up with $\hat{\mathcal{A}} = \{1, 2, 6, 8\}$ and $\hat{\mathcal{B}} = \{1, 2\}$. This suggests that to further study the multivariate associations between dietary habits and the plasma compound concentrations, we can focus only on six variables X_1, X_2, X_6, X_8, Y_1 and Y_2 instead of the original $\mathbf{x}$ and $\mathbf{y}$.

To demonstrate the effect of variable selection on canonical correlation analysis, we first calculate the first two pairs of canonical covariates (u_1, v_1) and (u_2, v_2) based on the original data, where $\mathbf{x} \in \mathbb{R}^9$ and $\mathbf{y} \in \mathbb{R}^2$. Then we calculate the first two pairs of canonical covariates $(\tilde{u}_1, \tilde{v}_1)$ and $(\tilde{u}_2, \tilde{v}_2)$ based on the reduced data, where $\mathbf{x}_{\hat{\mathcal{A}}} \in \mathbb{R}^4$ and $\mathbf{x}_{\hat{\mathcal{B}}} \in \mathbb{R}^2$. The plots of the canonical covariate from the original data versus the corresponding canonical covariate from the reduced data are provided in Figure 2.1. The scatterplots are close to the dotted 45 degree line, suggesting that the canonical covariates before and after the data reduction largely agree with each other. This implies that the reduced data keeps the canonical information from the original data.

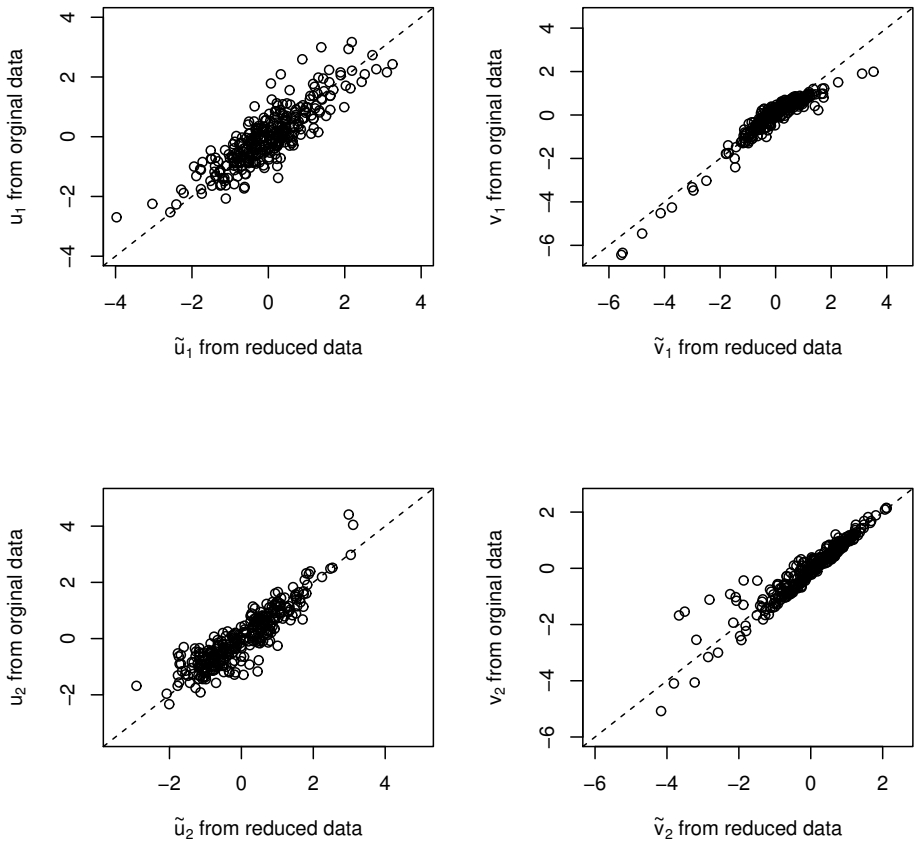

Figure 2.1. Scatterplots of the canonical covariates from the original data versus the canonical covariate from the reduced data.

CHAPTER 3

MARGINAL COORDINATE TEST WITH MULTIVARIATE RESPONSES THROUGH OLS

In this chapter, we study marginal coordinate test with multivariate responses. Specifically, for predictor $\mathbf{x} \in \mathbb{R}^p$ and response $\mathbf{y} \in \mathbb{R}^q$, we consider the following model

$$\mathbf{y} = \boldsymbol{f}(\mathbf{C}^{\mathrm{T}}\mathbf{x}) + \boldsymbol{\epsilon}, \tag{3.19}$$

where $\mathbf{C} \in \mathbb{R}^{p\times d}$ with $d \leq p$, $\boldsymbol{f} : \mathbb{R}^d \mapsto \mathbb{R}^q$, $\boldsymbol{\epsilon} \in \mathbb{R}^q$ with $E(\boldsymbol{\epsilon}) = \mathbf{0}$, and $\boldsymbol{\epsilon}$ is independent of $\mathbf{x}$. We aim to test whether there is significant contribution of one component of $\mathbf{x}$ to $\mathbf{y}$ in the presence of all the other components of $\mathbf{x}$. Under model (3.19), $\mathbf{y}$ depends on $\mathbf{x}$ only through $\mathbf{C}^{\mathrm{T}}\mathbf{x}$. There are several methods to estimate the column space of $\mathbf{C}$ in the sufficient dimension reduction literature, such as Li et al. (2003), Setodji and Cook (2004), Li et al. (2008), and Zhu et al. (2010). To the best of our knowledge, there are no formal tests of predictor contribution based on these existing multi-response sufficient dimension reduction methods.

Our contribution is two-fold. First, we propose a marginal coordinate test based on ordinary least squares (OLS). The asymptotic distribution of the sample level test statistic is carefully studied. Instead of assuming the specific forms of the link function $\boldsymbol{f}$, the asymptotic test is valid under the normality assumption of $\mathbf{x}$ as well as an asymmetry assumption. Second, we provide practical solutions when these assumptions are violated. When the normality assumption is violated, we demonstrate that the asymptotic test with elliptical trimming is still applicable. In the case when the asymmetry assumption is vio-

lated, clustering and elliptical trimming can be used to preprocess the data before applying the asymptotic test.

3.1 Marginal Coordinate Test Through OLS

For $k = 1, \ldots, p$, denote $\mathbf{x}_{-k} = (X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_p)^{\mathrm{T}}$, where the kth component X_k is removed from $\mathbf{x}$. To test the contribution of X_k to $\mathbf{y}$, we consider the following marginal coordinate hypotheses

$$H_0^{-k} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{-k} \text{ v.s. } H_a^{-k} : \mathbf{y} \not\perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{-k}, \tag{3.20}$$

where $\perp\!\!\!\perp$ means statistical independence. Under H_0^{-k}, $\mathbf{y}$ depends on $\mathbf{x}$ only through $\mathbf{x}_{-k}$, and the kth predictor does not contribute to $\mathbf{y}$ in the presence of all the other predictors. In another word, $\mathbf{y}$ does not depend on X_k under H_0^{-k}. An asymptotic test for (3.20) is developed in this section. Without loss of generality, we assume $E(\mathbf{x}) = \mathbf{0}$. Denote $\mathbf{\Sigma}_{\mathbf{x}}$ and $\mathbf{\Sigma}_{\mathbf{y}}$ as the covariance matrices for $\mathbf{x}$ and $\mathbf{y}$, respectively.

3.1.1 Population Level Development

OLS is among the most popular methods in multi-response regression. In the case when $\boldsymbol{f} : \mathbb{R}^d \mapsto \mathbb{R}^q$ in (3.19) is linear, this mapping can be described as premultiplying $\mathbf{W} \in \mathbb{R}^{q \times d}$. Model (3.19) thus becomes $\mathbf{y} = \mathbf{B}_0^{\mathrm{T}}\mathbf{x} + \boldsymbol{\epsilon}$, where $\mathbf{B}_0 = \mathbf{C}\mathbf{W}^{\mathrm{T}} \in \mathbb{R}^{p \times q}$. We denote the classical OLS estimator for $\mathbf{B}_0$ as $\mathbf{B} = \mathbf{\Sigma}_{\mathbf{x}}^{-1} E(\mathbf{x}\mathbf{y}^{\mathrm{T}})$. For $j = 1, \ldots, q$, let $\boldsymbol{\beta}_j = \mathbf{\Sigma}_{\mathbf{x}}^{-1} E(\mathbf{x} Y_j)$ be the jth column of $\mathbf{B}$, which is also the OLS estimator between the jth response Y_j and $\mathbf{x}$. Denote $\boldsymbol{\lambda}_k = (\beta_{1k}, \ldots, \beta_{qk})^{\mathrm{T}}$, where $\beta_{jk} \in \mathbb{R}$ is the kth component of $\boldsymbol{\beta}_j$. Define

$$t_k^{\mathbf{D}} = \boldsymbol{\lambda}_k^{\mathrm{T}} \mathbf{D} \boldsymbol{\lambda}_k, \text{ where } \mathbf{D} \in \mathbb{R}^{q \times q} \text{ is positive definite.} \tag{3.21}$$

With the linear model assumption, β_{jk} describes the regression information between Y_j and X_k, while $t_k^{\mathbf{D}}$ summarizes the regression information between $\mathbf{y}$ and X_k.

The next result bridges $t_k^{\mathbf{D}}$ with the marginal coordinate hypothesis (3.20).

Proposition 3.1.1 *Suppose (a) model (3.19) holds; (b)* $\mathbf{x} \sim N(\mathbf{0}, \mathbf{\Sigma_x})$*; and (c)* $E\{\partial E(Y_j|\mathbf{x})/\partial X_k\} \neq 0$ *for at least one of* $j = 1, 2, \ldots, q$ *under* H_a^{-k}*. Then (i)* H_0^{-k} *is true if and only if* $t_k^{\mathbf{D}} = 0$*; and (ii)* H_a^{-k} *is true if and only if* $t_k^{\mathbf{D}} > 0$*.*

Although we motivate the OLS estimator from the linear model assumption, we see from Proposition 3.1.1 that the linearity between $\mathbf{y}$ and $\mathbf{x}$ is not required for the marginal coordinate test. This result is well-known in the sufficient dimension reduction literature, and is referred to as regression analysis under link violation. See, for example, Li and Duan (1989). Instead of assuming specific link functions between $\mathbf{y}$ and $\mathbf{x}$, the normality assumption naturally bridges the marginal coordinate hypotheses and the OLS estimator through the regression mean function $E(\mathbf{y} \mid \mathbf{x})$. We refer to assumption (c) as the asymmetry assumption. It rules out the situation when $\mathbf{y}$ depends on X_k through a symmetric function between Y_j and X_k. We will revisit this assumption in Section 3.2.

Proposition 3.1.1 suggests that the utility $t_k^{\mathbf{D}}$ is nonzero if and only if $\mathbf{y}$ depends on X_k. To fix the idea, we consider a toy example as follows.

Example 3.1.1 *Let* $\mathbf{x} = (X_1, X_2, X_3)^T$*,* $\mathbf{y} = (Y_1, Y_2)^T$*, and* $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2)^T$*. Assume* $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I_3})$*,* $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I_2})$*, and* $\boldsymbol{\epsilon}$ *is independent of* $\mathbf{x}$*. Let* $Y_1 = X_1^3 + X_2 + \epsilon_1$ *and* $Y_2 = \epsilon_2$*. Obviously* H_0^{-k} *is true for* $k = 3$*, while* H_a^{-k} *is true for* $k = 1$ *and* $k = 2$*. It can be checked that assumption (c) holds in this example. Proposition 3.1.1 then guarantees* $t_3^{\mathbf{D}} = 0$ *and* $t_1^{\mathbf{D}} \neq 0 \neq t_2^{\mathbf{D}}$*. On the other hand, one can calculate* $t_k^{\mathbf{D}}$ *for* $k = 1, 2, 3$*. For the ease of demonstration, we set* $\mathbf{D} = \mathrm{I}_2$*. It follows that* $\boldsymbol{\beta}_1 = E(\mathbf{x}Y_1) = (3, 1, 0)^T$ *and* $\boldsymbol{\beta}_2 = E(\mathbf{x}Y_2) = (0, 0, 0)^T$*. Hence we have* $t_1^{\mathbf{D}} = 9$*,* $t_2^{\mathbf{D}} = 1$*, and* $t_3^{\mathbf{D}} = 0$*.*

Although Proposition 3.1.1 holds for any positive definite matrix $\mathbf{D}$, the matrix $\mathbf{D}$ has to be specified in practice. A natural choice of $\mathbf{D}$ in (3.21) is $\mathbf{\Sigma}_{\mathbf{y}}^{-1}$, and we denote the corresponding $t_k^{\mathbf{D}}$ as

$$t_k = \mathbf{e}_k^{\mathrm{T}} \mathbf{\Sigma}_{\mathbf{x}}^{-1} E(\mathbf{x}\mathbf{y}^{\mathrm{T}}) \mathbf{\Sigma}_{\mathbf{y}}^{-1} E(\mathbf{y}\mathbf{x}^{\mathrm{T}}) \mathbf{\Sigma}_{\mathbf{x}}^{-1} \mathbf{e}_k, \tag{3.22}$$

where $\mathbf{e}_k$ is the kth column of the identity matrix I_p. The next result states that t_k is invariant under full-rank linear transformation of $\mathbf{y}$.

Proposition 3.1.2 *Assume* $E(\mathbf{x}) = \mathbf{0}$. *Let* $\tilde{\mathbf{y}} = \mathbf{\Gamma}\mathbf{y} + \mathbf{a}$, *where* $\mathbf{\Gamma} \in \mathbb{R}^{q \times q}$ *is nonsingular and* $\mathbf{a} \in \mathbb{R}^q$. *Denote* $\tilde{t}_k = \mathbf{e}_k^{\tau} \mathbf{\Sigma}_{\mathbf{x}}^{-1} E(\mathbf{x}\tilde{\mathbf{y}}^{\tau}) \mathbf{\Sigma}_{\tilde{\mathbf{y}}}^{-1} E(\tilde{\mathbf{y}}\mathbf{x}^{\tau}) \mathbf{\Sigma}_{\mathbf{x}}^{-1} \mathbf{e}_k$ *with* $\mathbf{\Sigma}_{\tilde{\mathbf{y}}} = \mathrm{var}(\tilde{\mathbf{y}})$. *Then* $\tilde{t}_k = t_k$.

3.1.2 Sample Level Development

Let $\{(\mathbf{x}_1, \mathbf{y}_1), \cdots, (\mathbf{x}_n, \mathbf{y}_n)\}$ be an i.i.d sample. Denote $\bar{\mathbf{x}} = n^{-1} \sum_{i=1}^n \mathbf{x}_i$, $\widetilde{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}$, $\widehat{\mathbf{\Sigma}}_{\mathbf{x}} = n^{-1} \sum_{i=1}^n \widetilde{\mathbf{x}}_i \widetilde{\mathbf{x}}_i^{\mathrm{T}}$, and $\widehat{\mathbf{\Sigma}}_{\mathbf{xy}} = n^{-1} \sum_{i=1}^n \widetilde{\mathbf{x}}_i \mathbf{y}_i^{\mathrm{T}}$. The sample estimator of $t_k^{\mathbf{D}}$ in (3.21) becomes

$$\hat{t}_k^{\mathbf{D}} = \mathbf{e}_k^{\mathrm{T}} \widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1} \widehat{\mathbf{\Sigma}}_{\mathbf{xy}} \mathbf{D} \widehat{\mathbf{\Sigma}}_{\mathbf{xy}}^{\mathrm{T}} \widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1} \mathbf{e}_k. \tag{3.23}$$

The next theorem describes the asymptotic null distribution of $\hat{t}_k^{\mathbf{D}}$. Intuitively, it implies that large value of $\hat{t}_k^{\mathbf{D}}$ provides evidence against H_0^{-k}. In another word, we conclude $\mathbf{y}$ depends on X_k if $\hat{t}_k^{\mathbf{D}}$ exceeds certain threshold.

Theorem 3.1.1 *Assume all the moments involved are finite. Furthermore, suppose (a) model (3.19) holds; and (b)* $\mathbf{x} \sim N(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{x}})$. *Then*

$$n\hat{t}_k^{\mathbf{D}} \xrightarrow{L} \sum_{j=1}^q v_{k,j}^{\mathbf{D}} \chi_j^2(1)$$

under H_0^{-k}, where "$\xrightarrow{L}$" means convergence in distribution, $\chi_j^2(1)$'s are independent chi-square with one degree of freedom for $j = 1, \ldots, q$, $v_{k,1}^{\mathbf{D}} \geq \ldots \geq v_{k,q}^{\mathbf{D}}$ are the eigenvalues of $\mathbf{\Omega}_k^{\mathbf{D}}$, and the exact form of $\mathbf{\Omega}_k^{\mathbf{D}}$ is provided in the Appendix.

In the case with $\mathbf{D} = \mathbf{\Sigma}_{\mathbf{y}}^{-1}$, the sample estimator of t_k in (3.22) is

$$\hat{t}_k = \mathbf{e}_k^{\mathrm{T}} \widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1} \widehat{\mathbf{\Sigma}}_{\mathbf{xy}} \widehat{\mathbf{\Sigma}}_{\mathbf{y}}^{-1} \widehat{\mathbf{\Sigma}}_{\mathbf{xy}}^{\mathrm{T}} \widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1} \mathbf{e}_k, \tag{3.24}$$

where $\widehat{\mathbf{\Sigma}}_{\mathbf{y}} = n^{-1} \sum_{i=1}^{n} (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^{\mathrm{T}}$. The asymptotic distribution of $\hat{t}_k$ is described next.

Corollary 3.1.1 *Assume the same conditions as in Theorem 3.1.1. Then*

$$n\hat{t}_k \xrightarrow{L} \sum_{j=1}^{q} v_{k,j} \chi_j^2(1)$$

under H_0^{-k}, where $v_{k,1} \geq \ldots \geq v_{k,q}$ are the eigenvalues of $\mathbf{\Omega}_k$ and the exact form of $\mathbf{\Omega}_k$ is provided in the Appendix.

The asymptotic distribution in Corollary 3.1.1 needs to be estimated in practice. Specifically, let $\hat{\mathbf{\Omega}}_k$ be the sample estimator of $\mathbf{\Omega}_k$, and let $\hat{v}_{k,1} \geq \ldots \geq \hat{v}_{k,q}$ be the eigenvalues of $\hat{\mathbf{\Omega}}_k$. The asymptotic null distribution of $n\hat{t}_k$ can then be approximated by $\sum_{j=1}^{q} \hat{v}_{k,j} \chi_j^2(1)$. Denote $\boldsymbol{\zeta}_k = (\hat{v}_{k,1}, \ldots, \hat{v}_{k,q})^{\mathrm{T}}$ and let $\mathbf{\Xi} \in \mathbb{R}^{m \times q}$ consist of i.i.d. $\chi^2(1)$ realizations. Then the m elements of $\mathbf{\Xi}\boldsymbol{\zeta}_k$ become realizations of the approximate asymptotic distribution of $n\hat{t}_k$ under H_0^{-k}. The proportion of these m elements greater than $n\hat{t}_k$ become the approximate p-value for testing the marginal coordinate hypotheses in (3.20). For a given significance level α, we reject H_0^{-k} if this p-value is smaller than α. We use $m = 500$ in our numerical studies.

3.1.3 Algorithm for Model-Free Variable Selection

We propose a backward procedure to carry out the variable selection. Let $\{(\mathbf{x}_1, \mathbf{y}_1), \cdots, (\mathbf{x}_n, \mathbf{y}_n)\}$ be an i.i.d sample.

1. Calculate $\bar{\mathbf{x}} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i$ and $\bar{\mathbf{y}} = n^{-1}\sum_{i=1}^{n}\mathbf{y}_i$. Let $\widetilde{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}$, $\widetilde{\mathbf{y}}_i = \mathbf{y}_i - \bar{\mathbf{y}}$, $\widehat{\mathbf{\Sigma}}_{\mathbf{x}} = n^{-1}\sum_{i=1}^{n}\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^{\mathrm{T}}$, $\widehat{\mathbf{\Sigma}}_{\mathbf{xy}} = n^{-1}\sum_{i=1}^{n}\widetilde{\mathbf{x}}_i\widetilde{\mathbf{y}}_i^{\mathrm{T}}$, $\widehat{\mathbf{\Sigma}}_{\mathbf{yx}} = \widehat{\mathbf{\Sigma}}_{\mathbf{xy}}^{\mathrm{T}}$, $\widehat{\mathbf{\Sigma}}_{\mathbf{y}} = n^{-1}\sum_{i=1}^{n}\widetilde{\mathbf{y}}_i\widetilde{\mathbf{y}}_i^{\mathrm{T}}$.

2. For $k = 1, \ldots, q$, evaluate $\mathbf{p} = (p_1, \ldots, p_q)$ using the following steps

   (a) Calculate $\hat{t}_k = \mathbf{e}_k^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1}\widehat{\mathbf{\Sigma}}_{\mathbf{xy}}\widehat{\mathbf{\Sigma}}_{\mathbf{y}}^{-1}\widehat{\mathbf{\Sigma}}_{\mathbf{xy}}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1}\mathbf{e}_k$, where $\mathbf{e}_k$ is the k-th column of the identity matrix of size q.

   (b) For $i = 1, \ldots, n$, calculate $\widehat{\mathbf{C}}_i = \widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1}\widetilde{\mathbf{x}}_i(\widetilde{\mathbf{y}}_i - \widehat{\mathbf{\Sigma}}_{\mathbf{yx}}\widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1}\widetilde{\mathbf{x}}_i)^{\mathrm{T}}$.

   (c) Calculate $\widehat{\mathbf{\Gamma}}_k = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{C}}_i^{\mathrm{T}}\mathbf{e}_k\mathbf{e}_k^{\mathrm{T}}\widehat{\mathbf{C}}_i$.

   (d) Perform eigenvalue decomposition of $\widehat{\mathbf{\Sigma}}_{\mathbf{y}}^{-1/2}\widehat{\mathbf{\Gamma}}_k\widehat{\mathbf{\Sigma}}_{\mathbf{y}}^{-1/2}$. Denote the eigenvalues as $\hat{v}_{k,1}, \ldots, \hat{v}_{k,q}$.

   (e) Let $\hat{\mathbf{v}}_k = (\hat{v}_{k,1}, \ldots, \hat{v}_{k,q})^{\mathrm{T}}$. Let $\mathbf{\Xi}$ be an $m \times q$ matrix of i.i.d. $\chi^2(1)$ realizations. Then the m elements of $\mathbf{\Xi}\hat{\mathbf{v}}_k$ become realizations of the sum of weighted chi-squared distribution, and the proportion of $n\hat{t}_k$ less than these M elements is the approximate p-value, p_k.

3. For significance level α, denote K as the index of the least significant predictor, that is the one with the highest p-value. If $p_K > \alpha$, remove X_K from $\mathbf{x}$ and repeat steps 1 through 3. Otherwise, stop and report the current $\mathbf{x}$ as the final set of predictors.

Note that if our goal is to perform variable selection for the response, this procedure is still valid by simply switching the roles of $\mathbf{x}$ and $\mathbf{y}$.

3.2 Data Preprocessing for Practical Considerations

The asymptotic test we proposed in Section 2 relies on the normality assumption of $\mathbf{x}$ as well as the asymmetry assumption about the regression mean function. When these assumptions are violated, we need to preprocess the data before we can apply the asymptotic test.

3.2.1 Trimming Through Minimum Volume Ellipsoid

For the validity of the asymptotic test, we need the normality distribution assumption of $\mathbf{x}$. To fix the idea, consider the following toy example.

Example 3.2.1 *Denote* $\mathbf{x} = (X_1, X_2)^T$, $\mathbf{y} = (Y_1, Y_2)^T$, *and* $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2)^T$. *Let* $Y_1 = \exp(X_1) + \epsilon_1$ *and* $Y_2 = \epsilon_2$, *where* $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I_2})$ *and* $\boldsymbol{\epsilon}$ *is independent of* $\mathbf{x}$. *Then* H_0^{-k} *is true for* $k = 2$ *and* H_a^{-k} *is true for* $k = 1$. *Set* $\mathbf{D} = \mathrm{I}_2$. *Consider two cases for the distribution of* $\mathbf{x}$. *In case (i),* $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I_2})$. *It follows that* $\boldsymbol{\beta}_1 = (\exp(1/2), 0)^T$ *and* $\boldsymbol{\beta}_2 = (0, 0)^T$. *Hence we have* $t_1^{\mathbf{D}} = \exp(1)$ *and* $t_2^{\mathbf{D}} = 0$. *In case (ii), let* $X_1 \sim N(0, 1)$ *and* $X_2 = \exp(X_1) - \exp(1/2) + \delta$, *where* $\delta \sim N(0, 1)$ *and* δ *is independent of* X_1. *It can be calculated that* $\boldsymbol{\beta}_1 = c(\exp(1/2), \exp(2) - 2\exp(1))^T$ *and* $\boldsymbol{\beta}_2 = (0, 0)^T$, *where* $c = \{\exp(1) - 1\}^{-2}$. *As a result,* $t_1^{\mathbf{D}} \neq 0 \neq t_2^{\mathbf{D}}$.

From this example, we see when $\mathbf{x}$ is normal in case (i), $t_k^{\mathbf{D}} \neq 0$ if and only if $\mathbf{y}$ depends on X_k. However, when $\mathbf{x}$ is non-elliptical in case (ii), $t_2^{\mathbf{D}} \neq 0$ even though $\mathbf{y}$ does not depend on X_2. Direct application of the asymptotic test no longer works, as we are likely to falsely conclude that $\mathbf{y}$ depends on X_2 due to the unexpectedly large value of $\hat{t}_2^{\mathbf{D}}$.

This naturally motivates us to use the minimum volume ellipsoid (MVE) (Rousseeuw, 1985) to trim the predictors prior to the asymptotic test. MVE is the smallest volume ellipsoid containing a specified fraction of the data. In the robust statistics literature, MVE

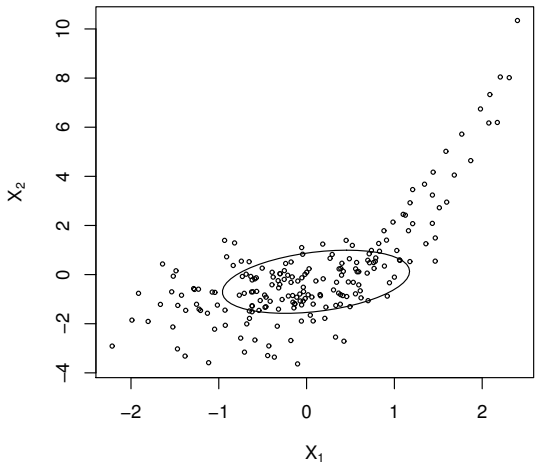

Figure 3.2. Scatterplot of $n = 200$ observations of $\mathbf{x}$ generated from case (ii) of Example 2. The ellipse is the MVE containing 50% of the data.

provides high-breakdown estimators of multivariate location and scatter. The MVE estimators can be computed through a resampling algorithm described in Maronna et al. (2006), which is implemented in the R package $rrcov$ as the function $CovMve$.

The effect of MVE trimming on the normality assumption is demonstrated in Figure 3.1. From case (ii) of Example 2, we generate 200 observations of $\mathbf{x} \in \mathbb{R}^2$. From the scatterplot of X_2 versus X_1 in Figure 3.1, we see the overall distribution of $\mathbf{x}$ is not normal. The ellipse in Figure 3.1 is the MVE containing 50% of the data, which can be well approximated by a normal distribution. The asymptotic test can then be applied to the observations after the elliptical trimming. We examine the numerical performance of asymptotic test with 50% MVE trimming in Section 4.

3.2.2 Clustering Through K-means

Besides the normality assumption, the performance of the asymptotic test also relies on the asymmetry assumption (c) from Proposition 3.1.1. The normality assumption is

required for the asymptotic null distributions of Theorem 3.1.1 and Corollary 3.1.1, and violation of the normality assumption will increase the Type I error. Violation of the asymmetry assumption, on the other hand, will increase the Type II error. To fix the idea, consider the following toy example.

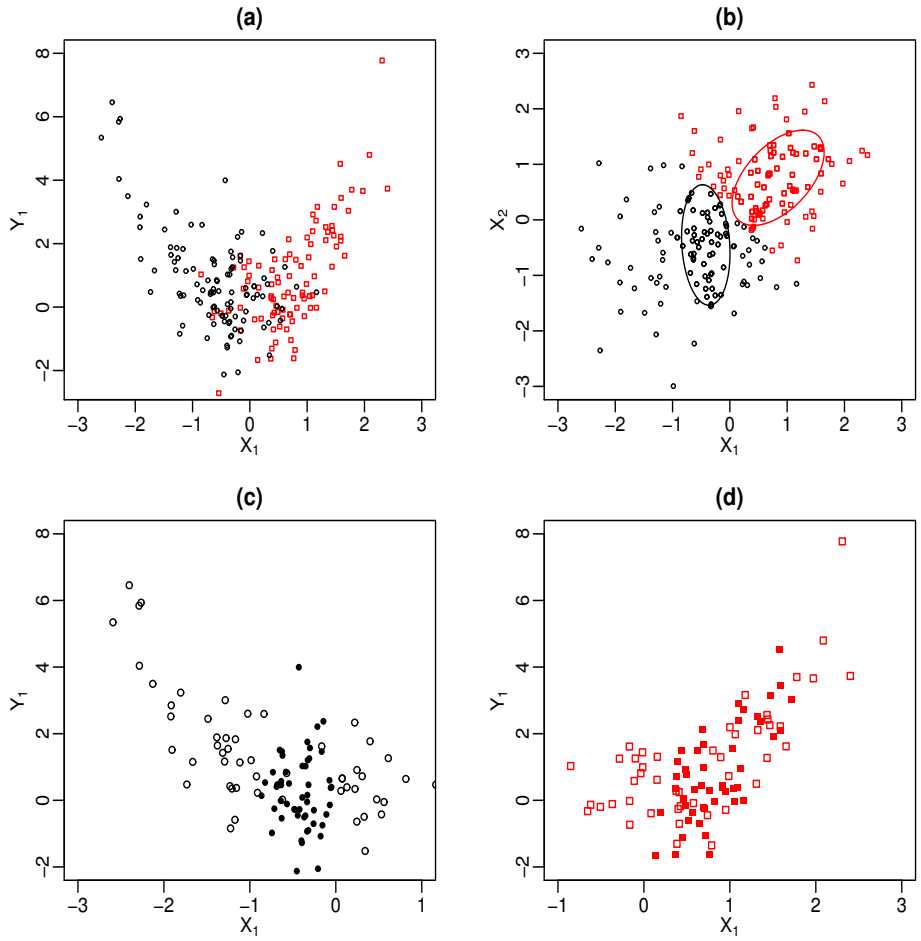

Figure 3.3. Scatterplots of $n = 200$ observations from Example 3. Panel (a): Y_1 versus X_1. Black dots and red squares denote two clusters from K-means. Panel (b): X_2 versus X_1. The ellipses are the MVE containing 50% of the data in each cluster. Panel (c): Y_1 versus X_1 for the first cluster. Solid black dots correspond to the data within the black ellipse in panel (b). Panel (d): Y_1 versus X_1 for the second cluster. Solid red squares correspond to the data within the red ellipse in panel (b).

Example 3.2.2 *Denote* $\mathbf{x} = (X_1, X_2)^T$, $\mathbf{y} = (Y_1, Y_2)^T$, *and* $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2)^T$. *Assume* $\mathbf{x} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{x})$ *with* $\text{var}(X_1) = \text{var}(X_2) = 1$ *and* $\text{cov}(X_1, X_2) = 1/2$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I_2})$, *and* $\boldsymbol{\epsilon}$ *is*

independent of $\mathbf{x}$*. Let* $Y_1 = X_1^2 + \epsilon_1$ *and* $Y_2 = \epsilon_2$*. Then* H_0^{-k} *is true for* $k = 2$ *and* H_a^{-k} *is true for* $k = 1$*. With* $\mathbf{D} = \mathrm{I}_2$*, it can be calculated that* $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = (0, 0)^T$ *and* $t_1^{\mathbf{D}} = t_2^{\mathbf{D}} = 0$*.*

From this example, we see that H_0^{-k} holds for $k = 2$. Theorem 3.1.1 thus guarantees small Type I error when we test the contribution of X_2 to $\mathbf{y}$. On the other hand, $E\{\partial E(Y_j|\mathbf{x})/\partial X_1\} = 0$ for $j = 1, 2$ and H_a^{-k} is true for $k = 1$. Thus the asymmetry assumption (c) is violated due to X_1. When we test the contribution of X_1, we are likely to falsely conclude that $\mathbf{y}$ does not depend on X_1 due to the unexpectedly small value of $\hat{t}_1^{\mathbf{D}}$.

A natural solution here is to use clustering to induce within cluster asymmetry. There are many clustering algorithms, and we use K-means clustering for its ease of implementation. For a comprehensive review of clustering algorithms, see Saxena et al. (2017). There are many criteria to decide the number of clusters. To suit our purpose of asymptotic test after clustering, we simply use $K = 2$ so that the within cluster sample size is relatively large.

The effect of K-means clustering on the asymmetry assumption is demonstrated in Figure 3.2. We generate 200 observations from Example 3. From the scatterplot of Y_1 versus X_1 in panel (a), we observe a U-shaped symmetry, which confirms the violation of the asymmetry assumption globally. After K-means clustering, we see that the asymmetry assumption is no longer violated locally within each cluster.

Although clustering provides remedy for violation of the asymmetry assumption, it may inadvertently lead to violation of the normality assumption. From the scatterplot of X_2 versus versus X_1 in panel (b), we see that although the overall distribution of $\mathbf{x}$ is normal, the normality assumption no longer holds within each cluster. Hence we perform an additional MVE trimming step after clustering. The ellipses in panel (b) contain 50% of the data for each cluster, respectively. Panel (c) and panel (d) provide scatterplots of Y_1 versus X_1 after clustering and trimming.

The numerical performance of asymptotic test with 2-means clustering and 50% MVE trimming is reported in Section 4. For $\ell = 1, 2$, let n_ℓ be the sample size of each cluster after trimming, and denote the corresponding test statistic for H_0^{-k} in (3.20) as $\hat{t}_{k,\ell}$. If the assumptions in Corollary 3.1.1 are satisfied for each cluster after trimming, we get $n_\ell \hat{t}_{k,\ell} \xrightarrow{L} \mathcal{D}_\ell$ for $\ell = 1, 2$, where $\mathcal{D}_\ell$ is the weighted sum of chi-squares specified in Corollary 3.1.1. Due to the independence between the two clusters, we have $\sum_{\ell=1}^{2} n_\ell \hat{t}_{k,\ell} \xrightarrow{L} \sum_{\ell=1}^{2} \mathcal{D}_\ell$. Following a similar procedure described after Corollary 3.1.1, we can approximate the distribution of $\sum_{\ell=1}^{2} \mathcal{D}_\ell$ and get the p-value for testing H_0^{-k} thereafter.

3.3 Numerical Studies

We use synthetic data to examine the performances of the proposed procedures in this section. Let $\mathbf{x} = (X_1, X_2, \ldots, X_p)^{\mathrm{T}}$ be multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma_x} = \{\sigma_{ij}\}$, where $\sigma_{ij} = .5^{|i-j|}$ for $1 \leq i, j \leq p-1$. To induce non-ellipticity between the predictors define $X_p = .5X_1 + ce^{X_1} + \lambda$, where $\lambda \sim \mathcal{N}(0, .8)$. Note that for $c = 0$, the relationship between X_1 and X_p is elliptical. Let $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_q)^{\mathrm{T}}$ be independent of $\mathbf{x}$ and have a multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}_q$, where $\mathbf{I}_q$ is the identity matrix of size q, then we generate $\mathbf{y} = (Y_1, \ldots, Y_q)^{\mathrm{T}}$ from the following models:

$$\text{Model 3: } Y_1 = \sin(X_1) + \sin(X_2) + \epsilon_1, Y_2 = X_2^3 + e^{X_3 + X_4} + \epsilon_2;$$

$$\text{Model 4: } Y_1 = X_1 + \sin(X_2) + \epsilon_1, Y_2 = e^{X_2} + X_3^2 + \epsilon_2.$$

In both models, we set $p = 8$ and $q = 2$. We use 1000 repetitions, and sample sizes $n = 300, 700$ to evaluate the frequencies of rejecting the marginal coordinate hypothesis (3.20) using the OLS-based test. For a fixed significance level $\alpha = .05$, the frequencies for models 1 and 2 are reported in Tables (3.4) and (3.5), respectively. The estimated

powers are given in a boldface font. For model 3, we can see that when the predictors are elliptical, our OLS-based test results in high frequencies for the important predictors X_1, X_2, X_3, and X_4. On the other hand, the estimated Type-I errors are controlled at the nominal level. As demonstrated in example (3.2.1), when the predictors are non-elliptical, we falsely reject the null for X_8 with higher frequencies. Using MVE trimming, the OLS-based test performs much better with frequencies of rejecting the null for X_8 closer to the nominal level. The estimated powers for $n = 300$ are lower when the predictor is clustered and trimmed, but this is potentially due to small samples produced in this case. For model

Model 3	n	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
OLS-based test $\hat{t}_k$									
$c = 0$	300	**1**	**1**	**0.951**	**0.938**	0.049	0.053	0.041	0.036
	700	**1**	**1**	**0.981**	**0.975**	0.042	0.048	0.052	0.047
$c = 0.5$	300	**1**	**1**	**0.94**	**0.936**	0.042	0.058	0.044	0.184
	700	**1**	**1**	**0.987**	**0.982**	0.037	0.039	0.041	0.306
OLS-based test $\hat{t}_k$ with trimmed data									
$c = 0$	300	**0.99**	**0.996**	**0.909**	**0.908**	0.059	0.057	0.049	0.054
	700	**1**	**1**	**0.988**	**0.992**	0.043	0.054	0.047	0.047
$c = 0.5$	300	**0.938**	**0.996**	**0.928**	**0.914**	0.057	0.062	0.063	0.056
	700	**0.999**	**1**	**0.989**	**0.988**	0.057	0.048	0.047	0.064
OLS-based test $\hat{t}_k$ with trimmed clusters									
$c = 0$	300	**0.961**	**0.999**	**0.961**	**0.966**	0.076	0.075	0.072	0.079
	700	**1**	**1**	**0.998**	**0.999**	0.081	0.062	0.047	0.063
$c = 0.5$	300	**0.854**	**0.996**	**0.955**	**0.968**	0.078	0.071	0.089	0.079
	700	**1**	**1**	**1**	**1**	0.074	0.065	0.066	0.074

Table 3.4. Results for model (3). Frequencies of rejecting H_0^{-k} based on 1000 repetitions for $k = 1, \ldots, p$.

4, when the asymmetry assumption is violated, the OLS-based test falsely rejects the null for X_3 with low frequency. Similar to model 3, when the normality assumption is violated, MVE trimming controls the frequency of rejecting the null for X_8 at the nominal level. On the other hand, OLS-based test with clustering and trimming produces high frequencies for X_3 and Type-I errors that are reasonably close to the nominal level. The results for models 1 and 2 from section (2.4.1) are summarized in table 3.6. We can see that the estimated

Model 4	n	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
OLS-based test $\hat{t}_k$									
$c=0$	300	**1**	**1**	**0.057**	0.056	0.047	0.062	0.057	0.061
	700	**1**	**1**	**0.053**	0.039	0.057	0.044	0.051	0.041
$c=0.5$	300	**1**	**1**	**0.059**	0.048	0.066	0.061	0.055	0.214
	700	**1**	**1**	**0.062**	0.051	0.044	0.059	0.046	0.411
OLS-based test $\hat{t}_k$ with trimmed data									
$c=0$	300	**1**	**0.999**	**0.137**	0.063	0.059	0.062	0.063	0.072
	700	**1**	**1**	**0.205**	0.059	0.053	0.064	0.056	0.059
$c=0.5$	300	**0.995**	**1**	**0.134**	0.062	0.062	0.057	0.062	0.059
	700	**1**	**1**	**0.231**	0.053	0.06	0.059	0.053	0.062
OLS-based test $\hat{t}_k$ with trimmed clusters									
$c=0$	300	**0.998**	**1**	**0.907**	0.089	0.089	0.085	0.084	0.095
	700	**1**	**1**	**0.998**	0.056	0.075	0.063	0.061	0.072
$c=0.5$	300	**0.978**	**1**	**0.928**	0.088	0.071	0.084	0.094	0.074
	700	**1**	**1**	**0.997**	0.081	0.071	0.072	0.057	0.067

Table 3.5. Results for model (4). Frequencies of rejecting H_0^{-k} based on 1000 repetitions for $k = 1, \ldots, p$.

powers of the important variables are satisfactory and the estimated Type-I errors are close to the nominal level for both models. In section (2.2), we defined $\mathcal{A}$ as the index set of the

Model	Test	X_1	X_2	X_3	X_4	X_5
Model 1	OLS-based test	**1**	**1**	0.059	0.049	0.051
	OLS-based test trimming	**1**	**1**	0.056	0.066	0.051
	OLS-based test trimmed clusters	**0.99**	**0.991**	0.078	0.07	0.059
Model 2	OLS-based test	0.058	0.051	**1**	**1**	0.058
	OLS-based test trimming	0.053	0.064	**0.992**	**0.966**	0.057
	OLS-based test trimmed clusters	0.075	0.06	**0.988**	**0.973**	0.066

Table 3.6. Results for models (1) and (2). Frequencies of rejecting H_0^{-k} based on 1000 repetitions for $k = 1, \ldots, p$.

active variables among the predictors. Parallel to OF, CF, UF, and CF defined in section (2.2), we define

$$UF = 1000^{-1} \sum_{i=1}^{1000} I(\mathcal{A} \not\subseteq \hat{\mathcal{A}}_i), CF = 1000^{-1} \sum_{i=1}^{1000} I(\mathcal{A} = \hat{\mathcal{A}}_i),$$
$$OF = 1000^{-1} \sum_{i=1}^{1000} I(\mathcal{A} \subset \hat{\mathcal{A}}_i), \text{and } MS = 1000^{-1} \sum_{i=1}^{1000} |\hat{\mathcal{A}}_i|,$$

based in the predictor only. Tables 3.7 and 3.8 summarize the results of the variable selection procedure proposed in section (3.1.3) for models 3 and 4, respectively. Each column represents the frequency a predictor is selected as being active based on 1000 repetitions. We can see that in model 3, the OLS-based test with clustering and trimming outperforms both the original test and the one with just trimming. Using clustering and trimming yields the highest frequencies of CF as well as lowest frequencies of picking up X_8 when $c = .5$. For model 4 we can see similar results, that is, clustering and trimming give the best results

Model 3	n	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	UF	CF	OF	MS
OLS-based test													
$c = 0$	300	**1**	**1**	**0.942**	**0.951**	0.047	0.051	0.048	0.047	0.081	0.759	0.16	4.086
	700	**1**	**1**	**0.982**	**0.983**	0.05	0.047	0.047	0.05	0.028	0.804	0.168	4.159
$c = 0.5$	300	**0.999**	**1**	**0.939**	**0.959**	0.049	0.056	0.04	0.193	0.079	0.649	0.272	4.235
	700	**1**	**1**	**0.985**	**0.986**	0.048	0.056	0.035	0.334	0.022	0.564	0.414	4.444
OLS-based test with trimming													
$c = 0$	300	**0.99**	**0.996**	**0.94**	**0.945**	0.07	0.08	0.072	0.059	0.107	0.684	0.209	4.152
	700	**1**	**1**	**0.988**	**0.991**	0.052	0.071	0.056	0.051	0.015	0.791	0.194	4.209
$c = 0.5$	300	**0.985**	**0.997**	**0.92**	**0.963**	0.065	0.067	0.058	0.078	0.121	0.693	0.186	4.133
	700	**1**	**1**	**0.985**	**0.99**	0.055	0.067	0.052	0.063	0.019	0.78	0.201	4.212
OLS-based test with trimmed clusters													
$c = 0$	300	**0.974**	**0.999**	**0.982**	**0.989**	0.044	0.036	0.044	0.03	0.052	0.83	0.118	4.098
	700	**1**	**1**	**1**	**1**	0.034	0.036	0.032	0.032	0	0.889	0.111	4.134
$c = 0.5$	300	**0.954**	**0.997**	**0.969**	**0.99**	0.033	0.048	0.033	0.049	0.084	0.806	0.11	4.073
	700	**1**	**1**	**1**	**1**	0.028	0.029	0.02	0.039	0	0.893	0.107	4.116

Table 3.7. Variable selection results for model (3) using the OLS-based test. Frequencies based on 1000 repetitions.

for the symmetric response and non-normal predictors.

Model 4	n	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	UF	CF	OF	MS
OLS-based test													
$c = 0$	300	**1**	**1**	**0.058**	0.065	0.064	0.043	0.055	0.064	0.942	0.041	0.017	2.349
	700	**1**	**1**	**0.052**	0.051	0.055	0.061	0.053	0.06	0.948	0.041	0.011	2.332
$c = 0.5$	300	**1**	**1**	**0.075**	0.07	0.061	0.058	0.065	0.2	0.925	0.044	0.031	2.529
	700	**1**	**1**	**0.064**	0.048	0.06	0.064	0.064	0.435	0.936	0.021	0.043	2.735
OLS-based test with trimming													
Model 2	n	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	UF	CF	OF	MS
$c = 0$	300	**0.999**	**0.999**	**0.172**	0.075	0.08	0.073	0.075	0.062	0.828	0.12	0.052	2.535
	700	**1**	**1**	**0.233**	0.079	0.068	0.067	0.064	0.065	0.767	0.157	0.076	2.576
$c = 0.5$	300	**0.997**	**1**	**0.156**	0.077	0.075	0.068	0.068	0.079	0.844	0.114	0.042	2.52
	700	**1**	**1**	**0.242**	0.077	0.06	0.074	0.079	0.095	0.758	0.168	0.074	2.627
OLS-based test with trimmed clusters													
Model 2	n	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	UF	CF	OF	MS
$c = 0$	300	**0.984**	**0.99**	**0.88**	0.037	0.03	0.026	0.033	0.026	0.126	0.749	0.125	3.006
	700	**1**	**1**	**0.992**	0.03	0.024	0.026	0.034	0.034	0.008	0.87	0.122	3.14
$c = 0.5$	300	**0.979**	**0.992**	**0.876**	0.027	0.038	0.037	0.03	0.028	0.136	0.75	0.114	3.007
	700	**1**	**1**	**0.99**	0.033	0.018	0.025	0.029	0.032	0.01	0.866	0.124	3.127

Table 3.8. Variable selection results for model (4) using the OLS-based test. Frequencies based on 1000 repetitions.

CHAPTER 4

DISCUSSION

In recent years, we have seen a rapid increase in computing power as well as huge drops in the cost of storage, which, in turn, led to a steady growth of databases' size across many disciplines. This development captured the attention of many researchers to reduce the complexity in datasets by reducing the number of variables for statistical analysis. Such goal can be achieved using both variable selection and dimension reduction.

In this dissertation we introduced two test statistics for marginal coordinate hypothesis based on sufficient dimension reduction concepts. Our tests are model-free and designed to work with multivariate responses. The first test is a CCA-based trace test for the dual marginal coordinate hypotheses, where the role of the predictor and the response is not important. Under the normality and asymmetry assumptions we derived and studied the asymptotic properties of test. The validity of the asymptotic test is then justified through simulation studies. Based on this novel test, we design a joint backward selection algorithm for dual model-free variable selection. The finite sample performances of the proposed test and the variable selection algorithm are very promising. The dual variable selection and feature screening in the case of diverging p and q is worth future investigation.

The second test is an OLS-based test for marginal coordinate hypotheses. The normality and asymmetry assumptions are necessary to derive the asymptotic distribution of the test. To relax the restrictions of the normality assumption, we preprocess the data by trimming via the minimum volume ellipsoid. The results of the asymptotic test after trimming are satisfying when the normality assumption is violated. In addition, when the link function between the predictor and the response is asymmetric, we proposed using K-means

clustering to break off the symmetry. Each cluster is then trimmed to avoid some side effects that may be introduced by clustering. Model-free variable selection can be applied based on our test using backward or forward algorithms.

REFERENCES

REFERENCES

Aragon, Y. (1997). A Gauss implementation of multivariate sliced inverse regression. *Computational Statistics*, **12**, 355-372.

Candés, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, **35**, 2313–2351.

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, **89**, 177-189.

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* (pp. 18-25). American Statistical Association Alexandria, VA. Chicago.

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

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

Cook, R. D. (1998). Principal Hessian directions revisited. *Journal of the American Statistical Association*, **93**, 84-94.

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

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

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

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

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.

Fung, W. K., He, X., Liu, L., & Shi, P. (2002). Dimension reduction based on canonical correlation. *Statistica Sinica*, 1093-1113.

Hotelling, H. (1936). Relations between two sets of variables. *Biometrika*, **58**, 433-451.

Hsing, T. (1999). Nearest neighbor inverse regression. *Annals of statistics*, 697-731.

Huang, J., Horowitz, J. L. and Wei, F. (2010). Variable selection in nonparametric additive models. *The Annals of Statistics*, **38**, 2282-2313.

Iaci, R., Yin, X. and Zhu, L. X. (2016). The dual central subspaces in dimension reduction. *Journal of Multivariate Analysis*, **145**, 178–189.

Jiang, B. and Liu, J. (2014). Variable selection for general index models via sliced inverse regression. *The Annals of Statistics*, **42**, 1751–1786.

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

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

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

Li, K. C., Aragon, Y., Shedden, K. and Agnan, C. T. (2003). Dimension reduction for multivariate response data. *Journal of the American Statistical Association*, **98**, 99–109.

Li, K. C. and Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics*, **17**, 1009-1052.

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

Li, L., Cook, R. D. and Nachtsheim, C. J. (2005). Model-free variable selection. *Journal of the Royal Statistical Society, Series B*, **67**, 285–299.

Liu, J. Y., Li, R. and Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh dimensional covariates. *Journal of the American Statistical Association*, **109**, 266-274.

Maronna R. A., Martin D. R. and Yohai V. J. (2006). *Robust Statistics: Theory and Methods*. New York: Wiley.

Nierenberg, D. W., Stukel, T. A., Baron, J. A., Dain, B. J., Greenberg, E. R. and Skin Cancer Prevention Study Group (1989). Determinants of plasma levels of beta-carotene and retinol. *American Journal of Epidemiology*, **130**, 511–521.

Rousseeuw P. J. (1985) *Multivariate estimation with high breakdown point*. In: Grossmann W., Pflug G., Vincze I., Wertz W., eds. *Mathematical Statistics and Applications*, Dordrecht: Reidel Publishing Company.

Saracco, J. (2005). Asymptotics for pooled marginal slicing estimator based on SIR_α approach. *Journal of multivariate Analysis*, **96**, 117-135.

Saxena, A., Prasad, M. Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W. P. and Lin, C. T. (2017). A review of clustering techniques and developments. *Neurocomputing*, **267**, 664–681.

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

Shao, Y., Cook, R. D. and Weisberg, S. (2007). Mariginal tests with sliced average variance estimation. *Biometrika*, **94**, 285–296.

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

Wang, L., Xue, L., Qu, A. and Liang, H. (2014). Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates. *The Annals of Statistics*, **42**, 592-624.

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

Yu, Z. and Dong, Y. (2016). Model-free coordinate test and variable selection via directional regression. *Statistica Sinica*, **26**, 1159–1174.

Yu, Z., Dong, Y. and Fang, Y. (2010). Marginal coordinate tests for central mean subspace with principal Hessian directions. *Chinese Journal of Applied Probability and Statistics*, **26**, 544–552.

Yu, Z., Dong, Y. and Zhu, L. X. (2016). Trace pursuit: a general framework for model-free variable selection. *Journal of the American Statistical Association*, **111**, 813–821.

Yu, Z. and Dong, Y. (2016). Model-free coordinate test and variable selection via directional regression. *Statistica Sinica*, **26**, 1159–1174.

Yu, Z., Dong, Y. and Fang, Y. (2010). Marginal coordinate tests for central mean subspace with principal Hessian directions. *Chinese Journal of Applied Probability and Statistics*, **26**, 544–552.

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

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

Zhu, L. P., Zhu, L. X. and Wen S. Q. (2010). On dimension reduction in regresssion with multivariate responses. *Statistica Sinica*, **20**, 1291–1307.

APPENDICES

APPENDIX A

PROOFS FOR PROPOSITIONS AND THEOREMS IN CHAPTER 1

PROOF OF PROPOSITION 1.3.1. Using law of total expectation

$$E(\mathbf{x}Y) = E(E(\mathbf{x}Y|\mathbf{x})) = E(\mathbf{x}E(Y|\mathbf{x})). \tag{A.25}$$

Suppose $\boldsymbol{\beta}$ is a basis in the central mean space, i.e. span($\boldsymbol{\beta}$) $= \mathcal{S}_{E(Y|\mathbf{x})}$. And because $E(Y|\mathbf{x}) = E(Y|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$, we have

$$E(\mathbf{x}E(Y|\mathbf{x})) = E(\mathbf{x}E(Y|\mathbf{x}, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})) = E(\mathbf{x}E(Y|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})). \tag{A.26}$$

Denote $\mathbf{P}_{\boldsymbol{\beta}} = \boldsymbol{\beta}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathrm{T}}$ as the projection matrix into span($\boldsymbol{\beta}$), the central mean space. Then, under the LCM assumption (1.3.1), we have

$$E(\mathbf{x}E(Y|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})) = E(YE(\mathbf{x}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})) = E(Y\mathbf{P}_{\boldsymbol{\beta}}\mathbf{x}) = \mathbf{P}_{\boldsymbol{\beta}}E(\mathbf{x}Y). \tag{A.27}$$

Thus, (A.25), (A.26), and (A.27) together imply $E(\mathbf{x}Y) = \mathbf{P}_{\boldsymbol{\beta}}E(\mathbf{x}Y)$. That is, the OLS estimator $E(\mathbf{x}Y)$ is its projection onto $\mathcal{S}(\boldsymbol{\beta})$, and thus $E(\mathbf{x}Y) \in \mathcal{S}_{E(Y|\mathbf{x})}$. □

PROOF OF PROPOSITION 1.3.2. Let $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ be a basis in the central space, i.e. span($\boldsymbol{\beta}$) $= \mathcal{S}_{Y|\mathbf{x}}$. We have

$$E(\mathbf{x}|Y) = E(E(\mathbf{x}|Y, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})|Y).$$

Since $Y \perp\!\!\!\perp \mathbf{x}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$, then

$$E(E(\mathbf{x}|Y, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})|Y) = E(E(\mathbf{x}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})|Y).$$

Denote $\mathbf{P}_{\boldsymbol{\beta}} = \boldsymbol{\beta}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathrm{T}}$ as the projection matrix into span($\boldsymbol{\beta}$), the central space. Then, under the LCM assumption (1.3.1), we have

$$E(E(\mathbf{x}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})|Y) = E(\mathbf{P}_{\boldsymbol{\beta}}\mathbf{x}|Y) = \mathbf{P}_{\boldsymbol{\beta}}E(\mathbf{x}|Y).$$

That is, the inverse regression $E(\mathbf{x}|Y)$ is in $\mathcal{S}(\boldsymbol{\beta})$, and thus $E(\mathbf{x}|Y) \in \mathcal{S}_{Y|\mathbf{x}}$. □

PROOF OF PROPOSITION 1.3.4. Let $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ be a basis in the central space, then by law of total variance

$$Var(\mathbf{x}|Y) = E(Var(\mathbf{x}|Y, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})|Y) + Var(E(\mathbf{x}|Y, \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})|Y).$$

Since $Y \perp\!\!\!\perp \mathbf{x}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$, then

$$Var(\mathbf{x}|Y) = E(Var(\mathbf{x}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})|Y) + Var(E(\mathbf{x}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})|Y).$$

Denote $\mathbf{P}_{\boldsymbol{\beta}} = \boldsymbol{\beta}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathrm{T}}$ as the projection matrix into span($\boldsymbol{\beta}$), the central space. Then, under the LCM assumption (1.3.1), we have

$$Var(E(\mathbf{x}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})|Y) = Var(\mathbf{P}_{\boldsymbol{\beta}}\mathbf{x}|Y) = \mathbf{P}_{\boldsymbol{\beta}}Var(\mathbf{x}|Y)\mathbf{P}_{\boldsymbol{\beta}}$$

Denote $\mathbf{Q}_{\boldsymbol{\beta}} = \mathrm{I}_p - \mathbf{P}_{\boldsymbol{\beta}}$ as the orthogonal projection matrix of span($\boldsymbol{\beta}$). The under the CCV assumption (1.3.2), we have

$$E(Var(\mathbf{x}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})|Y) = E((\mathrm{I}_p - \mathbf{P}_{\boldsymbol{\beta}})|Y) = \mathrm{I}_p - \mathbf{P}_{\boldsymbol{\beta}} = \mathbf{Q}_{\boldsymbol{\beta}}.$$

Therefore,

$$\mathrm{I}_p - Var(\mathbf{x}|Y) = \mathbf{P}_{\boldsymbol{\beta}} Var(\mathbf{x}|Y)\mathbf{P}_{\boldsymbol{\beta}},$$

That is, $\mathrm{I}_p - Var(\mathbf{x}|Y = y)$ is in the central space $\mathcal{S}_{Y|\mathbf{x}}$. □

PROOF OF PROPOSITION 1.3.5. Without loss of generality let $E(Y) = 0$. Let $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ be a basis in the central mean space, then by law of total expectation

$$\bar{\mathbf{H}}_{\mathbf{x}} = E(Y\mathbf{x}\mathbf{x}^{\mathrm{T}}) = E(E(Y\mathbf{x}\mathbf{x}^{\mathrm{T}}|\mathbf{x})) = E(E(Y|\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}).$$

Then, under the LCM assumption (1.3.1)

$$E(E(Y|\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}) = E(E(Y|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}).$$

By the self adjoint property, we have

$$E(E(Y|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\mathbf{x}\mathbf{x}^{\mathrm{T}}) = E(YE(\mathbf{x}\mathbf{x}^{\mathrm{T}}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})).$$

Denote $\mathbf{P}_{\boldsymbol{\beta}} = \boldsymbol{\beta}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathrm{T}}$ as the projection matrix into span($\boldsymbol{\beta}$), the central space and $\mathbf{Q}_{\boldsymbol{\beta}} = \mathrm{I}_p - \mathbf{P}_{\boldsymbol{\beta}}$ as the orthogonal projection matrix of span($\boldsymbol{\beta}$). The under the LCM and CCV assumptions, we have

$$\begin{aligned} E(\mathbf{x}\mathbf{x}^{\mathrm{T}}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) &= Var(\mathbf{x}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) + E(\mathbf{x}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})E(\mathbf{x}^{\mathrm{T}}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) \\ &= \mathbf{Q}_{\boldsymbol{\beta}} + \mathbf{P}_{\boldsymbol{\beta}}\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{P}_{\boldsymbol{\beta}}. \end{aligned}$$

Since $E(Y) = 0$,

$$\begin{aligned} E(E(\mathbf{x}\mathbf{x}^{\mathrm{T}}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})Y) &= E(Y[\mathbf{Q}_{\boldsymbol{\beta}} + \mathbf{P}_{\boldsymbol{\beta}}\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{P}_{\boldsymbol{\beta}}]) \\ &= E(Y\mathbf{Q}_{\boldsymbol{\beta}}) + E(Y(\mathbf{P}_{\boldsymbol{\beta}}\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{P}_{\boldsymbol{\beta}})) \\ &= \mathbf{P}_{\boldsymbol{\beta}}E(Y\mathbf{x}\mathbf{x}^{\mathrm{T}})\mathbf{P}_{\boldsymbol{\beta}}. \end{aligned}$$

That is, the average Hessian matrix $\bar{\mathbf{H}}_{\mathbf{x}}$ is in $\mathcal{S}(\boldsymbol{\beta})$, and thus $\bar{\mathbf{H}}_{\mathbf{x}} \in \mathcal{S}_{E(Y|\mathbf{x})}$. □

APPENDIX B

PROOFS FOR PROPOSITIONS AND THEOREMS IN CHAPTER 2

PROOF OF PROPOSITION 2.2.1. The proof is similar to Proposition 1 in Iaci et al. (2016), and is thus omitted. □

PROOF OF PROPOSITION 2.3.1. For part (i), note that $\mathrm{Span}(\mathbf{M}) = \mathrm{Span}\{\mathrm{E}(\mathbf{z}\mathbf{w}^{\mathrm{T}})\}$. Plug in $\mathbf{z} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}\mathbf{x}$ and $\mathbf{w} = \boldsymbol{\Sigma}_{\mathbf{y}}^{-1/2}\mathbf{y}$, and all we need to prove becomes

$$\mathrm{Span}\{\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathrm{E}(\mathbf{x}\mathbf{y}^{\mathrm{T}})\} = \mathrm{Span}\{\boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}\mathrm{E}(\mathbf{z}\mathbf{w}^{\mathrm{T}})\} \subseteq \mathcal{S}_{\mathbf{y}|\mathbf{x}} = \mathrm{Span}(\boldsymbol{\beta}). \tag{B.28}$$

From the law of iterated expectations and the fact that $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$, we have

$$\mathrm{E}(\mathbf{x}\mathbf{y}^{\mathrm{T}}) = \mathrm{E}\{\mathbf{x}\mathrm{E}^{\mathrm{T}}(\mathbf{y}|\mathbf{x})\} = \mathrm{E}\{\mathbf{x}\mathrm{E}^{\mathrm{T}}(\mathbf{y}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\}. \tag{B.29}$$

From the property of conditional expectation and the assumption that $\mathrm{E}(\mathbf{x}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ is linear in $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$, we have

$$\mathrm{E}\{\mathbf{x}\mathrm{E}^{\mathrm{T}}(\mathbf{y}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\} = \mathrm{E}\{\mathrm{E}(\mathbf{x}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\mathbf{y}^{\mathrm{T}}\} = \boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\beta}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathrm{T}}\mathrm{E}(\mathbf{x}\mathbf{y}^{\mathrm{T}}). \tag{B.30}$$

(B.29) and (B.30) together lead to

$$\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathrm{E}(\mathbf{x}\mathbf{y}^{\mathrm{T}}) = \boldsymbol{\beta}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathrm{T}}\mathrm{E}(\mathbf{x}\mathbf{y}^{\mathrm{T}}). \tag{B.31}$$

(B.28) follows from (B.31) and proof of part (i) is done. Proof of part (ii) is similar to the proof of part (i), and is thus omitted. □

PROOF OF PROPOSITION 2.3.2. For part (i), assume $\mathbf{x}_{\mathcal{F}} \in \mathbb{R}^{p_1}$ and $\mathbf{x}_{\mathcal{F}^c} \in \mathbb{R}^{p_2}$ with $p_1 + p_2 = p$. Let $\mathbf{x} = (\mathbf{x}_{\mathcal{F}}^{\mathrm{T}}, \mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}})^{\mathrm{T}}$. Define $\mathbf{C}$ and $\mathbf{D}$ as

$$\mathbf{C} = \begin{pmatrix} \mathrm{I}_{p_1} & \mathbf{0} \\ -\mathrm{E}(\mathbf{x}_{\mathcal{F}^c}\mathbf{x}_{\mathcal{F}}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1} & \mathrm{I}_{p_2} \end{pmatrix} \text{ and } \mathbf{D} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}} \end{pmatrix}.$$

Then $\mathbf{C}\mathbf{x} = (\mathbf{x}_{\mathcal{F}}^{\mathrm{T}}, \boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{C}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{C}^{\mathrm{T}} = \mathbf{D}$ and $\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} = \mathbf{C}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{C}$. It follows that

$$\begin{aligned}
&\mathrm{tr}(\mathbf{M}) = \mathrm{tr}\left\{\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathrm{E}(\mathbf{x}\mathbf{y}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\mathbf{x}^{\mathrm{T}})\right\} = \mathrm{tr}\left\{\mathbf{C}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{C}\mathrm{E}(\mathbf{x}\mathbf{y}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\mathbf{x}^{\mathrm{T}})\right\} \\
&= \mathrm{tr}\left\{\begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{-1} \end{pmatrix}\begin{pmatrix} \mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{y}^{\mathrm{T}}) \\ \mathrm{E}(\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}\mathbf{y}^{\mathrm{T}}) \end{pmatrix}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\left(\mathrm{E}(\mathbf{y}\mathbf{x}_{\mathcal{F}}^{\mathrm{T}}), \mathrm{E}(\mathbf{y}\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{\mathrm{T}})\right)\right\} \\
&= \mathrm{tr}\left\{\begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{y}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\mathbf{x}_{\mathcal{F}}^{\mathrm{T}}) & \boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{y}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{\mathrm{T}}) \\ \boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}\mathbf{y}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\mathbf{x}_{\mathcal{F}}^{\mathrm{T}}) & \boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}\mathbf{y}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{\mathrm{T}}) \end{pmatrix}\right\} \\
&= \mathrm{tr}\{\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{y}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\mathbf{x}_{\mathcal{F}}^{\mathrm{T}})\} + \mathrm{tr}\{\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}\mathbf{y}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{\mathrm{T}})\}.
\end{aligned}$$

Together with $\mathrm{tr}(\mathbf{M}_{\mathcal{F}}) = \mathrm{tr}\{\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{y}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\mathbf{x}_{\mathcal{F}}^{\mathrm{T}})\}$, we get

$$\delta_{-\mathcal{F}} = \mathrm{tr}(\mathbf{M}) - \mathrm{tr}(\mathbf{M}_{\mathcal{F}}) = \mathrm{tr}\{\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}\mathbf{y}^{\mathrm{T}})\}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{\mathrm{T}})\}. \tag{B.32}$$

For part (ii), the assumption that $\mathrm{E}(\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}})$ is a linear function of $\mathbf{x}_{\mathcal{F}}$ implies $\mathrm{E}(\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}) = \mathrm{E}(\mathbf{x}_{\mathcal{F}^c}\mathbf{x}_{\mathcal{F}}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathbf{x}_{\mathcal{F}}$. It follows that

$$\mathrm{E}\{\mathrm{E}(\mathbf{x}_{\mathcal{F}^c}\mathbf{x}_{\mathcal{F}}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathbf{x}_{\mathcal{F}}\mathbf{y}^{\mathrm{T}}\} = \mathrm{E}\{\mathrm{E}(\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}})\mathbf{y}^{\mathrm{T}}\} = \mathrm{E}\{\mathbf{x}_{\mathcal{F}^c}\mathrm{E}^{\mathrm{T}}(\mathbf{y}|\mathbf{x}_{\mathcal{F}})\}. \tag{B.33}$$

Under $H_0^{\mathcal{F}} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$, we have $\mathrm{E}(\mathbf{y}|\mathbf{x}) = \mathrm{E}(\mathbf{y}|\mathbf{x}_{\mathcal{F}})$. Thus

$$\mathrm{E}\{\mathbf{x}_{\mathcal{F}^c}\mathrm{E}^{\mathrm{T}}(\mathbf{y}|\mathbf{x}_{\mathcal{F}})\} = \mathrm{E}\{\mathbf{x}_{\mathcal{F}^c}\mathrm{E}^{\mathrm{T}}(\mathbf{y}|\mathbf{x})\} = \mathrm{E}(\mathbf{x}_{\mathcal{F}^c}\mathbf{y}^{\mathrm{T}}). \tag{B.34}$$

The definition $\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}} = \mathbf{x}_{\mathcal{F}^c} - \mathrm{E}(\mathbf{x}_{\mathcal{F}^c}\mathbf{x}_{\mathcal{F}}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathbf{x}_{\mathcal{F}}$ together with (B.33) and (B.34) leads to $\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}\mathbf{y}^{\mathrm{T}}) = \mathbf{0}$. It follows from part (i) that $\delta_{-\mathcal{F}} = 0$ under $H_0^{\mathcal{F}}$. □

PROOF OF PROPOSITION 2.3.3. The proof is similar to the proof of Proposition 2.3.2, and is thus omitted. □

PROOF OF PROPOSITION 2.3.4. For part (i), assume $\mathbf{y}_{\mathcal{G}} \in \mathbb{R}^{q_1}$ and $\mathbf{y}_{\mathcal{G}^c} \in \mathbb{R}^{q_2}$ with $q_1 + q_2 = q$. Let $\mathbf{y} = (\mathbf{y}_{\mathcal{G}}^{\mathrm{T}}, \mathbf{y}_{\mathcal{G}^c}^{\mathrm{T}})^{\mathrm{T}}$. Define $\mathbf{K}$ and $\mathbf{O}$ as

$$\mathbf{K} = \begin{pmatrix} \mathrm{I}_{q_1} & \mathbf{0} \\ -\mathrm{E}(\mathbf{y}_{\mathcal{G}^c}\mathbf{y}_{\mathcal{G}}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}}}^{-1} & \mathrm{I}_{q_2} \end{pmatrix} \text{ and } \mathbf{O} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}} \end{pmatrix}.$$

Then $\mathbf{K}\mathbf{y} = (\mathbf{y}_{\mathcal{G}}^{\mathrm{T}}, \boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{K}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{K}^{\mathrm{T}} = \mathbf{O}$ and $\boldsymbol{\Sigma}_{\mathbf{y}}^{-1} = \mathbf{K}^{\mathrm{T}}\mathbf{O}^{-1}\mathbf{K}$. Thus

$$\begin{aligned}
&\mathrm{tr}(\mathbf{M}_{\mathcal{F}}) = \mathrm{tr}\left\{\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{y}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathrm{E}(\mathbf{y}\mathbf{x}_{\mathcal{F}}^{\mathrm{T}})\right\} = \mathrm{tr}\left\{\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}\mathbf{y}^{\mathrm{T}})\mathbf{K}^{\mathrm{T}}\mathbf{O}^{-1}\mathbf{K}\mathrm{E}(\mathbf{y}\mathbf{x}^{\mathrm{T}})\right\} \\
&= \mathrm{tr}\left\{\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\left(\mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{y}_{\mathcal{G}}^{\mathrm{T}}), \mathrm{E}(\mathbf{x}_{\mathcal{F}}\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{\mathrm{T}})\right)\begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}}}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{-1} \end{pmatrix}\begin{pmatrix} \mathrm{E}(\mathbf{y}_{\mathcal{G}}\mathbf{x}_{\mathcal{F}}^{\mathrm{T}}) \\ \mathrm{E}(\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}\mathbf{x}_{\mathcal{F}}^{\mathrm{T}}) \end{pmatrix}\right\} \\
&= \mathrm{tr}\{\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{y}_{\mathcal{G}}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}}}^{-1}\mathrm{E}(\mathbf{y}_{\mathcal{G}}\mathbf{x}_{\mathcal{F}}^{\mathrm{T}})\} + \mathrm{tr}\{\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{-1}\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}\mathbf{x}_{\mathcal{F}}^{\mathrm{T}})\} \\
&= \mathrm{tr}(\mathbf{M}_{\mathcal{F}}^{\mathcal{G}}) + \mathrm{tr}\{\boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{-1}\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}\mathbf{x}_{\mathcal{F}}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{\mathrm{T}})\}.
\end{aligned}$$

Together with (B.32) from the proof of Proposition 2.3.2, we get the desired result in part (i).

For part (ii), we have seen that $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$ leads to $\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}\mathbf{y}^{\mathrm{T}}) = \mathbf{0}$ from the proof of Proposition 2.3.2. Following similar steps, we can show that $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}$ leads to $\mathrm{E}(\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}\mathbf{x}_{\mathcal{F}}^{\mathrm{T}}) = \mathbf{0}$. It follows from part (i) that $\delta_{-\mathcal{F}}^{-\mathcal{G}} = 0$ under $H_0^{\mathcal{F},[\mathcal{G}]} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$ and $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}$. □

We need Lemma B.1 and Lemma B.2 before we prove Theorem 2.3.1. Let $\bar{\mathbf{x}}_{\mathcal{F}} = n^{-1}\sum_{i=1}^n \mathbf{x}_{\mathcal{F}}^{(i)}$, $\tilde{\mathbf{x}}_{\mathcal{F}}^{(i)} = \mathbf{x}_{\mathcal{F}}^{(i)} - \bar{\mathbf{x}}_{\mathcal{F}}$, and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}_{\mathcal{F}}} = n^{-1}\sum_{i=1}^n \tilde{\mathbf{x}}_{\mathcal{F}}^{(i)}(\tilde{\mathbf{x}}_{\mathcal{F}}^{(i)})^{\mathrm{T}}$. Similarly we define $\tilde{\mathbf{y}}^{(i)}$ and $\tilde{\mathbf{x}}_{\mathcal{F}^c}^{(i)}$. Let $\mathrm{E}_n(\mathbf{x}_{\mathcal{F}}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}}) = n^{-1}\sum_{i=1}^n \tilde{\mathbf{x}}_{\mathcal{F}}^{(i)}\left(\tilde{\mathbf{x}}_{\mathcal{F}^c}^{(i)}\right)^{\mathrm{T}}$, $\hat{\boldsymbol{\gamma}}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{(i)} = \tilde{\mathbf{x}}_{\mathcal{F}^c}^{(i)} - \mathrm{E}_n(\mathbf{x}_{\mathcal{F}^c}\mathbf{x}_{\mathcal{F}}^{\mathrm{T}})\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}_{\mathcal{F}}}^{-1}\tilde{\mathbf{x}}_{\mathcal{F}}^{(i)}$, and $\mathrm{E}_n(\mathbf{y}\hat{\boldsymbol{\gamma}}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{\mathrm{T}}) = n^{-1}\sum_{i=1}^n \tilde{\mathbf{y}}^{(i)}\left(\hat{\boldsymbol{\gamma}}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{(i)}\right)^{\mathrm{T}}$. Define

$$\boldsymbol{\Pi}^{(i)} = \left\{\tilde{\mathbf{y}}^{(i)} - \mathrm{E}(\mathbf{y}\mathbf{x}_{\mathcal{F}}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\tilde{\mathbf{x}}_{\mathcal{F}}^{(i)}\right\}\left\{(\tilde{\mathbf{x}}_{\mathcal{F}^c}^{(i)})^{\mathrm{T}} - (\tilde{\mathbf{x}}_{\mathcal{F}}^{(i)})^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}})\right\}$$

and we have

Lemma B.1 *Suppose* $\mathrm{E}(\mathbf{x}) = \mathbf{0}$, $\mathrm{E}(\mathbf{y}) = \mathbf{0}$, *and* $\mathrm{E}(\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}})$ *is a linear function of* $\mathbf{x}_{\mathcal{F}}$. *If* $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$, *then*

$$\mathrm{E}_n(\mathbf{y}\hat{\boldsymbol{\gamma}}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{T}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\Pi}^{(i)} + O_p(n^{-1}),$$

where the first term on the right hand side is of order $O_p(n^{-1/2})$.

PROOF OF LEMMA B.1. From the definition of $\hat{\boldsymbol{\gamma}}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}$ and $\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}$, we have

$$\begin{aligned}\mathrm{E}_n(\mathbf{y}\hat{\boldsymbol{\gamma}}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{\mathrm{T}}) - \mathrm{E}(\mathbf{y}\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{\mathrm{T}}) &= \{\mathrm{E}_n(\mathbf{y}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}}) - \mathrm{E}(\mathbf{y}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}})\} \\ &\quad - [\mathrm{E}_n\{\mathbf{y}\mathbf{x}_{\mathcal{F}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}_n(\mathbf{x}_{\mathcal{F}}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}})\} - \mathrm{E}\{\mathbf{y}\mathbf{x}_{\mathcal{F}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}})\}].\end{aligned} \tag{B.35}$$

Because $\mathrm{E}(\mathbf{x}) = \mathbf{0}$ and $\mathrm{E}(\mathbf{y}) = \mathbf{0}$, it can be shown that

$$\mathrm{E}_n(\mathbf{y}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}}) - \mathrm{E}(\mathbf{y}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}}) = \frac{1}{n}\sum_{i=1}^n \{\tilde{\mathbf{y}}^{(i)}(\tilde{\mathbf{x}}_{\mathcal{F}^c}^{(i)})^{\mathrm{T}} - \mathrm{E}(\mathbf{y}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}})\} + O_p(n^{-1}). \tag{B.36}$$

The asymptotic expansions of $\widehat{\mathbf{\Sigma}}_{\mathbf{x}_\mathcal{F}}^{-1}$ and $\mathrm{E}_n(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}})$ are, respectively,

$$\widehat{\mathbf{\Sigma}}_{\mathbf{x}_\mathcal{F}}^{-1} - \mathbf{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1} = -\mathbf{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\tilde{\mathbf{x}}_\mathcal{F}^{(i)}(\tilde{\mathbf{x}}_\mathcal{F}^{(i)})^{\mathrm{T}} - \mathbf{\Sigma}_{\mathbf{x}_\mathcal{F}}\right\}\right]\mathbf{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1} + O_p(n^{-1}) \text{ and} \tag{B.37}$$

$$\mathrm{E}_n(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}}) - \mathrm{E}(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\tilde{\mathbf{x}}_\mathcal{F}^{(i)}(\tilde{\mathbf{x}}_{\mathcal{F}^c}^{(i)})^{\mathrm{T}} - \mathrm{E}(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}})\right\} + O_p(n^{-1}). \tag{B.38}$$

(B.37) and (B.38) together lead to

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_{\mathbf{x}_\mathcal{F}}^{-1}\mathrm{E}_n(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}}) - \mathbf{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\mathrm{E}(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\tilde{\mathbf{x}}_\mathcal{F}^{(i)}\Big\{&(\tilde{\mathbf{x}}_{\mathcal{F}^c}^{(i)})^{\mathrm{T}} - \\ &(\tilde{\mathbf{x}}_\mathcal{F}^{(i)})^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\mathrm{E}(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}})\Big\} + O_p(n^{-1}).\end{aligned} \tag{B.39}$$

It follows from (B.39) that

$$\begin{aligned}&\mathrm{E}_n\{\mathbf{y}\mathbf{x}_\mathcal{F}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_{\mathbf{x}_\mathcal{F}}^{-1}\mathrm{E}_n(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}})\} - \mathrm{E}\{\mathbf{y}\mathbf{x}_\mathcal{F}^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\mathrm{E}(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}})\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(\mathbf{y}\mathbf{x}_\mathcal{F}^{\mathrm{T}})\mathbf{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\tilde{\mathbf{x}}_\mathcal{F}^{(i)}\left\{(\tilde{\mathbf{x}}_{\mathcal{F}^c}^{(i)})^{\mathrm{T}} - (\tilde{\mathbf{x}}_\mathcal{F}^{(i)})^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\mathrm{E}(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}})\right\} \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\{\tilde{\mathbf{y}}^{(i)}(\tilde{\mathbf{x}}_\mathcal{F}^{(i)})^{\mathrm{T}} - \mathrm{E}(\mathbf{y}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}})\}\mathbf{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\mathrm{E}(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}}) + O_p(n^{-1}).\end{aligned} \tag{B.40}$$

(B.35), (B.36) and (B.40) together lead to

$$\begin{aligned}&\mathrm{E}_n(\mathbf{y}\hat{\boldsymbol{\gamma}}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_\mathcal{F}}^{\mathrm{T}}) = \mathrm{E}_n(\mathbf{y}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}}) - \mathrm{E}_n\{\mathbf{y}\mathbf{x}_\mathcal{F}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_{\mathbf{x}_\mathcal{F}}^{-1}\mathrm{E}_n(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}})\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big[\tilde{\mathbf{y}}^{(i)}(\tilde{\mathbf{x}}_{\mathcal{F}^c}^{(i)})^{\mathrm{T}} - \tilde{\mathbf{y}}^{(i)}(\tilde{\mathbf{x}}_\mathcal{F}^{(i)})^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\mathrm{E}(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}}) \\ &\quad -\mathrm{E}(\mathbf{y}\mathbf{x}_\mathcal{F}^{\mathrm{T}})\mathbf{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\tilde{\mathbf{x}}_\mathcal{F}^{(i)}\left\{(\tilde{\mathbf{x}}_{\mathcal{F}^c}^{(i)})^{\mathrm{T}} - (\tilde{\mathbf{x}}_\mathcal{F}^{(i)})^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\mathrm{E}(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}})\right\}\Big] + O_p(n^{-1}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{\tilde{\mathbf{y}}^{(i)} - \mathrm{E}(\mathbf{y}\mathbf{x}_\mathcal{F}^{\mathrm{T}})\mathbf{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\tilde{\mathbf{x}}_\mathcal{F}^{(i)}\right\}\left\{(\tilde{\mathbf{x}}_{\mathcal{F}^c}^{(i)})^{\mathrm{T}} - (\tilde{\mathbf{x}}_\mathcal{F}^{(i)})^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{x}_\mathcal{F}}^{-1}\mathrm{E}(\mathbf{x}_\mathcal{F}\mathbf{x}_{\mathcal{F}^c}^{\mathrm{T}})\right\} + O_p(n^{-1}),\end{aligned}$$

which is the desired result. □

Similarly, let $\mathrm{E}_n(\mathbf{y}_{\mathcal{G}}\mathbf{y}_{\mathcal{G}^c}^{\mathrm{T}}) = n^{-1}\sum_{i=1}^n \tilde{\mathbf{y}}_{\mathcal{G}}^{(i)}\left(\tilde{\mathbf{y}}_{\mathcal{G}^c}^{(i)}\right)^{\mathrm{T}}$, $\hat{\boldsymbol{\gamma}}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{(i)} = \tilde{\mathbf{y}}_{\mathcal{G}^c}^{(i)} - \mathrm{E}_n(\mathbf{y}_{\mathcal{G}^c}\mathbf{y}_{\mathcal{G}}^{\mathrm{T}})\widehat{\boldsymbol{\Sigma}}_{\mathbf{y}_{\mathcal{G}}}^{-1}\tilde{\mathbf{y}}_{\mathcal{G}}^{(i)}$, and $\mathrm{E}_n(\mathbf{x}_{\mathcal{F}}\hat{\boldsymbol{\gamma}}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{\mathrm{T}}) = n^{-1}\sum_{i=1}^n \tilde{\mathbf{x}}_{\mathcal{F}}^{(i)}\left(\hat{\boldsymbol{\gamma}}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{(i)}\right)^{\mathrm{T}}$. Define

$$\boldsymbol{\Lambda}^{(i)} = \left\{\tilde{\mathbf{x}}_{\mathcal{F}}^{(i)} - \mathrm{E}(\mathbf{x}_{\mathcal{F}}\mathbf{y}_{\mathcal{G}}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}}}^{-1}\tilde{\mathbf{y}}_{\mathcal{G}}^{(i)}\right\}\left\{(\tilde{\mathbf{y}}_{\mathcal{G}^c}^{(i)})^{\mathrm{T}} - (\tilde{\mathbf{y}}_{\mathcal{G}}^{(i)})^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}}}^{-1}\mathrm{E}(\mathbf{y}_{\mathcal{G}}\mathbf{y}_{\mathcal{G}^c}^{\mathrm{T}})\right\}$$

and we have

Lemma B.2 *Suppose* $\mathrm{E}(\mathbf{x}) = \mathbf{0}$, $\mathrm{E}(\mathbf{y}) = \mathbf{0}$, *and* $\mathrm{E}(\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}})$ *is a linear function of* $\mathbf{y}_{\mathcal{G}}$. *If* $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}$, *then*

$$\mathrm{E}_n(\mathbf{x}_{\mathcal{F}}\hat{\boldsymbol{\gamma}}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{T}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\Lambda}^{(i)} + O_p(n^{-1}),$$

where the first term on the right hand side is of order $O_p(n^{-1/2})$.

PROOF OF LEMMA B.2. The proof is similar to the proof of Lemma B.1, and is thus omitted. □

PROOF OF THEOREM 2.3.1. Recall that $|\mathcal{F}| = p_1$, $|\mathcal{F}^c| = p_2$, $|\mathcal{G}| = q_1$, $|\mathcal{G}^c| = q_2$, $p_1 + p_2 = p$ and $q_1 + q_2 = q$. Let $\boldsymbol{\phi}_1 = \{\boldsymbol{\Sigma}_{\mathbf{y}}^{-1/2}\mathrm{E}(\mathbf{y}\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{-1/2}\} \in \mathbb{R}^{qp_2}$, $\boldsymbol{\phi}_2 = \{\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1/2}\mathrm{E}(\mathbf{x}_{\mathcal{F}}\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{\mathrm{T}})\boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{-1/2}\} \in \mathbb{R}^{p_1q_2}$, and $\boldsymbol{\psi} = (\boldsymbol{\phi}_1^{\mathrm{T}}, \boldsymbol{\phi}_2^{\mathrm{T}})^{\mathrm{T}}$, where $(\cdot)$ denotes vectorization. Then we have $\delta_{-\mathcal{F}}^{-\mathcal{G}} = \boldsymbol{\psi}^{\mathrm{T}}\boldsymbol{\psi}$. At the sample level, let $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\phi}}_1^{\mathrm{T}}, \hat{\boldsymbol{\phi}}_2^{\mathrm{T}})^{\mathrm{T}}$, where $\hat{\boldsymbol{\phi}}_1 = \{\widehat{\boldsymbol{\Sigma}}_{\mathbf{y}}^{-1/2}\mathrm{E}_n(\mathbf{y}\hat{\boldsymbol{\gamma}}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{\mathrm{T}})\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{-1/2}\}$ and $\hat{\boldsymbol{\phi}}_2 = \{\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}_{\mathcal{F}}}^{-1/2}\mathrm{E}_n(\mathbf{x}_{\mathcal{F}}\hat{\boldsymbol{\gamma}}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{\mathrm{T}})\widehat{\boldsymbol{\Sigma}}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{-1/2}\}$. Then we have

$$\hat{\delta}_{-\mathcal{F}}^{-\mathcal{G}} = \hat{\boldsymbol{\psi}}^{\mathrm{T}}\hat{\boldsymbol{\psi}}. \tag{B.41}$$

Under $H_0^{\mathcal{F},[\mathcal{G}]}$, we have $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$. It follows that $\mathrm{E}(\mathbf{y}\boldsymbol{\gamma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{\mathrm{T}}) = \mathbf{0}$ and $\boldsymbol{\phi}_1 = \mathbf{0}$. Together with Lemma B.1, we have

$$\hat{\boldsymbol{\phi}}_1 = \frac{1}{n}\sum_{i=1}^{n} \{\boldsymbol{\Sigma}_{\mathbf{y}}^{-1/2}\boldsymbol{\Pi}^{(i)}\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{-1/2}\} + O_p(n^{-1}), \tag{B.42}$$

where the first term on the right hand side is of order $O_p(n^{-1/2})$. Similarly, we have $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}$ under $H_0^{\mathcal{F},[\mathcal{G}]}$. It follows that $\mathrm{E}(\mathbf{x}_{\mathcal{F}}\boldsymbol{\gamma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{\mathrm{T}}) = \mathbf{0}$ and $\boldsymbol{\phi}_2 = \mathbf{0}$. Together with Lemma B.2, we have

$$\hat{\boldsymbol{\phi}}_2 = \frac{1}{n}\sum_{i=1}^{n} \{\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1/2}\boldsymbol{\Lambda}^{(i)}\boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{-1/2}\} + O_p(n^{-1}), \tag{B.43}$$

where the first term on the right hand side is of order $O_p(n^{-1/2})$. It follows from (B.42) and (B.43) that

$$\hat{\boldsymbol{\psi}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\vartheta}^{(i)} + O_p(n^{-1}), \tag{B.44}$$

where $\boldsymbol{\vartheta}^{(i)} = \left\{{}^{\mathrm{T}}(\boldsymbol{\Sigma}_{\mathbf{y}}^{-1/2}\boldsymbol{\Pi}^{(i)}\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{-1/2}), {}^{\mathrm{T}}(\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}}}^{-1/2}\boldsymbol{\Lambda}^{(i)}\boldsymbol{\Sigma}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}^{-1/2})\right\}^{\mathrm{T}} \in \mathbb{R}^L$ with $\mathrm{E}(\boldsymbol{\vartheta}^{(i)}) = \mathbf{0}$ and $L = qp_2 + p_1q_2 = pq - p_1q_1$. As a result of (B.44), we have

$$\sqrt{n}\hat{\boldsymbol{\psi}} \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Omega}) \tag{B.45}$$

as $n \to \infty$, where $\boldsymbol{\Omega} = \mathrm{E}\{\boldsymbol{\vartheta}^{(i)}(\boldsymbol{\vartheta}^{(i)})^{\mathrm{T}}\}$. (B.45) and (B.41) lead to the desired result. □

As a special case, $\delta_{-\mathcal{F}}^{-\mathcal{G}}$ reduces to $\delta_{-\mathcal{F}}$ when we set $\mathcal{G} = \mathcal{I}_{\mathbf{y}}$. Then $\delta_{-\mathcal{F}} = \boldsymbol{\phi}_1^{\mathrm{T}}\boldsymbol{\phi}_1$ and $\hat{\delta}_{-\mathcal{F}} = \hat{\boldsymbol{\phi}}_1^{\mathrm{T}}\hat{\boldsymbol{\phi}}_1$. It follows from (B.42) that $\hat{\boldsymbol{\phi}}_1 = n^{-1}\sum_{i=1}^{n}\boldsymbol{\vartheta}_1^{(i)} + O_p(n^{-1})$, where $\boldsymbol{\vartheta}_1^{(i)} = \{\boldsymbol{\Sigma}_{\mathbf{y}}^{-1/2}\boldsymbol{\Pi}^{(i)}\boldsymbol{\Sigma}_{\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}}}^{-1/2}\} \in \mathbb{R}^{p_2q}$. Thus $\sqrt{n}\hat{\boldsymbol{\phi}}_1 \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Gamma})$, where $\boldsymbol{\Gamma} = \mathrm{E}\{\boldsymbol{\vartheta}_1^{(i)}(\boldsymbol{\vartheta}_1^{(i)})^{\mathrm{T}}\}$. The asymptotic distribution of $\hat{\delta}_{-\mathcal{F}}$ is summarized in the next result.

Corollary B.1 *Suppose* $\mathrm{E}(\mathbf{x}) = \mathbf{0}$*,* $\mathrm{E}(\mathbf{y}) = \mathbf{0}$*, and* $\mathrm{E}(\mathbf{x}_{\mathcal{F}^c}|\mathbf{x}_{\mathcal{F}})$ *is a linear function of* $\mathbf{x}_{\mathcal{F}}$*. Then under* $H_0^{\mathcal{F}} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{F}}$*,*

$$n\hat{\delta}_{-\mathcal{F}} \xrightarrow{D} \sum_{\ell=1}^{p_2 q} \rho_\ell \chi^2_\ell(1)$$

as $n \to \infty$*, where "*$\xrightarrow{D}$*" means convergence in distribution,* $\chi^2_\ell(1)$ *are independent chi-square with one degree of freedom for* $\ell = 1, \ldots, p_2 q$*, and* $\rho_1 \geq \ldots \geq \rho_{p_2 q}$ *are the eigenvalues of* $\mathbf{\Gamma}$*.*

Similarly, $\delta^{-\mathcal{G}}_{-\mathcal{F}}$ becomes $\delta^{-\mathcal{G}}$ when we set $\mathcal{F} = \mathcal{I}_{\mathbf{x}}$. Note that $\delta^{-\mathcal{G}} = \boldsymbol{\phi}_3^{\mathrm{T}} \boldsymbol{\phi}_3$ with $\boldsymbol{\phi}_3 = \{\mathbf{\Sigma}_{\mathbf{x}}^{-1/2} \mathrm{E}(\mathbf{x} \boldsymbol{\gamma}^{\mathrm{T}}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}) \mathbf{\Sigma}^{-1/2}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}\} \in \mathbb{R}^{pq_2}$, and $\hat{\delta}^{-\mathcal{G}} = \hat{\boldsymbol{\phi}}_3^{\mathrm{T}} \hat{\boldsymbol{\phi}}_3$ with $\hat{\boldsymbol{\phi}}_3 = \{\widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1/2} \mathrm{E}_n(\mathbf{x} \hat{\boldsymbol{\gamma}}^{\mathrm{T}}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}) \widehat{\mathbf{\Sigma}}^{-1/2}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}\}$. Similar to (B.43), it can be shown that $\hat{\boldsymbol{\phi}}_3 = n^{-1} \sum_{i=1}^n \boldsymbol{\vartheta}_3^{(i)} + O_p(n^{-1})$, where $\boldsymbol{\vartheta}_3^{(i)} = \{\mathbf{\Sigma}_{\mathbf{x}}^{-1/2} \mathbf{\Lambda}^{(i)} \mathbf{\Sigma}^{-1/2}_{\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}}}\}$. Thus $\sqrt{n}\hat{\boldsymbol{\phi}}_3 \xrightarrow{D} N(\mathbf{0}, \mathbf{\Upsilon})$, where $\mathbf{\Upsilon} = \mathrm{E}\{\boldsymbol{\vartheta}_3^{(i)} (\boldsymbol{\vartheta}_3^{(i)})^{\mathrm{T}}\}$. The asymptotic distribution of $\hat{\delta}^{-\mathcal{G}}$ is summarized in the next result.

Corollary B.2 *Suppose* $\mathrm{E}(\mathbf{x}) = \mathbf{0}$*,* $\mathrm{E}(\mathbf{y}) = \mathbf{0}$*, and* $\mathrm{E}(\mathbf{y}_{\mathcal{G}^c}|\mathbf{y}_{\mathcal{G}})$ *is a linear function of* $\mathbf{y}_{\mathcal{G}}$*. Then under* $H_0^{[\mathcal{G}]} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{y}_{\mathcal{G}}$*,*

$$n\hat{\delta}^{-\mathcal{G}} \xrightarrow{D} \sum_{\ell=1}^{pq_2} \omega_\ell \chi^2_\ell(1)$$

as $n \to \infty$*, where "*$\xrightarrow{D}$*" means convergence in distribution,* $\chi^2_\ell(1)$ *are independent chi-square with one degree of freedom for* $\ell = 1, \ldots, pq_2$*, and* $\omega_1 \geq \ldots \geq \omega_{pq_2}$ *are the eigenvalues of* $\mathbf{\Upsilon}$*.*

DERIVATION OF $\mathcal{B}$ FOR MODEL 1. Let

$$\mathbf{\Sigma}_{\boldsymbol{\epsilon}} = \begin{bmatrix} 1 & .5 & .25 & .125 \\ .5 & 1 & .5 & .25 \\ .25 & .5 & 1 & .5 \\ .125 & .25 & .5 & 1 \end{bmatrix} \text{ and } \mathbf{H} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

Because $\mathbf{y} = \mathbf{H}\mathbf{x} + \boldsymbol{\epsilon}$, we have

$$\boldsymbol{\Sigma}_{\mathbf{y}} = \mathbf{H}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{H}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \begin{bmatrix} 1 & .5 & .25 & .125 \\ .5 & 1 & .5 & .25 \\ .25 & .5 & 1 & .5 \\ .125 & .25 & .5 & 3 \end{bmatrix} \text{ and } \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} = \begin{bmatrix} \frac{4}{3} & -\frac{2}{3} & 0 & 0 \\ -\frac{2}{3} & \frac{5}{3} & -\frac{2}{3} & 0 \\ 0 & -\frac{2}{3} & \frac{47}{33} & -\frac{2}{11} \\ 0 & 0 & -\frac{2}{11} & \frac{4}{11} \end{bmatrix}.$$

It follows that

$$\mathrm{E}(\mathbf{x}|\mathbf{y}) = \boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{y} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{4}{3} & -\frac{2}{3} & 0 & 0 \\ -\frac{2}{3} & \frac{5}{3} & -\frac{2}{3} & 0 \\ 0 & -\frac{2}{3} & \frac{47}{33} & -\frac{2}{11} \\ 0 & 0 & -\frac{2}{11} & \frac{4}{11} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} = \frac{2}{11} \begin{bmatrix} 2Y_4 - Y_3 \\ 2Y_4 - Y_3 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Thus we have $\mathcal{B} = \{3, 4\}$ for Model 1. □

APPENDIX C

PROOFS FOR PROPOSITIONS AND THEOREMS IN CHAPTER 3

PROOF OF PROPOSITION 3.1.1. Because $t_k^{\mathbf{D}} \geq 0$, statement (i) and statement (ii) are equivalent. It suffices to prove statement (i). By Stein's Lemma, the normality of $\mathbf{x}$ guarantees that

$$E\left\{\frac{\partial E(Y_j|\mathbf{x})}{\partial \mathbf{x}}\right\} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} E\{E(Y_j|\mathbf{x})\mathbf{x}\} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} E(\mathbf{x}Y_j) = \boldsymbol{\beta}_j.$$

It follows that

$$E\{\partial E(Y_j|\mathbf{x})/\partial X_k\} = \beta_{jk}. \tag{C.46}$$

For the "only if" part, under $H_0^{-k} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{-k}$ and model (3.19), we have $\partial E(Y_j|\mathbf{x})/\partial X_k = 0$, $j = 1, \ldots, q$. This fact together with (C.46) and the definition of $\boldsymbol{\lambda}_k$ leads to $\boldsymbol{\lambda}_k = (\beta_{1k}, \ldots, \beta_{qk})^{\mathrm{T}} = \mathbf{0}$. As a result, $t_k^{\mathbf{D}} = \boldsymbol{\lambda}_k^{\mathrm{T}} \mathbf{D} \boldsymbol{\lambda}_k = 0$.

For the "if" part, assumption (c) and (C.46) together lead to

$$\boldsymbol{\lambda}_k = (\beta_{1k}, \ldots, \beta_{qk})^{\mathrm{T}} \neq \mathbf{0} \text{ under } \mathbf{H_a^{-k}} : \mathbf{y} \not\perp\!\!\!\perp \mathbf{x} \mid \mathbf{x_{-k}}, \tag{C.47}$$

On the other hand, $t_k^{\mathbf{D}} = 0$ implies $\boldsymbol{\lambda}_k = \mathbf{0}$ due to the positive definiteness of $\mathbf{D}$. It follows from (C.47) that $H_0^{-k} : \mathbf{y} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{-k}$ must be true due to the contradiction otherwise. □

PROOF OF PROPOSITION 3.1.2. First note that $\mathbf{\Sigma}_{\tilde{\mathbf{y}}} = \mathbf{\Gamma}\mathbf{\Sigma}_{\mathbf{y}}\mathbf{\Gamma}^{\mathrm{T}}$ and $\mathbf{\Sigma}_{\tilde{\mathbf{y}}}^{-1} = (\mathbf{\Gamma}^{\mathrm{T}})^{-1}\mathbf{\Sigma}_{\mathbf{y}}^{-1}\mathbf{\Gamma}^{-1}$. We also have $E(\mathbf{x}\tilde{\mathbf{y}}^{\mathrm{T}}) = E(\mathbf{x}\mathbf{y}^{\mathrm{T}})\mathbf{\Gamma}^{\mathrm{T}}$ as $E(\mathbf{x}) = \mathbf{0}$. It follows that

$$\tilde{t}_k = \mathbf{e}_k^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{x}}^{-1}E(\mathbf{x}\mathbf{y}^{\mathrm{T}})\mathbf{\Gamma}^{\mathrm{T}}(\mathbf{\Gamma}^{\mathrm{T}})^{-1}\mathbf{\Sigma}_{\mathbf{y}}^{-1}\mathbf{\Gamma}^{-1}\mathbf{\Gamma}E(\mathbf{y}\mathbf{x}^{\mathrm{T}})\mathbf{\Sigma}_{\mathbf{x}}^{-1}\mathbf{e}_k = t_k.$$

□

PROOF OF THEOREM 3.1.1. The asymptotic expansion of $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1}$ is

$$\widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1} - \mathbf{\Sigma}_{\mathbf{x}}^{-1} = -\mathbf{\Sigma}_{\mathbf{x}}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}(\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^{\mathrm{T}} - \mathbf{\Sigma}_{\mathbf{x}})\right)\mathbf{\Sigma}_{\mathbf{x}}^{-1} + O_p(n^{-1}). \tag{C.48}$$

Denote $\mathbf{\Sigma}_{\mathbf{xy}} = E(\mathbf{x}\mathbf{y}^{\mathrm{T}})$. Then the asymptotic expansion of $\widehat{\mathbf{\Sigma}}_{\mathbf{xy}}$ is

$$\widehat{\mathbf{\Sigma}}_{\mathbf{xy}} - \mathbf{\Sigma}_{\mathbf{xy}} = \frac{1}{n}\sum_{i=1}^{n}(\widetilde{\mathbf{x}}_i\mathbf{y}_i^{\mathrm{T}} - \mathbf{\Sigma}_{\mathbf{xy}}) + O_p(n^{-1}). \tag{C.49}$$

For $\mathbf{B} = \mathbf{\Sigma}_{\mathbf{x}}^{-1}E(\mathbf{x}\mathbf{y}^{\mathrm{T}})$, denote its estimator as $\hat{\mathbf{B}} = \widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1}\widehat{\mathbf{\Sigma}}_{\mathbf{xy}}$. Let $\mathbf{B}_i^* = -\mathbf{\Sigma}_{\mathbf{x}}^{-1}(\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^{\mathrm{T}} - \mathbf{\Sigma}_{\mathbf{x}})\mathbf{\Sigma}_{\mathbf{x}}^{-1}\mathbf{\Sigma}_{\mathbf{xy}} + \mathbf{\Sigma}_{\mathbf{x}}^{-1}(\widetilde{\mathbf{x}}_i\mathbf{y}_i^{\mathrm{T}} - \mathbf{\Sigma}_{\mathbf{xy}}) = \mathbf{\Sigma}_{\mathbf{x}}^{-1}\widetilde{\mathbf{x}}_i(\mathbf{y}_i - \mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{x}}^{-1}\widetilde{\mathbf{x}}_i)^{\mathrm{T}}$, where $\mathbf{\Sigma}_{\mathbf{yx}} = E(\mathbf{y}\mathbf{x}^{\mathrm{T}})$. From (C.48) and (C.49), we have

$$\hat{\mathbf{B}} - \mathbf{B} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{B}_i^* + O_p(n^{-1}). \tag{C.50}$$

It follows from (C.50) that

$$\mathbf{D}^{1/2}(\hat{\mathbf{B}} - \mathbf{B})^{\mathrm{T}}\mathbf{e}_k = \frac{1}{n}\sum_{i=1}^{n}\mathbf{D}^{1/2}(\mathbf{B}_i^*)^{\mathrm{T}}\mathbf{e}_k + O_p(n^{-1}). \tag{C.51}$$

By the central limit theorem, (C.51) leads to

$$\sqrt{n}\mathbf{D}^{1/2}(\hat{\mathbf{B}} - \mathbf{B})^{\mathrm{T}}\mathbf{e}_k \xrightarrow{L} N(\mathbf{0}, \mathbf{\Omega}_{\mathbf{k}}^{\mathbf{D}}), \tag{C.52}$$

where $\mathbf{\Omega}_k^{\mathbf{D}} = E\left\{\mathbf{D}^{1/2}(\mathbf{B}_i^*)^{\mathrm{T}}\mathbf{e}_k\mathbf{e}_k^{\mathrm{T}}\mathbf{B}_i^*\mathbf{D}^{1/2}\right\}$. It is implied by (C.52) that

$$n\mathbf{e}_k^{\mathrm{T}}(\hat{\mathbf{B}}-\mathbf{B})\mathbf{D}(\hat{\mathbf{B}}-\mathbf{B})^{\mathrm{T}}\mathbf{e}_k \xrightarrow{L} \sum_{j=1}^{q} v_{k,j}^{\mathbf{D}}\chi^2(1), \tag{C.53}$$

where $v_{k,j}^{\mathbf{D}}$ is the jth eigenvalue of $\mathbf{\Omega}_k^{\mathbf{D}}$, $j = 1, \ldots, q$. From the proof of the "only if" part of Proposition 3.1.1, we have shown that with assumptions (a) and (b),

$$\mathbf{e}_k^{\mathrm{T}}\mathbf{B} = \boldsymbol{\lambda}_k^{\mathrm{T}} = \mathbf{0} \text{ under } \mathbf{H_0^{-k}}. \tag{C.54}$$

(C.53), (C.54), and the definition of $\hat{t}_k^{\mathbf{D}}$ in (3.23) together lead to the desired result. □

PROOF OF COROLLARY 3.1.1. Let $\mathbf{\Omega}_k = E\left\{\mathbf{\Sigma}_{\mathbf{y}}^{-1/2}(\mathbf{B}_i^*)^{\mathrm{T}}\mathbf{e}_k\mathbf{e}_k^{\mathrm{T}}\mathbf{B}_i^*\mathbf{\Sigma}_{\mathbf{y}}^{-1/2}\right\}$. It follows from (C.53) in the proof of Theorem 3.1.1 that

$$n\mathbf{e}_k^{\mathrm{T}}(\hat{\mathbf{B}}-\mathbf{B})\mathbf{\Sigma}_{\mathbf{y}}^{-1}(\hat{\mathbf{B}}-\mathbf{B})^{\mathrm{T}}\mathbf{e}_k \xrightarrow{L} \sum_{j=1}^{q} v_{k,j}\chi^2(1), \tag{C.55}$$

where $v_{k,j}$ is the jth eigenvalue of $\mathbf{\Omega}_k$. On the other hand, because $\hat{\mathbf{B}} - \mathbf{B} = O_p(n^{-1/2})$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{y}}^{-1} - \mathbf{\Sigma}_{\mathbf{y}}^{-1} = O_p(n^{-1/2})$, we have

$$\begin{aligned}(\hat{\mathbf{B}}-\mathbf{B})\widehat{\mathbf{\Sigma}}_{\mathbf{y}}^{-1}(\hat{\mathbf{B}}-\mathbf{B})^{\mathrm{T}} &= (\hat{\mathbf{B}}-\mathbf{B})\mathbf{\Sigma}_{\mathbf{y}}^{-1}(\hat{\mathbf{B}}-\mathbf{B})^{\mathrm{T}} + (\hat{\mathbf{B}}-\mathbf{B})(\widehat{\mathbf{\Sigma}}_{\mathbf{y}}^{-1} - \mathbf{\Sigma}_{\mathbf{y}}^{-1})(\hat{\mathbf{B}}-\mathbf{B})^{\mathrm{T}} \\ &= (\hat{\mathbf{B}}-\mathbf{B})\mathbf{\Sigma}_{\mathbf{y}}^{-1}(\hat{\mathbf{B}}-\mathbf{B})^{\mathrm{T}} + o_p(n^{-1}),\end{aligned} \tag{C.56}$$

where the leading term in (C.56) is $O_p(n^{-1})$. (C.54), (C.55), (C.56), and the definition of $\hat{t}_k$ in (3.24) together lead to the desired result. □