DIMENSION REDUCTION WITH INVERSE REGRESSION: A MINIMUM DISCREPANCY APPROACH

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

LIQIANG NI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

R. DENNIS COOK, Adviser

June 2003

UNIVERSITY OF MINNESOTA

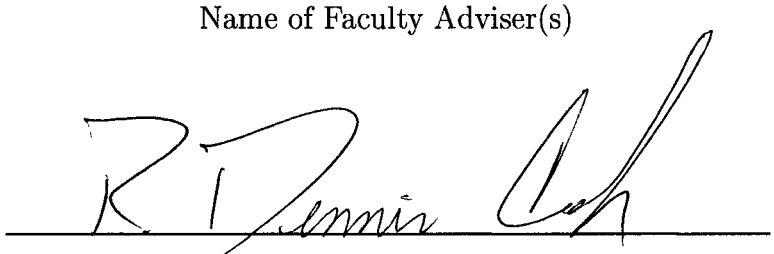
This is to certify that I have examined this copy of a doctoral thesis by

Liqiang Ni

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

R. DENNIS COOK

Name of Faculty Adviser(s)



Signature of Faculty Adviser(s)

June 23, 2003

Date

GRADUATE SCHOOL

Abstract

The goal of this thesis is to develop sufficient dimension reduction methods in regressions via a minimum discrepancy approach. The thesis includes two main parts.

Part I examines regressions for a single population. A family of dimension reduction methods is developed by minimizing a quadratic objective function. An optimal member of this family called *optimal inverse regression estimation* (Optimal IRE) is proposed, along with inference methods and a computational algorithm. Optimal IRE is optimal in two respects: Its estimated basis of the central dimension reduction subspace (Cook 1994) is asymptotically efficient and its test statistic for dimension has an asymptotic chi-squared distribution. Current methods like *sliced inverse regression* (SIR; Li 1991) and the *weighted chi-squared test* (WCT; Bura and Cook 2001b) belong to a sub-optimal class of this family. Another member of this sub-optimal class–*simple inverse regression estimation* (Simple IRE) proposed in this thesis—often performs better than SIR and WCT. Comparison of these methods is reported through simulation studies.

Part II focuses on sufficient partial dimension reduction in regression across multiple subpopulations. We rederive and extend *partial sliced inverse regression* (partial SIR; Chiaromonte, Cook and Li 2002) by the minimum discrepancy approach. A new method, *general partial SIR*, is proposed in this thesis, which removes the restriction in partial SIR that the predictor covariances matrices are constant across subpopulations. This extension significantly expands the applicability of dimension reduction methodologies.

Acknowledgments

I would like to thank my advisor, Professor R. Dennis Cook, for his constant help, clear guidance, and enormous patience during my graduate studies in University of Minnesota. I also thank Professors Douglas Hawkins, Glen Meeden, and Christopher Nachtsheim on my examination committee.

I dedicate this thesis to my parents for their endless love and support. Thanks to my wife, Rong Yang. Without her, life is just a random walk.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Data are indispensable in modern life. Important information is hiding in the data. With the development of technology, more measurements can be collected with higher accuracy than ever before. As information technology flourishes, people are overwhelmed by huge amounts of data. Meanwhile even with so much data, when we put them in a high dimensional space, they are still isolated and sparse. This makes many techniques that work fairly well in lower dimensions lose their edge in high dimensions. Even with unlimited computing power, we still have to address the central issue: how to get the most information from the data without letting details obstruct our eyes. Help is available from dimension reduction techniques that summarize the data in a much lower dimensional space while preserve as much information as possible. Dimension reduction makes visualization of the data become possible. A statistical model in the lower dimensional space is often far more parsimonious.

While people agree on its importance, they have different understandings about dimension reduction. Many dimension reduction methods have been developed. Principal components analysis, factor analysis, independent components analysis, projection pursuit, Fourier transformation, and wavelets are among a long list of dimension reduction tools. These methods try to find the "most important" features or patterns in the data. In this thesis, we employ the distinct but related notion of *sufficient dimension reduction* in regression.

Consider a typical regression of a response Y on a vector $\mathbf{X}$ of p predictors. Generally the object of interest is the conditional distribution of $Y|\mathbf{X}$, which in many cases is identical to the conditional distribution of Y given $q < p$ linear combinations of $\mathbf{X}$, $\boldsymbol{\eta}^T\mathbf{X} = (\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_q)^T\mathbf{X}$, where $\boldsymbol{\eta}_i \in \mathbb{R}^p$. Letting $\mathbf{P_A}$ be the orthogonal projection operator onto the space spanned by the columns of a matrix $\mathbf{A}$, we can restrict our attention to a lower dimensional projection $\mathbf{P}_{\boldsymbol{\eta}}\mathbf{X}$ without loss of information on $Y|\mathbf{X}$. We call this *sufficient dimension reduction.* Cook (1994) formulated *sufficient dimension reduction* via the concept of the *central dimension reduction subspace* (CS), which is the intersection of all spaces $\mathcal{S}$ such that

$$Y \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_{\mathcal{S}}\mathbf{X},$$

where $\perp\!\!\!\perp$ represents independence. Let $\mathcal{S}_{Y|\mathbf{X}}$ denote the central subspace. It is easy to see that a linear transformation of $\mathbf{X}$ leads to a linear transformation of $\mathcal{S}_{Y|\mathbf{X}}$. Suppose $\mathcal{S}_{Y|\mathbf{X}}$ is d-dimensional with a basis $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)$. We call these $\boldsymbol{\beta}_i^T\mathbf{X}$ *sufficient predictors.* According to the above definition, given $\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X}$, $\mathbf{X}$ is independent of Y, which means that $\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X}$ carries

all the information about Y that is available from $\mathbf{X}$. Another unique feature of sufficient dimension reduction is that it is model-free. We do not assume any particular parametric or nonparametric model for $Y|\mathbf{X}$ when we are considering dimension reduction. Naturally, when modeling is necessary, we may construct models based on the sufficient predictors after we find the CS. Model construction is not within the scope of this thesis.

1.1 Central Dimension Reduction Subspace

Methods for estimation of the central subspace include *sliced inverse regression* (SIR; Li 1991), *sliced average variance estimation* (SAVE; Cook and Weisberg 1991), graphical regression (Cook 1994, 1998), and parametric inverse regression (Bura and Cook 2001a). Among them, SIR is perhaps the most widely used method. Let $\mathbf{Z} \equiv \mathrm{Cov}(\mathbf{X})^{-\frac{1}{2}}(\mathbf{X} - \mathrm{E}[\mathbf{X}])$ denote the standardized predictor. Under mild conditions on the marginal distribution of the predictor vector, $\mathrm{Span}(\mathrm{Cov}(\mathrm{E}[\mathbf{Z}|Y])) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$. SIR uses a sample version of $\mathrm{Cov}(\mathrm{E}[\mathbf{Z}|Y])$ as a kernel matrix, where columns are intended to span $\mathcal{S}_{Y|\mathbf{Z}}$ in the population. Then, SIR produces a spectral decomposition of the kernel matrix, where the sum of the smallest eigenvalues is utilized to construct test statistics for estimating the dimension of $\mathcal{S}_{Y|\mathbf{Z}}$. Suppose we decide $\dim(\mathcal{S}_{Y|\mathbf{Z}}) = d$. Then the d eigenvectors corresponding to the d largest eigenvalues constitute an estimate of a basis of $\mathcal{S}_{Y|\mathbf{Z}}$, since the space spanned by these eigenvectors is the subspace "closest" to the kernel matrix's columns. We call this estimation approach the *spectral decomposition approach.* Then

an estimate of $\mathcal{S}_{Y|\mathbf{X}}$ can be obtained by simple linear transformation. The mild conditions under which SIR works best are well-studied in the literature. Bura and Cook (2001b) proposed the *weighted chi-squared test* (WCT), which extended SIR for more general situations, while still using the same test statistics. We will discuss SIR and WCT in detail in later chapters.

A simple example using SIR

We consider a simple example using SIR. Let $\mathbf{X} = (X_1, X_2, \ldots, X_5)^T$ be a 5-dimensional multivariate normal random vector with zero mean and identity covariance matrix. Suppose the response Y is generated according to the model:

$$Y = \exp[-(X_1 + X_2 + X_3)] + \epsilon,$$

where ϵ is a standard normal variable that is independent of $\mathbf{X}$. In this case, the central subspace is 1-dimensional with $\boldsymbol{\beta}_1 = (1, 1, 1, 0, 0)^T$, since

$$Y \perp\!\!\!\perp \mathbf{X} | \boldsymbol{\beta}_1^T \mathbf{X}.$$

We generated 400 data points according to the model. SIR detected one sufficient predictor: The estimated sufficient predictor is

$$\hat{\boldsymbol{\beta}}_1^T \mathbf{X} = (0.57, 0.59, 0.57, -0.05, 0.02)^T \mathbf{X},$$

which has a sample correlation of 0.9987 with the true sufficient predictor $\boldsymbol{\beta}_1^T \mathbf{X}$. SIR did a good job in this example. A plot of Y versus the estimated sufficient predictor $\hat{\boldsymbol{\beta}}_1^T \mathbf{X}$ is shown in Figure 1.1.

In Chapter 5 we show that SIR's spectral decomposition approach is a special case of a minimum discrepancy approach, which is in the form of a

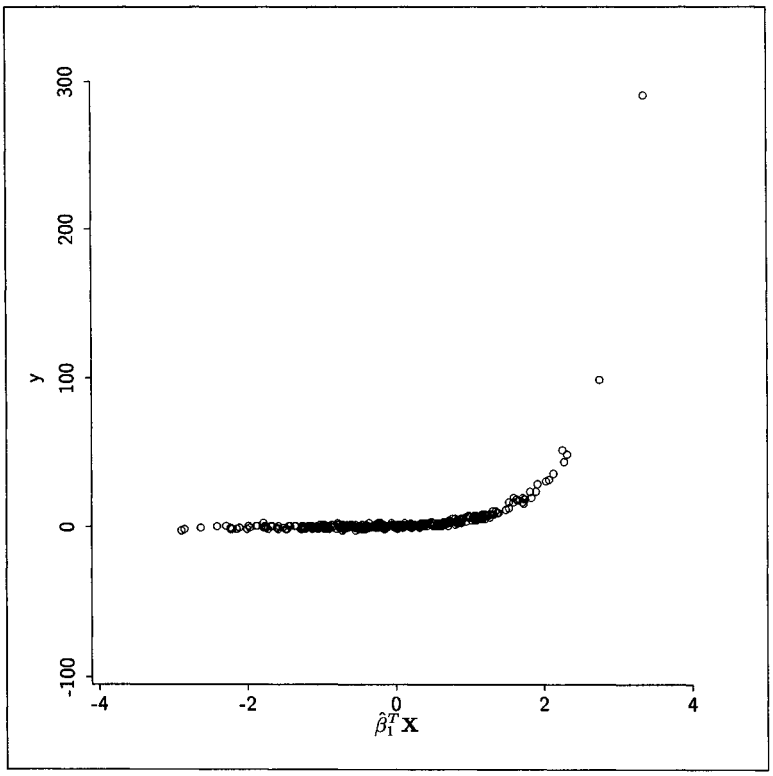


Figure 1.1: Summary plot from a simple example using SIR.

quadratic inference function. In this thesis, we develop a class of dimension reduction methods via the minimum discrepancy approach. We call it the MDA family. SIR belongs to this family. Is SIR the optimal member? The answer is no! We propose a new method, *optimal inverse regression estimation* (Optimal IRE), which is optimal in terms of asymptotic efficiency and its test statistic for dimension has an asymptotic chi-squared distribution. It turns out that SIR belongs to a sub-optimal class of the MDA family. Even within this sub-optimal class, another member *simple inverse regression estimation* (Simple IRE), which is also proposed in this thesis, often has better

performance than SIR.

1.2 Partial Central Dimension Reduction Subspace

Later we change our focus to the multiple subpopulation case. Suppose W is a random variable which indicates the subpopulation. SIR may do well with continuous or many-valued predictors, but it fails to deal with this situation. Chiaromonte, Cook, and Li (2002; herein after CCL) extended sufficient dimension reduction to regressions across multiple subpopulations. Parallel to the central dimension reduction subspace, they define the *partial central subspace* (PCS) as the intersection of all spaces $\mathcal{S}$ such that

$$Y \perp\!\!\!\perp \mathbf{X} | (\mathbf{P}_{\mathcal{S}}\mathbf{X}, W).$$

CCL proposed a methodology, *partial SIR*, to estimate the PCS. Let us look at one example to fix the idea.

Lean Body Mass Regression

We revisit one of the regressions discussed by CCL. For $n = 202$ athletes at the Australian Institute of Sport, consider the regression of lean body mass L on $p = 5$ continuous or many-values predictors, the logarithms of height, weight, red cell count, white cell count and hemoglobin, represented by $\mathbf{X}$ and gender indicated by $W = m$ or f. Partial SIR estimated the PCS as 1-dimensional. A plot of L versus the estimated sufficient predictor $\hat{\beta}^T\mathbf{X}$ is shown in Figure 1.2. The ordinary least squares fits are shown for males

and females as visual aids. The interpretation of the plot is that while males and females have different regressions they both depend on one and the same linear combination of the predictors $\mathbf{X}$.

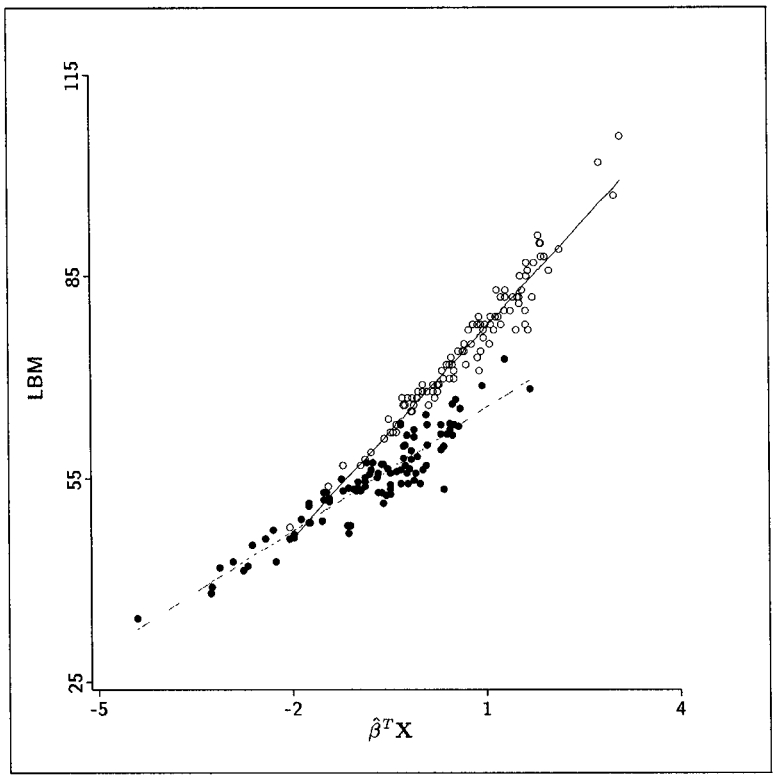


Figure 1.2: Summary plot from application of partial SIR to the lean body mass regression. ○ males. ● females.

However, partial SIR has an important limiting condition: the predictors must have the same covariance structure across subpopulations; that is, $\mathrm{Cov}(\mathbf{X}|W)$ must be constant. Since there is no big difference between the covariance matrices of $\mathbf{X}$ for males and females in the above regression, partial

SIR works fine. But usually we do not expect homogenous subpopulation covariances. Therefore, this restriction should not be neglected in practice. In this thesis, we develop dimension reduction methods for heterogenous subpopulations via the minimum discrepancy approach. We propose a new method, *general partial SIR* (GP.SIR), to estimate the partial central subspace. GP.SIR only requires the same conditions as SIR does. Therefore, we can use GP.SIR in far more situations than partial SIR, which expands substantially the application domain of sufficient dimension reduction. When all subpopulations share the same covariance matrix, partial SIR arises naturally as a special case. When there is only one population, GP.SIR reduces to SIR.

1.3 Outline of the Thesis

This thesis includes two main parts. Part I investigates sufficient dimension reduction methods in single population regression. Chapter 2 reviews sufficient dimension reduction via the concept of the central dimension reduction subspace, where dimension reduction and inverse regression are connected, setting the stage for developing methodologies by a minimum discrepancy approach. An MDA family of dimension reduction methods is proposed there. Chapter 3 develops an optimal method—Optimal IRE. Within this MDA family, Chapter 4 outlines a sub-optimal class that includes SIR, WCT, and Simple IRE, which are discussed in subsequent chapters. Comparisons between SIR, WCT, Simple IRE, and Optimal IRE are reported in Chapter 8.

In Part II we shift our focus to dimension reduction in regression across multiple subpopulations. Chapter 9 reviews the frame work of partial dimension reduction via the partial central dimension reduction subspace, which parallels the central subspace discussed in Chapter 2. Then, we propose a new method—general partial SIR, of which partial SIR is a special case. Inference about partial dimension reduction is addressed in Chapter 10. Partial SIR and general partial SIR are compared by simulations and a real example. At the end of the thesis, we briefly discuss a plan for future research.

Part I

Dimension Reduction for Regression in a Single Population

Chapter 2

Sufficient Dimension Reduction

In a regression of Y on $\mathbf{X} \in \mathbb{R}^p$, usually we consider the conditional distribution of $Y|\mathbf{X}$. Common practice is to model the relation between Y and $\mathbf{X}$ as

$$Y = f(\boldsymbol{\eta}_1^T\mathbf{X}, \ldots, \boldsymbol{\eta}_d^T\mathbf{X}, \epsilon) \tag{2.1}$$

where $\boldsymbol{\eta}_i \in \mathbb{R}^p$, and ϵ is an error term that is independent of $\mathbf{X}$. The function $f(\cdot)$ can be estimated either by a parametric model or nonparametric smoothing techniques. There are two overarching questions: one is how to estimate $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_d)$, the other is how to specify $f(\cdot)$. Often we solve the problem iteratively, by assuming we know the answer for one question and then estimating the answer for the other. For example, projection pursuit regression (Friedman and Stuetzle 1981) estimates the regression surface by a sum of univariate functions:

$$f(\mathbf{X}) = \sum_{i=1}^{d} f_i(\boldsymbol{\eta}_i^T\mathbf{X}),$$

where $f_i(\cdot)$'s can be empirically determined by assuming that they belong to a particular family, say, quadratic functions. Recursive partitioning regression depends on using the $\boldsymbol{\eta}_i^T\mathbf{X}$'s to split the predictor space and to estimate the function within each space. However, when $\mathbf{X}$ is high dimensional, we encounter the *curse of dimensionality.* The data are so sparse in any region of interest that it is very difficult to estimate $\boldsymbol{\eta}$, even with assumption of a function form. Meanwhile, the estimation process is so heavily data-driven that often some $\boldsymbol{\eta}_i$'s found are mainly artifacts.

In this thesis, we directly consider Span($\boldsymbol{\eta}$), rather than $\boldsymbol{\eta}$, without assuming any statistical model. More generally our analysis does not rely on a particular function like (2.1). Instead, we consider conditional independence between $\mathbf{X}$ and Y. We focus on the intersection of all subspaces $\mathcal{S} \subseteq R^p$ such that

$$Y \perp\!\!\!\perp \mathbf{X} | \mathbf{P}_{\mathcal{S}}\mathbf{X} \tag{2.2}$$

where $\perp\!\!\!\perp$ indicates independence. When the intersection itself satisfies (2.2), it is called the *central dimension reduction subspace* or *central subspace* (CS) of the regression and denoted as $\mathcal{S}_{Y|\mathbf{X}}$. Therefore, we can focus on a lower dimensional projection $\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X}$ instead of $\mathbf{X}$ without losing any information on the regression. For background on the existence of the CS and related issues, see Cook (1998, Ch. 6). The dimension $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$ is called the structural dimension of the regression. It is easy to see that a linear transformation of the predictor $\mathbf{X}$ leads to a linear transformation of the CS. For example, define the standardized predictor $\mathbf{Z} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{X} - \mathrm{E}[\mathbf{X}])$, where

$\boldsymbol{\Sigma} = \text{Cov}(\mathbf{X})$. Then,

$$\mathcal{S}_{Y|\mathbf{X}} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathcal{S}_{Y|\mathbf{Z}}. \tag{2.3}$$

Therefore, without loss of generality, we may work on the $\mathbf{Z}$-scale and transform back to the original $\mathbf{X}$-scale when necessary.

2.1 Inverse Regression

We make one important assumption about the distribution of $\mathbf{X}$: The *linearity condition* requires that the standardized predictor $\mathbf{Z}$ satisfies

$$\text{E}[\mathbf{Z}|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{Z}}}\mathbf{Z}] = \mathbf{P}_{\mathcal{S}_{Y|\mathbf{Z}}}\mathbf{Z}.$$

This condition connects the central subspace with the inverse regression of $\mathbf{Z}$ on Y. When it holds, $\text{Span}\{\text{E}[\mathbf{Z}|Y]\} \subseteq \mathcal{S}_{Y|\mathbf{Z}}$ (Li 1991). Based on this result, we are able to estimate at least part of $\mathcal{S}_{Y|\mathbf{X}}$. When Y is discrete, it is easy to construct sample versions of $\text{E}[\mathbf{X}|Y]$. When Y is continuous, we consider a discrete version $\tilde{Y}$ of Y by partitioning the range of Y. One important fact is that $\mathcal{S}_{\tilde{Y}|\mathbf{X}} \subseteq \mathcal{S}_{Y|\mathbf{X}}$. When the number of values that $\tilde{Y}$ may take is reasonable large, we typically have $\mathcal{S}_{\tilde{Y}|\mathbf{X}} = \mathcal{S}_{Y|\mathbf{X}}$. Thus, without loss of generality, we assume Y is discrete and has a finite support $\{1, 2, \ldots, h\}$ unless specified otherwise. A value y of Y is called a slice.

Let us define a target space

$$\mathcal{S}_{\boldsymbol{\xi}} \equiv \oplus_{y=1}^{h}\text{Span}\{\boldsymbol{\xi}_y\}, \tag{2.4}$$

where

$$\boldsymbol{\xi}_y = \boldsymbol{\Sigma}^{-1}(\text{E}[\mathbf{X}|Y = y] - \text{E}[\mathbf{X}]) = \boldsymbol{\Sigma}^{-\frac{1}{2}}\text{E}[\mathbf{Z}|y],$$

and $\oplus$ indicates the sum of subspaces ($V_1 \oplus V_2 = \{v_1 + v_2 | v_1 \in V_1, v_2 \in V_2\}$). When the linearity condition holds, $\mathcal{S}_{\boldsymbol{\xi}}$ is a subset of $\mathcal{S}_{Y|\mathbf{X}}$. We often take this a step further and assume the *coverage condition*:

$$\oplus_{y=1}^{h} \text{Span}\{\text{E}[\mathbf{Z}|Y = y]\} = \mathcal{S}_{Y|\mathbf{Z}}.$$

Then, $\mathcal{S}_{\boldsymbol{\xi}} = \mathcal{S}_{Y|\mathbf{X}}$. Let $\boldsymbol{\beta}$ denote a basis of $\mathcal{S}_{\boldsymbol{\xi}}$, where $\boldsymbol{\beta}$ is a $p \times d$ matrix. An estimate of $\boldsymbol{\beta}$ provides an estimate of a basis of the CS. Throughout this thesis, when an basis of the CS or the *partial central subspace* (cf. Section 1.2) is involved, we implicitly assume the coverage condition. Inference about $\mathcal{S}_{\boldsymbol{\xi}}$ itself does not require the linearity condition or coverage condition.

By definition, for each y, there exists a vector $\boldsymbol{\gamma}_y$ such that $\boldsymbol{\xi}_y = \boldsymbol{\beta}\boldsymbol{\gamma}_y$. Define

$$\boldsymbol{\xi} \equiv (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_h) = \boldsymbol{\beta}\boldsymbol{\gamma},$$

where

$$\boldsymbol{\gamma} \equiv (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_h).$$

Let

$$\mathbf{f} \;=\; (f_1, f_2, \ldots, f_h)^T \tag{2.5}$$

where $f_y = \Pr(Y = y)$, and let $\mathbf{g} = \sqrt{\mathbf{f}}$. It is easy to see that

$$\boldsymbol{\xi}\mathbf{f} \;=\; \boldsymbol{\beta}\boldsymbol{\gamma}\mathbf{f} = 0, \tag{2.6}$$

which we call the *intrinsic location constraint*. We notice immediately that the parameterization $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is not identifiable, but $\boldsymbol{\beta}\boldsymbol{\gamma}$ is identifiable. However, this is not an issue since any basis $\boldsymbol{\beta}$ suffices to specify the CS. When

necessary, we may impose constraints on $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ to make the parameterization unique. For illustration, we describe one unique reparameterization here. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, where $\boldsymbol{\beta}_1 \in \mathbb{R}^{d\times d}$, $\boldsymbol{\beta}_2 \in \mathbb{R}^{(p-d)\times d}$. Without loss of generality, we assume that $\boldsymbol{\beta}_1$ is nonsingular. Otherwise, we only need to change the order of the elements in $\mathbf{X}$. Then,

$$\boldsymbol{\beta}\boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \boldsymbol{\gamma} = \begin{pmatrix} \mathbf{I}_d \\ \boldsymbol{\beta}_2\boldsymbol{\beta}_1^{-1} \end{pmatrix} \boldsymbol{\beta}_1\boldsymbol{\gamma}.$$

Therefore, we can impose $\boldsymbol{\beta} = (\mathbf{I}_d, \boldsymbol{\beta}^{*T})^T$. Now we have new parameters: $\boldsymbol{\beta}^* \in \mathbb{R}^{(p-d)\times d}$ and $\boldsymbol{\gamma} \in \mathbb{R}^{d\times h}$. This parameterization brings full rank Jacobian matrices and open parameter spaces, which is helpful in justification of some theoretical results. At the same time, any reparameterization does not affect the estimation of $\mathcal{S}_{\boldsymbol{\xi}}$ or the estimate's asymptotic properties. Therefore, we still use the overparameterized setting like $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ throughout this thesis and only visit the constrainted setting when necessary.

2.2 Minimum Discrepancy Approach

We now start out to develop dimension reduction methods via a minimum discrepancy approach. Suppose we have a sample of total size n for $(\mathbf{X}, Y)$, among which n_y points have $Y = y$. Let $\bar{\mathbf{X}}_{\bullet\bullet}$ be the total average of $\mathbf{X}$, and let $\bar{\mathbf{X}}_{y\bullet}$ be the average of the n_y points with $Y = y$. Let

$$\hat{\mathbf{f}} = (\frac{n_1}{n}, \ldots, \frac{n_h}{n})^T,$$

$\hat{\mathbf{g}} = \sqrt{\hat{\mathbf{f}}}$, and let $\hat{\mathbf{\Sigma}}$ denote the sample covariance. The sample version of $\boldsymbol{\xi}_y$ is

$$\hat{\boldsymbol{\xi}}_y = \hat{\mathbf{\Sigma}}^{-1}(\bar{\mathbf{X}}_{y\bullet} - \bar{\mathbf{X}}_{\bullet\bullet}) = \hat{\mathbf{\Sigma}}^{-\frac{1}{2}}\bar{\mathbf{Z}}_{y\bullet},$$

where $\bar{\mathbf{Z}}_{y\bullet} = \hat{\mathbf{\Sigma}}^{-\frac{1}{2}}(\bar{\mathbf{X}}_{y\bullet} - \bar{\mathbf{X}}_{\bullet\bullet})$. Let

$$\hat{\boldsymbol{\xi}} = (\hat{\boldsymbol{\xi}}_1, \ldots, \hat{\boldsymbol{\xi}}_h). \tag{2.7}$$

We know that the columns of $\boldsymbol{\xi}$ span a d-dimensional subspace. It is natural to estimate this subspace with a d-dimensional subspace that is closest to the columns of $\hat{\boldsymbol{\xi}}$ that is $\boldsymbol{\xi}$'s moment estimate. There are many ways to define "closeness". In this thesis, we consider quadratic discrepancy functions:

$$F_d(\mathbf{B}, \mathbf{C}) = (\operatorname{vec}(\hat{\boldsymbol{\xi}}\mathbf{M}_n) - \operatorname{vec}(\mathbf{BC}))^T \mathbf{V}_n (\operatorname{vec}(\hat{\boldsymbol{\xi}}\mathbf{M}_n) - \operatorname{vec}(\mathbf{BC})),$$

where $\mathbf{M}_n \in \mathbb{R}^{h\times l}$, $\mathbf{V}_n \in \mathbb{R}^{pl\times pl}$ is a positive definite matrix, $\mathbf{B} \in \mathbb{R}^{p\times d}$, and $\mathbf{C} \in \mathbb{R}^{d\times l}$. The matrix $\mathbf{M}_n$ decides how we organize the columns of $\hat{\boldsymbol{\xi}}$. Both $\mathbf{M}_n$ and $\mathbf{V}_n$ can be fixed or stochastic. The value of $\mathbf{B} \in \mathbb{R}^{p\times d}$ that minimizes the function provides an estimate of a subset of $\operatorname{Span}(\boldsymbol{\beta})$. The minimum value $\hat{F}_m$ of the function F_m can be used to construct a test statistic for the hypothesis $d = m$. Therefore, one pair of $(\mathbf{M}_n, \mathbf{V}_n)$ corresponds to one dimension reduction method. We call these methods the *MDA family*. Obviously, given $(\mathbf{M}_n, \mathbf{V}_n)$, solutions of this minimization are not unique because of the over-parameterization of the setting. However, this is not an issue since we are searching for $\operatorname{Span}(\boldsymbol{\beta})$ not $\boldsymbol{\beta}$ itself.

One way to estimate $\operatorname{Span}(\boldsymbol{\beta})$ is by letting $\mathbf{M}_n = I_h$ and $\mathbf{V}_n = \operatorname{diag}\{\mathbf{V}_{ny}\}$ a positive definite block diagonal matrix that converges to $\mathbf{V}$ in probability.

Consider

$$\begin{aligned} F_d(\mathbf{B},\mathbf{C}) &= (\operatorname{vec}(\hat{\boldsymbol{\xi}}) - \operatorname{vec}(\mathbf{BC}))^T \mathbf{V}_n (\operatorname{vec}(\hat{\boldsymbol{\xi}}) - \operatorname{vec}(\mathbf{BC})) \\ &= \sum_y (\hat{\boldsymbol{\xi}}_y - \mathbf{BC}_y)^T \mathbf{V}_{ny} (\hat{\boldsymbol{\xi}}_y - \mathbf{BC}_y) \end{aligned} \tag{2.8}$$

where $\mathbf{C}_y \in \mathbb{R}^d$. In Section 3.1 we will show $\sqrt{n}(\operatorname{vec}(\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}) - \operatorname{vec}(\boldsymbol{\beta}\boldsymbol{\gamma}\mathbf{D}_{\mathbf{f}}))$ converges to a normal random vector with zero mean and covariance matrix $\boldsymbol{\Gamma}^*$, where $\boldsymbol{\Gamma}^* \in \mathbb{R}^{ph \times ph}$ is singular because of the intrinsic location constraint. Thus, we rewrite (2.8) as

$$\begin{aligned} F_d(\mathbf{B},\widetilde{\mathbf{C}}) &= (\operatorname{vec}(\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}) - \operatorname{vec}(\mathbf{B}\widetilde{\mathbf{C}}))^T (\mathbf{D}_{\hat{\mathbf{f}}}^{-1} \otimes \mathbf{I}) \mathbf{V}_n (\mathbf{D}_{\hat{\mathbf{f}}}^{-1} \otimes \mathbf{I}) \\ &\quad (\operatorname{vec}(\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}) - \operatorname{vec}(\mathbf{B}\widetilde{\mathbf{C}})), \end{aligned}$$

where $\widetilde{\mathbf{C}} = \mathbf{C}\mathbf{D}_{\hat{\mathbf{f}}}$, and $\mathbf{D}_{\hat{\mathbf{f}}}$ is the diagonal matrix with elements of $\hat{\mathbf{f}}$ on the diagonal. Let

$$\mathbf{V}^* = (\mathbf{D}_{\mathbf{f}}^{-1} \otimes \mathbf{I}) \mathbf{V} (\mathbf{D}_{\mathbf{f}}^{-1} \otimes \mathbf{I}),$$

and let $\boldsymbol{\Delta}^*$ denote the Jacobian matrix for this discrepancy function

$$\boldsymbol{\Delta}^* = \left(\frac{\partial \operatorname{vec}(\mathbf{B}\widetilde{\mathbf{C}})}{\partial \operatorname{vec}(\mathbf{B})}, \frac{\partial \operatorname{vec}(\mathbf{B}\widetilde{\mathbf{C}})}{\partial \operatorname{vec}(\widetilde{\mathbf{C}})} \right) \Bigg|_{(\mathbf{B}=\boldsymbol{\beta}, \widetilde{\mathbf{C}}=\boldsymbol{\gamma}\mathbf{D}_{\mathbf{f}})}.$$

Then, the test statistic $n\hat{F}_d$ has an asymptotic chi-squared distribution only when

$$\boldsymbol{\Gamma}^* \mathbf{U} \boldsymbol{\Gamma}^* \mathbf{U} \boldsymbol{\Gamma}^* = \boldsymbol{\Gamma}^* \mathbf{U} \boldsymbol{\Gamma}^*, \tag{2.9}$$

where $\mathbf{U} = \mathbf{V}^* - \mathbf{V}^* \boldsymbol{\Delta}^* (\boldsymbol{\Delta}^{*T} \mathbf{V}^* \boldsymbol{\Delta}^*)^- \boldsymbol{\Delta}^{*T} \mathbf{V}^*$ (Shapiro 1986).

Unfortunately, generally condition (2.9) is not satisfied, i.e. there may not exist such a $\mathbf{V}$ for a particular regression of Y on $\mathbf{X}$. Even when such

a $\mathbf{V}$ exists, we need an estimate of $\boldsymbol{\Delta}^*$ that includes $\boldsymbol{\beta}$. But the estimation of $\boldsymbol{\beta}$ is the very reason we are looking for $\mathbf{V}$ in the first place. We may consider some iterative scheme; however this may unnecessarily complicate the straightforward idea. Therefore, the discrepancy function (2.8) is not optimal generally. We call the methods using (2.8) the *sub-optimal class* within the MDA family. One prominent member of this class is SIR. In Chapter 5, we rederive SIR in the minimum discrepancy approach. Even in this sub-optimal class, SIR is not the best method. Taking into account variation in $\mathrm{Cov}(\mathbf{X}|Y)$, we propose *simple inverse regression estimation* (Simple IRE) in Chapter 7. Simple IRE often beats SIR when we encounter large variation among the conditional covariances of $\mathbf{X}|Y$.

Is there an optimal discrepancy function we can use for estimation of the CS? The answer is yes. We know

$$\hat{\boldsymbol{\xi}}\hat{\mathbf{f}} = \hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}\mathbf{1} = 0, \tag{2.10}$$

where $\mathbf{1}$ is a vector with all elements being 1. This is a sample version of intrinsic location constraint (cf. (2.6)). Since one linear combination of the columns of $\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}$ is always a null vector that does not provide any information about $\mathcal{S}_{\boldsymbol{\xi}}$ (cf. (2.4)), an efficient objective function should examine only the orthogonal complement of $\mathrm{Span}(\mathbf{1})$. Therefore, instead of $\hat{\boldsymbol{\xi}}$ we may consider $\hat{\boldsymbol{\zeta}} \equiv \hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}\mathbf{A}$ in the construction of discrepancy functions, where $\mathbf{A} \in \mathbb{R}^{h\times(h-1)}$ such that $\mathbf{A}^T\mathbf{A} = \mathbf{I}_{h-1}$ and $\mathbf{A}^T\mathbf{1} = 0$. Thus, $\hat{\boldsymbol{\zeta}}$ converges to

$$\boldsymbol{\beta}\boldsymbol{\nu} \equiv \boldsymbol{\beta}\boldsymbol{\gamma}\mathbf{D}_{\mathbf{f}}\mathbf{A} \tag{2.11}$$

in probability, where $\boldsymbol{\nu} = \boldsymbol{\gamma}\mathbf{D}_{\mathbf{f}}\mathbf{A}$. We prove the asymptotic normality of

$\sqrt{n}(\,\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\boldsymbol{\beta\nu}))$ in Section 3.1:

$$\sqrt{n}(\,\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\boldsymbol{\beta\nu})) \xrightarrow{\mathcal{D}} \mathrm{Normal}(0, \boldsymbol{\Gamma}_{\hat{\boldsymbol{\zeta}}}),$$

where $\boldsymbol{\Gamma}_{\hat{\boldsymbol{\zeta}}} \in \mathbb{R}^{p(h-1)\times p(h-1)}$ is a nonsingular matrix. Now we construct a discrepancy function:

$$F_d(\mathbf{B}, \mathbf{C}) \;=\; (\,\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\mathbf{BC}))^T \mathbf{V}_n (\,\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\mathbf{BC})),$$

where $\mathbf{B} \in \mathbb{R}^{p\times d}$, $\mathbf{C} \in \mathbb{R}^{d\times(h-1)}$, and $\mathbf{V}_n \in \mathbb{R}^{p(h-1)\times p(h-1)}$ is a consistent estimate of $\boldsymbol{\Gamma}_{\hat{\boldsymbol{\zeta}}}^{-1}$. We call the method using this discrepancy function *optimal inverse regression estimation* (Optimal IRE). Optimal IRE is optimal: Its test statistic $n\hat{F}_d$ has an asymptotic chi-squared distribution, and its estimate of $\mathrm{vec}(\boldsymbol{\beta\nu})$ is *asymptotically efficient.* Notice that a function of $(\boldsymbol{\beta}, \boldsymbol{\nu})$ is uniquely defined only when it is a function of $\mathrm{vec}(\boldsymbol{\beta\nu})$. Borrowing the terminology from linear models, only functions of $\mathrm{vec}(\boldsymbol{\beta\nu})$ are estimable. The asymptotic efficiency we consider here means that the estimate of any function of $\mathrm{vec}(\boldsymbol{\beta\nu})$ that is obtained from this particular choice of $\mathbf{V}$ has smallest asymptotic variance among estimates from all possible $\mathbf{V}$. See Section 3.2 for the details. Asymptotic properties and computation of Optimal IRE will be addressed in Chapter 3.

Chapter 3

Optimal Inverse Regression Estimation

The essence of inverse regression estimation is to estimate the target space $\mathcal{S}_{\xi}$ by minimizing an appropriate objective function that measures the discrepancy between $\hat{\boldsymbol{\xi}}$ (cf. (2.7)) and the estimated space $\mathrm{Span}(\hat{\boldsymbol{\beta}})$. An efficient discrepancy function should take advantage of the sample intrinsic location constraint (cf. (2.10)). Therefore, instead of $\hat{\boldsymbol{\xi}}$ we consider $\hat{\boldsymbol{\zeta}} = \hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}\mathbf{A}$ in the construction of discrepancy functions, where $\mathbf{A} \in \mathbb{R}^{h\times(h-1)}$ such that $\mathbf{A}^T\mathbf{A} = \mathbf{I}_{h-1}$ and $\mathbf{A}^T\mathbf{1} = 0$. Thus, $\hat{\boldsymbol{\zeta}}$ converges to $\boldsymbol{\beta\nu} = \boldsymbol{\beta\gamma}\mathbf{D}_{\mathbf{f}}\mathbf{A}$ in probability. Let us state a fact and defer its proof until Section 3.1:

$$\sqrt{n}(\,\mathrm{vec}(\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}) - \mathrm{vec}(\boldsymbol{\beta\gamma}\mathbf{D}_{\mathbf{f}})) \xrightarrow{\mathcal{D}} \mathrm{Normal}(0, \boldsymbol{\Gamma}^*), \tag{3.1}$$

where $\boldsymbol{\Gamma}^* \in \mathbb{R}^{ph\times ph}$ is a singular matrix. Then,

$$\sqrt{n}(\,\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\boldsymbol{\beta\nu})) \xrightarrow{\mathcal{D}} \mathrm{Normal}(0, \boldsymbol{\Gamma}_{\hat{\boldsymbol{\zeta}}}),$$

where

$$\Gamma_{\hat{\zeta}} = (\mathbf{A}^T \otimes \mathbf{I})\Gamma^*(\mathbf{A} \otimes \mathbf{I}) \tag{3.2}$$

is nonsingular. Suppose a positive definite matrix $\mathbf{V}_n$ is a consistent estimate of $\Gamma_{\hat{\zeta}}^{-1}$. Then, the optimal discrepancy function is

$$F_d(\mathbf{B}, \mathbf{C}) = (\operatorname{vec}(\hat{\zeta}) - \operatorname{vec}(\mathbf{BC}))^T \mathbf{V}_n (\operatorname{vec}(\hat{\zeta}) - \operatorname{vec}(\mathbf{BC})), \tag{3.3}$$

where $\mathbf{B} \in \mathbb{R}^{p \times d}$, $\mathbf{C} \in \mathbb{R}^{d \times (h-1)}$. The values of $\mathbf{B}$ and $\mathbf{C}$ that minimize (3.3) are the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$. The discrepancy function (3.3) is optimal in two respects as we shall see. First, the estimate of $\operatorname{vec}(\boldsymbol{\beta\nu})$ is asymptotically efficient. Secondly, the test statistic for dimension—sample size times the minimum discrepancy value— has an asymptotic chi-squared distribution. We call the method using (3.3) *optimal inverse regression estimation* (Optimal IRE). Asymptotic properties and computation will be addressed in following sections.

3.1 Asymptotic Normality

From the above discussion, we understand that the establishment of the Optimal IRE discrepancy function hinges on the asymptotic normality of $\sqrt{n}(\operatorname{vec}(\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}) - \operatorname{vec}(\boldsymbol{\beta\gamma}\mathbf{D}_{\mathbf{f}}))$ and the estimation of the its limiting covariance matrix. In this section asymptotic normality is proved. Meanwhile, the limiting covariance matrix Γ^* is expressed as a covariance matrix of a random vector. Before we can report the results, some preparation is needed. First, we define h random variables J_y such that J_y equals 1 if a point is in the y-th

slice and 0 otherwise, $y = 1, 2, \ldots, h$. Then, $\mathrm{E}[J_y] = f_y$, where $f_y = \Pr(Y = y)$. Also define a random vector

$$\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_h)^T \tag{3.4}$$

where the random elements

$$\varepsilon_y = J_y - \mathrm{E}[J_y] - \mathbf{Z}^T \mathrm{E}[\mathbf{Z} J_y], \; y = 1, 2, \ldots, h,$$

are the population residuals from the ordinary least square fit of J_y on $\mathbf{Z}$. Now the asymptotic distribution of $\sqrt{n}(\,\mathrm{vec}(\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}) - \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\gamma}\mathbf{D}_{\mathbf{f}}))$ is specified in the following theorem. Since $\hat{\boldsymbol{\zeta}} = \hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}\mathbf{A}$ and $\mathbf{A}$ is a constant matrix, the asymptotic normality of $\sqrt{n}(\,\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\nu}))$ is easily obtained.

Theorem 1. *Assume that the data* $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, *are a simple random sample of* $(\mathbf{X}, Y)$. *All notation is as defined previously. Then*

$$\sqrt{n}(\,\mathrm{vec}(\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}) - \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\gamma}\mathbf{D}_{\mathbf{f}})) \xrightarrow{\mathcal{D}} \mathrm{Normal}(0, \boldsymbol{\Gamma}^*),$$

where $\boldsymbol{\Gamma}^* = \mathrm{Cov}(\,\mathrm{vec}(\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{Z}\boldsymbol{\varepsilon}^T)) \in \mathbb{R}^{ph \times ph}$.

Proof:

The strategy to showing asymptotic normality is to decompose $\sqrt{n}(\,\mathrm{vec}(\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}) - \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\gamma}\mathbf{D}_{\mathbf{f}}))$ as a summation of i.i.d. observations plus a remainder converging to 0 in probability. Then, by the central limit theorem, we obtain the desired results. In this decomposition process we need the following lemma that decomposes the difference between the inverse of a sample covariance matrix and its population value.

Lemma 1. *Suppose a random vector* $\mathbf{X}$ *has covariance matrix* $\mathbf{\Sigma} > 0$. *Then,*

$$\hat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1} = -n^{-1}\mathbf{\Sigma}^{-\frac{1}{2}} \sum_{j=1}^{n} (\mathbf{Z}_j \mathbf{Z}_j^T - \mathbf{I}) \mathbf{\Sigma}^{-\frac{1}{2}} + O_p(n^{-1}).$$

Here $\hat{\mathbf{\Sigma}}$ *is the sample covariance calculated from a sample of size* n *and* $\mathbf{Z} = \mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{X} - \mathrm{E}[\mathbf{X}])$ *is the standardized version of* $\mathbf{X}$.

Proof:

Note that

$$\begin{aligned}
\sqrt{n}(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}) &= n^{-\frac{1}{2}} \sum_{j=1}^{n} [(\mathbf{X}_j - \bar{\mathbf{X}})(\mathbf{X}_j - \bar{\mathbf{X}})^T - \mathbf{\Sigma}] \\
&= n^{-\frac{1}{2}} \sum_{j=1}^{n} [(\mathbf{X}_j - \boldsymbol{\mu})(\mathbf{X}_j - \boldsymbol{\mu})^T - \mathbf{\Sigma}] + O_p(n^{-\frac{1}{2}}) \\
&= n^{-\frac{1}{2}} \mathbf{\Sigma}^{\frac{1}{2}} \sum_{j=1}^{n} (\mathbf{Z}_j \mathbf{Z}_j^T - \mathbf{I}) \mathbf{\Sigma}^{\frac{1}{2}} + O_p(n^{-\frac{1}{2}}) \\
&= \mathbf{\Delta}_n + O_p(n^{-\frac{1}{2}}),
\end{aligned}$$

where $\boldsymbol{\mu} = \mathrm{E}[\mathbf{X}]$, $\mathbf{Z}_j = \mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{X}_j - \boldsymbol{\mu})$ and $\mathbf{\Delta}_n$ is defined implicitly. Denote $\hat{\mathbf{\Sigma}}^{-1} = \mathbf{\Sigma}^{-1} + n^{-\frac{1}{2}}\mathbf{D}_n + O_p(n^{-1})$. Here $\mathbf{D}_n$ is an $O_p(1)$ random matrix. Since $\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{\Sigma}} = \mathbf{I}$, by simple algebra, we have

$$\mathbf{D}_n = -\mathbf{\Sigma}^{-1}\mathbf{\Delta}_n\mathbf{\Sigma}^{-1} = -n^{-\frac{1}{2}}\mathbf{\Sigma}^{-\frac{1}{2}} \sum_{j=1}^{n} (\mathbf{Z}_j \mathbf{Z}_j^T - \mathbf{I}) \mathbf{\Sigma}^{-\frac{1}{2}}.$$

Therefore,

$$\hat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1} = -n^{-1}\mathbf{\Sigma}^{-\frac{1}{2}} \sum_{j=1}^{n} (\mathbf{Z}_j \mathbf{Z}_j^T - \mathbf{I}) \mathbf{\Sigma}^{-\frac{1}{2}} + O_p(n^{-1}). \quad \square$$

Recall that $\bar{\mathbf{X}}_{y\bullet}$ is the average of the n_y observations in the yth slice and $\bar{\mathbf{X}}_{\bullet\bullet}$ is the grand average of all n observations. Letting $\boldsymbol{\mu}_y = \mathrm{E}[\bar{\mathbf{X}}_{y\bullet}]$, $\boldsymbol{\mu} = \mathrm{E}[\bar{\mathbf{X}}_{\bullet\bullet}]$, consider

$$\begin{aligned}
& \sqrt{n}(\hat{f}_y\hat{\boldsymbol{\xi}}_y - f_y\boldsymbol{\xi}_y) \\
= \; & \sqrt{n}\hat{f}_y\hat{\boldsymbol{\Sigma}}^{-1}(\bar{\mathbf{X}}_{y\bullet} - \bar{\mathbf{X}}_{\bullet\bullet}) - \sqrt{n}f_y\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_y - \boldsymbol{\mu}) \\
= \; & \sqrt{n}(\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1})f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu}) + \sqrt{n}\boldsymbol{\Sigma}^{-1}[\hat{f}_y(\bar{\mathbf{X}}_{y\bullet} - \bar{\mathbf{X}}_{\bullet\bullet}) - f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu})] \\
& + \sqrt{n}(\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1})[\hat{f}_y(\bar{\mathbf{X}}_{y\bullet} - \bar{\mathbf{X}}_{\bullet\bullet}) - f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu})] \\
= \; & \sqrt{n}(\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1})f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu}) + \sqrt{n}\boldsymbol{\Sigma}^{-1}[\hat{f}_y(\bar{\mathbf{X}}_{y\bullet} - \bar{\mathbf{X}}_{\bullet\bullet}) - f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu})] \\
& + O_p(n^{-\frac{1}{2}}). \qquad (3.5)
\end{aligned}$$

By Lemma 1, we have

$$\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1} = -n^{-1}\boldsymbol{\Sigma}^{-\frac{1}{2}}\sum_{j=1}^{n}(\mathbf{Z}_j\mathbf{Z}_j^T - \mathbf{I})\boldsymbol{\Sigma}^{-\frac{1}{2}} + O_p(n^{-1}).$$

Therefore, the first term in (3.5) can be simplified as

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1})f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu}) &= -n^{-\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\sum_{j=1}^{n}(\mathbf{Z}_j\mathbf{Z}_j^T - \mathbf{I})\boldsymbol{\Sigma}^{-\frac{1}{2}}f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu}) + O_p(n^{-\frac{1}{2}}) \\
&= -n^{-\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\sum_{j=1}^{n}(\mathbf{Z}_j\mathbf{Z}_j^T - \mathbf{I})\mathrm{E}[\mathbf{Z}J_y] + O_p(n^{-\frac{1}{2}}). \qquad (3.6)
\end{aligned}$$

Meanwhile, letting J_{yj} denote the value of J_y for the jth observation, $j = 1, 2, \ldots, n$, we have

$$\begin{aligned}
\hat{f}_y(\bar{\mathbf{X}}_{y\bullet} - \bar{\mathbf{X}}_{\bullet\bullet}) &= \frac{1}{n}\sum_{j=1}^{n}[(\mathbf{X}_j - \bar{\mathbf{X}}_{\bullet\bullet})J_{yj}] \\
&= \frac{1}{n}\sum_{j=1}^{n}[(\mathbf{X}_j - \bar{\mathbf{X}}_{\bullet\bullet})(J_{yj} - \mathrm{E}[J_y])] \\
&= \frac{1}{n}\sum_{j=1}^{n}[(\mathbf{X}_j - \boldsymbol{\mu})(J_{yj} - \mathrm{E}[J_y])] - \frac{1}{n}\sum_{j=1}^{n}[(\bar{\mathbf{X}}_{\bullet\bullet} - \boldsymbol{\mu})(J_{yj} - E[J_y])] \\
&= \frac{1}{n}\sum_{j=1}^{n}[(\mathbf{X}_j - \boldsymbol{\mu})(J_{yj} - \mathrm{E}[J_y])] - \frac{1}{n}(\bar{\mathbf{X}}_{\bullet\bullet} - \boldsymbol{\mu})\sum_{j=1}^{n}(J_{yj} - \mathrm{E}[J_y]) \\
&= \frac{1}{n}\sum_{j=1}^{n}[(\mathbf{X}_j - \boldsymbol{\mu})(J_{yj} - \mathrm{E}[J_y])] + O_p(n^{-1}).
\end{aligned}$$

Therefore, the second term in (3.5) can be simplified as

$$\begin{aligned}
&\sqrt{n}\boldsymbol{\Sigma}^{-1}[\hat{f}_y(\bar{\mathbf{X}}_{y\bullet} - \bar{\mathbf{X}}_{\bullet\bullet}) - f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu})] \\
= \; & n^{-\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\sum_{j=1}^{n}[\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{X}_j - \boldsymbol{\mu})(J_{yj} - \mathrm{E}[J_y])] - \sqrt{n}\boldsymbol{\Sigma}^{-1}f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu}) + O_p(n^{-\frac{1}{2}}) \\
= \; & n^{-\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\sum_{j=1}^{n}[\mathbf{Z}_j(J_{yj} - \mathrm{E}[J_y])] - \sqrt{n}\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathrm{E}[\mathbf{Z}J_y] + O_p(n^{-\frac{1}{2}}) \\
= \; & n^{-\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\sum_{j=1}^{n}[\mathbf{Z}_j(J_{yj} - \mathrm{E}[J_y]) - \mathrm{E}[\mathbf{Z}J_y]] + O_p(n^{-\frac{1}{2}}). \qquad (3.7)
\end{aligned}$$

We plug (3.6) and (3.7) into (3.5) and obtain

$$\begin{aligned}
& \sqrt{n}(\hat{f}_y\hat{\boldsymbol{\xi}}_y - f_y\boldsymbol{\xi}_y) \\
= \; & n^{-\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\sum_{j=1}^{n}[\mathbf{Z}_j(J_{yj} - \mathrm{E}[J_y]) - \mathrm{E}[\mathbf{Z}J_y] - (\mathbf{Z}_j\mathbf{Z}_j^T - \mathbf{I})\mathrm{E}[\mathbf{Z}J_y]] + O_p(n^{-\frac{1}{2}}) \\
= \; & n^{-\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\sum_{j=1}^{n}[\mathbf{Z}_j(J_{yj} - \mathrm{E}[J_y] - \mathbf{Z}_j^T\mathrm{E}[\mathbf{Z}J_y])] + O_p(n^{-\frac{1}{2}}) \\
= \; & n^{-\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\sum_{j=1}^{n}[\mathbf{Z}_j\varepsilon_{yj}] + O_p(n^{-\frac{1}{2}}),
\end{aligned}$$

where $\varepsilon_{yj} = J_{yj} - \mathrm{E}[J_y] - \mathbf{Z}_j^T\mathrm{E}[\mathbf{Z}J_y]$ is the jth value for ε_y. Let $\boldsymbol{\epsilon}_j = [\varepsilon_{1j}, \cdots, \varepsilon_{hj}]^T$ be the jth value for the random vector $\boldsymbol{\varepsilon}$ (cf. (3.4)). We have

$$\sqrt{n}(\,\mathrm{vec}(\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}) - \,\mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\gamma}\mathbf{D}_{\mathbf{f}})) = n^{-\frac{1}{2}}\sum_{j=1}^{n}\mathrm{vec}(\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{Z}_j\boldsymbol{\epsilon}_j^T) + O_p(n^{-\frac{1}{2}})$$

where $(\mathbf{Z}_j, \boldsymbol{\epsilon}_j)$ are i.i.d. random vectors. Thus,

$$\sqrt{n}(\,\mathrm{vec}(\hat{\boldsymbol{\xi}}\mathbf{D}_{\hat{\mathbf{f}}}) - \,\mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\gamma}\mathbf{D}_{\mathbf{f}})) \xrightarrow{\mathcal{D}} \mathrm{Normal}(0, \boldsymbol{\Gamma}^*),$$

where

$$\boldsymbol{\Gamma}^* = \mathrm{Cov}(\,\mathrm{vec}(\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{Z}\boldsymbol{\varepsilon}^T)).$$

3.2 Asymptotic Properties of Optimal IRE

In this section, we show that the value of $\mathbf{B}$ that minimizes (3.3) is a consistent estimate of $\boldsymbol{\beta}$ that is a basis of the CS under special conditions we discussed previously. Meanwhile, we use $n\hat{F}_m$ as the test statistic for the null hypothesis that $\dim(\mathcal{S}_{\boldsymbol{\xi}}) = m$, where $\hat{F}_m$ is the minimum discrepancy value.

Before we report the asymptotic properties of Optimal IRE, a little setup is necessary. We need the $p(h-1) \times (p+h-1)d$ matrix

$$\boldsymbol{\Delta}_{\boldsymbol{\zeta}} \equiv (\boldsymbol{\nu}^T \otimes \mathbf{I}_p, \mathbf{I}_{h-1} \otimes \boldsymbol{\beta}). \tag{3.8}$$

The matrix $\boldsymbol{\Delta}_{\boldsymbol{\zeta}}$ is the Jacobian matrix for the discrepancy function

$$\boldsymbol{\Delta}_{\boldsymbol{\zeta}} = \left(\frac{\partial \operatorname{vec}(\mathbf{BC})}{\partial \operatorname{vec}(\mathbf{B})}, \frac{\partial \operatorname{vec}(\mathbf{BC})}{\partial \operatorname{vec}(\mathbf{C})} \right) \Bigg|_{(\mathbf{B}=\boldsymbol{\beta}, \mathbf{C}=\boldsymbol{\nu})},$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$ are as defined previously in Section 2.2 (cf. (2.11)). Asymptotic properties of Optimal IRE are given in the following theorem.

Theorem 2. *Assume that the data* $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, *are a simple random sample of* $(\mathbf{X}, Y)$. *Let* $\mathcal{S}_{\boldsymbol{\xi}} = \oplus_{y=1}^{h} \operatorname{Span}\{\boldsymbol{\xi}_y\}$, *let* $d = \dim(\mathcal{S}_{\boldsymbol{\xi}})$ *and let* $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\nu}}) = \arg_{\mathbf{B},\mathbf{C}} \min F_d(\mathbf{B}, \mathbf{C})$, *where*

$$F_d(\mathbf{B}, \mathbf{C}) = (\operatorname{vec}(\hat{\boldsymbol{\zeta}}) - \operatorname{vec}(\mathbf{BC}))^T \mathbf{V}_n (\operatorname{vec}(\hat{\boldsymbol{\zeta}}) - \operatorname{vec}(\mathbf{BC})),$$

where $\mathbf{V}_n$ *converges in probability to* $\mathbf{V} = \boldsymbol{\Gamma}_{\hat{\boldsymbol{\zeta}}}^{-1}$ *(cf. (3.2)). Then*

1. *The estimate* $\operatorname{vec}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\nu}})$ *is asymptotically efficient, and* $\sqrt{n}(\operatorname{vec}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\nu}}) - \operatorname{vec}(\boldsymbol{\beta}\boldsymbol{\nu}))$ *is asymptotically normal with zero mean and covariance matrix* $\boldsymbol{\Delta}_{\boldsymbol{\zeta}}(\boldsymbol{\Delta}_{\boldsymbol{\zeta}}^T \mathbf{V} \boldsymbol{\Delta}_{\boldsymbol{\zeta}})^{-} \boldsymbol{\Delta}_{\boldsymbol{\zeta}}^T$.

2. $n\hat{F}_d$ *has an asymptotic chi-squared distribution with degrees of freedom* $(p-d)(h-d-1)$.

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

This theorem is quite general, requiring none of the special conditions discussed previously. A value $\hat{\boldsymbol{\beta}}$ of $\mathbf{B}$ that minimizes the discrepancy function $F_d(\mathbf{B}, \mathbf{C})$ always provides a consistent estimate of a basis for $\mathcal{S}_{\boldsymbol{\xi}}$, and

this theorem allows us to test hypotheses about its dimension. However, without some of the special conditions, $\mathcal{S}_{\xi}$ might not be a useful population parameter and therefore tests on its dimension might not be of interest.

If the linearity condition holds then $\mathcal{S}_{\xi} \subseteq \mathcal{S}_{Y|\mathbf{X}}$. The subspace spanned by $\hat{\boldsymbol{\beta}}$ is still a consistent estimate of $\mathcal{S}_{\xi}$, which is now a subspace of the CS. In this case we are able to use Theorem 2 to infer about a possibly proper subset of the CS. If the linearity and coverage conditions both hold, $\mathcal{S}_{\xi} = \mathcal{S}_{Y|\mathbf{X}}$, and we can use Theorem 2 to infer about the full CS.

For a sequence of $\mathbf{V}_n > 0$ that converges to $\mathbf{V} > 0$, the minimization of the function (3.3) always provides a consistent estimate of $\mathrm{vec}(\boldsymbol{\beta\nu})$. But the particular choice of $\mathbf{V}_n$ in Theorem 2 makes the estimate have the smallest asymptotic covariance. The proof of Theorem 2 hinges on a theorem by Shapiro (1986) on the asymptotics of over-parameterized structural models. Shapiro's results are given in next section, along with a proof of Theorem 2.

Optimal IRE that incorporates a right nonsingular transformation of $\hat{\boldsymbol{\zeta}}$ provides the same test statistic and the same asymptotic efficiency. Suppose $\mathbf{S} \in \mathbb{R}^{(h-1)\times(h-1)}$ is a fixed nonsingular matrix. We consider a new discrepancy function associated with $\hat{\boldsymbol{\zeta}}\mathbf{S}$:

$$G_d(\mathbf{B},\mathbf{C}) \;=\; (\,\mathrm{vec}(\hat{\boldsymbol{\zeta}}\mathbf{S}) - \mathrm{vec}(\mathbf{BC}))^T \mathcal{V}_n (\,\mathrm{vec}(\hat{\boldsymbol{\zeta}}\mathbf{S}) - \mathrm{vec}(\mathbf{BC})). \quad (3.9)$$

The limit of $n\mathrm{Cov}(\,\mathrm{vec}(\hat{\boldsymbol{\zeta}}\mathbf{S}) - \mathrm{vec}(\boldsymbol{\beta\nu}\mathbf{S}))$ is $(\mathbf{S}^T \otimes \mathbf{I})\boldsymbol{\Gamma}_{\hat{\zeta}}(\mathbf{S} \otimes \mathbf{I})$. For efficiency,

$\mathcal{V}_n$ should converge to

$$\mathcal{V} = (\mathbf{S}^{-1} \otimes \mathbf{I}_p)\boldsymbol{\Gamma}_{\hat{\boldsymbol{\zeta}}}^{-1}(\mathbf{S}^{-T} \otimes \mathbf{I}_p).$$

Let $\mathcal{V}_n = (\mathbf{S}^{-1} \otimes \mathbf{I}_p)\mathbf{V}_n(\mathbf{S}^{-T} \otimes \mathbf{I}_p)$. Then

$$\begin{aligned}
\min_{\mathbf{B},\mathbf{C}} G_d(\mathbf{B},\mathbf{C}) &= \min_{\mathbf{B},\mathbf{C}} (\operatorname{vec}(\hat{\boldsymbol{\zeta}}\mathbf{S}) - \operatorname{vec}(\mathbf{BC}))^T \mathcal{V}_n (\operatorname{vec}(\hat{\boldsymbol{\zeta}}\mathbf{S}) - \operatorname{vec}(\mathbf{BC})) \\
&= \min_{\mathbf{B},\mathbf{C}} (\operatorname{vec}(\hat{\boldsymbol{\zeta}}) - \operatorname{vec}(\mathbf{BC}))^T \mathbf{V}_n (\operatorname{vec}(\hat{\boldsymbol{\zeta}}) - \operatorname{vec}(\mathbf{BC})) \\
&= \min_{\mathbf{B},\mathbf{C}} F_d(\mathbf{B},\mathbf{C}).
\end{aligned}$$

The minimum values of F and G are the same. Meanwhile, the new Jacobian matrix

$$\Delta \equiv (\mathbf{S}^T\boldsymbol{\nu}^T \otimes \mathbf{I}_p, \mathbf{I}_{h-1} \otimes \boldsymbol{\beta}) = (\mathbf{S}^T \otimes \mathbf{I}_p)\boldsymbol{\Delta}_{\boldsymbol{\zeta}} \begin{bmatrix} \mathbf{I}_{pd} & \\ & \mathbf{S}^{-T} \otimes \mathbf{I}_d \end{bmatrix}.$$

Therefore,

$$\begin{aligned}
\Delta\mathcal{V}\Delta &= \begin{bmatrix} \mathbf{I}_{pd} & \\ & \mathbf{S}^{-1} \otimes \mathbf{I}_d \end{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{\zeta}}^T (\mathbf{S} \otimes \mathbf{I}_p)\mathcal{V}(\mathbf{S}^T \otimes \mathbf{I}_p)\boldsymbol{\Delta}_{\boldsymbol{\zeta}} \begin{bmatrix} \mathbf{I}_{pd} & \\ & \mathbf{S}^{-T} \otimes \mathbf{I}_d \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{I}_{pd} & \\ & \mathbf{S}^{-1} \otimes \mathbf{I}_d \end{bmatrix} \boldsymbol{\Delta}_{\boldsymbol{\zeta}}^T \mathbf{V} \boldsymbol{\Delta}_{\boldsymbol{\zeta}} \begin{bmatrix} \mathbf{I}_{pd} & \\ & \mathbf{S}^{-T} \otimes \mathbf{I}_d \end{bmatrix}
\end{aligned}$$

and

$$\begin{aligned}
&\Delta(\Delta^T\mathcal{V}\Delta)^-\Delta^T \\
&= \Delta \begin{bmatrix} \mathbf{I}_{pd} & \\ & \mathbf{S}^T \otimes \mathbf{I}_d \end{bmatrix} (\boldsymbol{\Delta}_{\boldsymbol{\zeta}}^T \mathbf{V} \boldsymbol{\Delta}_{\boldsymbol{\zeta}})^- \begin{bmatrix} \mathbf{I}_{pd} & \\ & \mathbf{S} \otimes \mathbf{I}_d \end{bmatrix} \Delta^T \\
&= (\mathbf{S}^T \otimes \mathbf{I}_p)\boldsymbol{\Delta}_{\boldsymbol{\zeta}}(\boldsymbol{\Delta}_{\boldsymbol{\zeta}}^T \mathbf{V} \boldsymbol{\Delta}_{\boldsymbol{\zeta}})^- \boldsymbol{\Delta}_{\boldsymbol{\zeta}}^T (\mathbf{S} \otimes \mathbf{I}_p).
\end{aligned}$$

Let $(\hat{\mathbf{B}}, \hat{\mathbf{C}})$ be the values of $(\mathbf{B}, \mathbf{C})$ that minimize $G_d(\mathbf{B}, \mathbf{C})$. Based on Theorem 2, $\sqrt{n}(\operatorname{vec}(\hat{\mathbf{B}}\hat{\mathbf{C}}) - \operatorname{vec}(\boldsymbol{\beta}\boldsymbol{\nu}\mathbf{S}))$ has an asymptotically normal distribution

with zero mean and covariance matrix $(\mathbf{S}^T \otimes \mathbf{I}_p)\boldsymbol{\Delta}_\zeta(\boldsymbol{\Delta}_\zeta^T \mathbf{V} \boldsymbol{\Delta}_\zeta)^- \boldsymbol{\Delta}_\zeta^T (\mathbf{S} \otimes \mathbf{I}_p)$. It results in the same asymptotic covariance matrix if we use the estimates from $F_d(\mathbf{B}, \mathbf{C})$ to estimate $\text{vec}(\boldsymbol{\beta\nu}\mathbf{S})$.

3.3 About Theorem 2

3.3.1 Preparations

The proof of Theorem 2 hinges on Shapiro's (1986) results on the asymptotics of over-parameterized discrepancy functions and two supplemental lemmas. We first give these results and then show how they can be used to prove the theorem.

Proposition 1. *(Shapiro 1986, Prop. 3.1, 3.2, 4.1,and 5.1) Suppose θ is a q-dimensional parameter vector which lies in an open and connected parameter space $\Theta \subseteq \mathbb{R}^q$. Let θ_0 denote the true value of θ. Define $g(\theta) = (g_1(\theta), \ldots, g_m(\theta))^T : \Theta \to \mathbb{R}^m$, where $g_i(\theta)$ is twice continuously differentiable on Θ, $i = 1, \ldots, m$. The Jacobian matrix $\boldsymbol{\Delta} = \frac{\partial g(\theta)}{\partial \theta}|_{\theta=\theta_0}$ need not be of full rank, so g can be over-parameterized. Also assume*

1. *$\boldsymbol{\tau}_n$ is an asymptotically normal estimate of the population value $g(\theta_0)$: $\sqrt{n}(\boldsymbol{\tau}_n - g(\theta_0)) \xrightarrow{\mathcal{D}} \text{Normal}(0, \boldsymbol{\Gamma})$, where n is the sample size.*

2. *For a known inner product matrix $\mathbf{V}$, the discrepancy function*

$$H(\boldsymbol{\tau}_n, g(\theta)) = (\boldsymbol{\tau}_n - g(\theta))^T \mathbf{V} (\boldsymbol{\tau}_n - g(\theta))$$

satisfies following properties:

p1 $H(a,b) \geq 0 \; \forall \; a, \; b \in \mathbb{R}^m$.

p2 $H(a,b) = 0$ *if and only if* $a = b$.

p3 H *is at least twice continuously differentiable in* a *and* b.

p4 *There are positive constants* δ *and* ϵ *such that* $H(a,b) \geq \epsilon$ *whenever* $\|a - b\| \geq \delta$, *where* $\|\cdot\|$ *means ordinary Euclidean distance.*

3. *The point* θ_0 *is regular.*

4. $\text{rank}(\boldsymbol{\Delta}) = \text{rank}(\boldsymbol{\Delta}^T \mathbf{V} \boldsymbol{\Delta})$.

Then

1. *Letting* $\hat{H} = H(\boldsymbol{\tau}_n, g(\hat{\theta}))$ *denote the value of the discrepancy function minimized over* Θ, *the asymptotic distribution of* $n\hat{H}$ *is the same as the distribution of the quadratic form* $\mathbf{W}^T \mathbf{U} \mathbf{W}$, *where* $\mathbf{W} \sim \text{Normal}(0, \boldsymbol{\Gamma})$,

$$\mathbf{U} = \mathbf{V} - \mathbf{V}\boldsymbol{\Delta}(\boldsymbol{\Delta}^T \mathbf{V} \boldsymbol{\Delta})^- \boldsymbol{\Delta}^T \mathbf{V} = \mathbf{V}^{\frac{1}{2}} \mathbf{Q}_{\boldsymbol{\Phi}} \mathbf{V}^{\frac{1}{2}}$$

   and $\boldsymbol{\Phi} = \mathbf{V}^{\frac{1}{2}} \boldsymbol{\Delta}$.

2. *If* $\boldsymbol{\Gamma}\mathbf{U}\boldsymbol{\Gamma}\mathbf{U}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}\mathbf{U}\boldsymbol{\Gamma}$, *then* $n\hat{H} \xrightarrow{\mathcal{D}} \chi^2_D$, *where the degrees of freedom* $D = \text{trace}(\mathbf{U}\boldsymbol{\Gamma})$.

3. *The estimate* $g(\hat{\theta})$ *that minimizes the discrepancy function is a consistent estimator of* $g(\theta_0)$ *and* $\sqrt{n}(g(\hat{\theta}) - g(\theta_0))$ *has an asymptotically normal distribution with zero mean and covariance matrix* $\mathbf{V}^{-\frac{1}{2}} \mathbf{P}_{\boldsymbol{\Phi}} \mathbf{V}^{\frac{1}{2}} \boldsymbol{\Gamma} \mathbf{V}^{\frac{1}{2}} \mathbf{P}_{\boldsymbol{\Phi}} \mathbf{V}^{-\frac{1}{2}}$.

4. *When* $\boldsymbol{\Gamma}$ *is nonsingular,* $g(\hat{\theta})$ *is asymptotically efficient and* $n\hat{H} \xrightarrow{\mathcal{D}} \chi^2_k$, *where the degrees of freedom* $k = m - \text{rank}(\boldsymbol{\Delta})$, *if and only if* $\mathbf{V} = (\boldsymbol{\Gamma} + \boldsymbol{\Delta}\mathbf{D}\boldsymbol{\Delta}^T)^{-1}$, *where* $\mathbf{D}$ *is an arbitrary symmetric matrix.*

In our adaptation of Shapiro's results, the inner product matrix $\mathbf{V}$ is often random rather than fixed as required in Proposition 1. The next lemmas allow us to connect minimum discrepancy functions with fixed inner products to those with random inner products. The first lemma deals with the asymptotic distribution of the minimum discrepancy value. The second lemma is about asymptotic properties of the estimate of Span($\boldsymbol{\beta}$).

Lemma 2. *Let* $\{\mathbf{Y}_n\} \in \mathbb{R}^s$ *be a sequence of random vectors, and let* $\boldsymbol{\xi} \in \Xi \subseteq \mathbb{R}^s$. *Suppose* $\{\mathbf{V}_n > 0\}$ *is a sequence of* $s \times s$ *matrices that converges to* $\mathbf{V} > 0$ *in probability. If*

$$n\hat{H}_{\mathbf{V}} = \min_{\xi \in \Xi} n(\mathbf{Y}_n - \boldsymbol{\xi})^T \mathbf{V}(\mathbf{Y}_n - \boldsymbol{\xi}) \xrightarrow{\mathcal{D}} \Psi,$$

then $n\hat{H}_{\mathbf{V}_n} = \min_{\xi \in \Xi} n(\mathbf{Y}_n - \boldsymbol{\xi})^T \mathbf{V}_n(\mathbf{Y}_n - \boldsymbol{\xi})$ *also converges in distribution to* Ψ *and vice versa.*

Moreover, let $\hat{\boldsymbol{\xi}}_1$ *and* $\hat{\boldsymbol{\xi}}_2$ *be the values of* $\boldsymbol{\xi}$ *which reach* $n\hat{H}_{\mathbf{V}}$ *and* $n\hat{H}_{\mathbf{V}_n}$ *respectively. If* $\mathbf{V}^{\frac{1}{2}}\mathbf{Y}_n \xrightarrow{p} \boldsymbol{\alpha}$, *then both* $\mathbf{V}^{\frac{1}{2}}\hat{\boldsymbol{\xi}}_1$ *and* $\mathbf{V}_n^{\frac{1}{2}}\hat{\boldsymbol{\xi}}_2$ *converge to* $\boldsymbol{\alpha}$ *in probability.*

Proof:

Since $\mathbf{V}_n \to \mathbf{V}$ in probability, $\Pr[(1-\epsilon)\mathbf{V} < \mathbf{V}_n < (1+\epsilon)\mathbf{V}] \to 1 \ \forall \epsilon > 0$. For any $\boldsymbol{\xi} \in \Xi$, if $(1-\epsilon)\mathbf{V} < \mathbf{V}_n < (1+\epsilon)\mathbf{V}$, then

$$(\mathbf{Y}_n - \boldsymbol{\xi})^T(1-\epsilon)\mathbf{V}(\mathbf{Y}_n - \boldsymbol{\xi}) \leq (\mathbf{Y}_n - \boldsymbol{\xi})^T\mathbf{V}_n(\mathbf{Y}_n - \boldsymbol{\xi}) \leq (\mathbf{Y}_n - \boldsymbol{\xi})^T(1+\epsilon)\mathbf{V}(\mathbf{Y}_n - \boldsymbol{\xi}).$$

Hence, the minimum of these functions keeps the same ordering:

$$(1-\epsilon)n\hat{H}_{\mathbf{V}} \leq n\hat{H}_{\mathbf{V}_n} \leq (1+\epsilon)n\hat{H}_{\mathbf{V}}.$$

Therefore, $\Pr[|\frac{n\hat{H}_{\mathbf{V}_n}}{n\hat{H}_{\mathbf{V}}} - 1| \leq \epsilon] \to 1$, i.e. $\frac{\hat{H}_{\mathbf{V}_n}}{\hat{H}_{\mathbf{V}}} \xrightarrow{p} 1$. By Slutsky's theorem, $n\hat{H}_{\mathbf{V}_n} \xrightarrow{\mathcal{D}} \Psi$.

Furthermore, since $n\hat{H}_{\mathbf{V}} = n\|\mathbf{V}^{\frac{1}{2}}\mathbf{Y}_n - \mathbf{V}^{\frac{1}{2}}\hat{\boldsymbol{\xi}}_1\|^2 \to \Psi$, $\forall \epsilon > 0$,

$$\lim_{n\to\infty} \Pr[\|\mathbf{V}^{\frac{1}{2}}\mathbf{Y}_n - \mathbf{V}^{\frac{1}{2}}\hat{\boldsymbol{\xi}}_1\|^2 > \epsilon] = \lim \Pr[n\hat{H}_{\mathbf{V}} > n\epsilon] = \lim \Pr[\Psi > n\epsilon] = 0.$$

Since $\mathbf{V}_n^{\frac{1}{2}} \xrightarrow{p} \mathbf{V}^{\frac{1}{2}}$, $\lim_{n\to\infty} \mathbf{V}^{\frac{1}{2}}\mathbf{Y}_n = \lim \mathbf{V}_n^{\frac{1}{2}}\hat{\boldsymbol{\xi}}_1 = \lim \mathbf{V}_n^{\frac{1}{2}}\hat{\boldsymbol{\xi}}_2$ if the first limit exits. □

Lemma 3. *Let $\mathcal{X}_n$ denote a simple random sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, where $\mathbf{X}_i$ can be a scalar or a vector. The distribution of $\mathbf{X}$ depends on parameters that include a vector θ in $\Theta \subseteq \mathbb{R}^k$. Let θ_0 be the true value of θ. Assume that*

1. *Θ is an open set.*
2. *The mapping $p(\theta)$ from Θ into $\mathbb{R}^s$ is one-to-one, bicontinuous, and twice continuously differentiable. Let $\mathbf{D}(\theta) = \frac{\partial p(\theta)}{\partial \theta} \in \mathbb{R}^{s\times k}$, and $\mathbf{D}_0 = \mathbf{D}(\theta_0)$.*
3. *$\mathbf{Y}_n = \mathbf{Y}_n(\mathcal{X}_n) \in \mathbb{R}^s$ is a consistent estimate of $p(\theta_0)$ with*
$$\sqrt{n}(\mathbf{Y}_n - p(\theta_0)) \xrightarrow{\mathcal{D}} \text{Normal}(0, \boldsymbol{\Gamma}).$$
4. *$\mathbf{V}_n = \mathbf{V}_n(\mathcal{X}_n)$ is a positive definite matrix that converges to a constant matrix $\mathbf{V}$ in probability.*

Define a discrepancy function as

$$F(\mathbf{Y}_n, p(\theta)) = (\mathbf{Y}_n - p(\theta))^T \mathbf{V}_n (\mathbf{Y}_n - p(\theta)).$$

Let $\hat{\theta} = \hat{\theta}(\mathcal{X}_n)$ be the value of θ that minimizes F. Then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} \text{Normal}(0, (\mathbf{D}_0^T\mathbf{V}\mathbf{D}_0)^{-1}\mathbf{D}_0^T\mathbf{V}\boldsymbol{\Gamma}\mathbf{V}\mathbf{D}_0(\mathbf{D}_0^T\mathbf{V}\mathbf{D}_0)^{-1}),$$

and

$$\sqrt{n}(p(\hat{\theta}) - p(\theta_0)) \xrightarrow{\mathcal{D}} \text{Normal}(0, \mathbf{D}_0(\mathbf{D}_0^T\mathbf{V}\mathbf{D}_0)^{-1}\mathbf{D}_0^T\mathbf{V}\boldsymbol{\Gamma}\mathbf{V}\mathbf{D}_0(\mathbf{D}_0^T\mathbf{V}\mathbf{D}_0)^{-1}\mathbf{D}_0).$$

Proof:

Define the functions $G(\mathcal{X}_n, \theta) = \mathbf{D}^T(\hat{\theta})\mathbf{V}_n(\mathbf{Y}_n - p(\theta))$. Let $G_\theta(\mathcal{X}_n, \theta) = -\mathbf{D}^T(\hat{\theta})\mathbf{V}_n\mathbf{D}(\theta)$ be the partial derivative of G with respect to θ. We expand $G(\mathcal{X}_n, \theta_0)$ about the point $\hat{\theta}$:

$$G(\mathcal{X}_n, \theta_0) = G(\mathcal{X}_n, \hat{\theta}) + [\int_0^1 G_\theta\{\mathcal{X}_n, \hat{\theta} + \lambda(\theta_0 - \hat{\theta})\}d\lambda](\theta_0 - \hat{\theta}), \quad (3.10)$$

where the integral of a matrix proceeds elementwise. According to the definition of $\hat{\theta}$ we have

$$\left.\frac{\partial F(\mathbf{Y}_n, p(\theta))}{\partial \theta^T}\right|_{\hat{\theta}} = -2\mathbf{D}^T(\hat{\theta})\mathbf{V}_n(\mathbf{Y}_n - p(\hat{\theta})) = 0$$

i.e. $G(\mathcal{X}_n, \hat{\theta}) = 0$. We multiply both sides of (3.10) by $\sqrt{n}$:

$$\sqrt{n}\mathbf{D}^T(\hat{\theta})\mathbf{V}_n(\mathbf{Y}_n - p(\theta_0)) = -\sqrt{n}[\int_0^1 G_\theta\{\mathcal{X}_n, \hat{\theta} + \lambda(\theta_0 - \hat{\theta})\}d\lambda](\hat{\theta} - \theta_0).$$

By Lemma 2, we know $\hat{\theta}$ converges to θ_0, thus for any λ,

$$G_\theta\{\mathcal{X}_n, \hat{\theta} + \lambda(\theta_0 - \hat{\theta})\} = -\mathbf{D}^T(\hat{\theta})\mathbf{V}_n\mathbf{D}(\hat{\theta} + \lambda(\theta_0 - \hat{\theta}))$$

converges to $\mathbf{D}_0^T\mathbf{V}\mathbf{D}_0$ in probability. Therefore, we have

$$\lim_{n\to\infty}\int_0^1 G_\theta\{\mathcal{X}_n, \hat{\theta}+\lambda(\theta_0-\hat{\theta})\}d\lambda = \int_0^1 \lim_{n\to\infty} G_\theta\{\mathcal{X}_n, \hat{\theta}+\lambda(\theta_0-\hat{\theta})\}d\lambda = \mathbf{D}_0\mathbf{V}\mathbf{D}_0^T$$

by bounded convergence theorem. By Slutsky's theorem and delta methods, we reach the conclusions. □

Remark:

Lemma 3 is similar to the modified χ^2 method (Ferguson 1958). However, the latter only deals with the cases where $\mathbf{V}_n$ is a function of the statistic

$\mathbf{Y}_n(\mathcal{X}_n)$. Lemma 3 generalized the result for all $\mathbf{V}_n$ that is a general function of the whole sample $\mathcal{X}_n$.

Lemma 3 shows that the asymptotic distribution of $p(\hat{\theta})$ is the same for different series of $\mathbf{V}_n$ as long as $\mathbf{V}_n$ converges to the same $\mathbf{V}$ in probability. It is easy to see that parameterization θ does not affect the asymptotic properties of $p(\hat{\theta})$. Thus, for simplicity, we can impose $\mathbf{V}_n = \mathbf{V}$ when considering asymptotic properties of $p(\hat{\theta})$, if we can show there exists one parameterization that satisfies the conditions in the statement of the lemma.

3.3.2 Proof of Theorem 2

In Theorem 2, we consider a discrepancy function:

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

We first address the issue that the inner product matrix in Proposition 1 is assumed to be known while the inner product matrix in F_d is estimated. Since $\mathbf{V}_n$ converges to $\mathbf{V}$ in probability, it follows from Lemma 2 that the asymptotic distribution of $n\hat{F}_d$ is the same as that of $n\hat{H}_d$, where

$$H_d(\mathbf{B}, \mathbf{C}) = (\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\mathbf{BC}))^T \mathbf{V} (\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\mathbf{BC})).$$

Furthermore, we want to show the asymptotic distribution of $\mathrm{vec}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\nu}})$ of $F_d(\mathbf{B}, \mathbf{C})$ is the same as that of $H_d(\mathbf{B}, \mathbf{C})$. Based on the remarks about Lemma 3, we only need to show there is one parameterization that satisfies the conditions in the statement of Lemma 3. We can use the parameterization discussed at the end of Section 2.1. Let $\boldsymbol{\beta} = (\mathbf{I}_d, \boldsymbol{\beta}^{*T})^T$. Therefore,

we have parameters: $\boldsymbol{\beta}^* \in \mathbb{R}^{(p-d)\times d}$ and $\boldsymbol{\nu} \in \mathbb{R}^{d\times(h-1)}$, and (vec($\boldsymbol{\beta}^*$), vec($\boldsymbol{\nu}$)) corresponds to the θ in Lemma 3. The new setting provides a full rank Jacobian matrix and an open parameter space in $\mathbb{R}^{d(h+p-d-1)}$, thus satisfying the conditions in Lemma 3. At same time, the reparameterization affects neither our algorithm for minimization nor asymptotic results. From now on, we only need to prove the conclusions for H_d.

$H_d(\mathbf{B}, \mathbf{C})$ is in the form of Shapiro's discrepancy function H. This can be seen by setting

$$\begin{aligned} \theta &= \begin{pmatrix} \text{vec}(\mathbf{B}) \\ \text{vec}(\mathbf{C}) \end{pmatrix} \in \mathbb{R}^{d(p+h-1)} \\ g(\theta) &= \text{vec}(\mathbf{BC}) \in \mathbb{R}^{p(h-1)} \\ \boldsymbol{\tau}_n &= \text{vec}(\hat{\boldsymbol{\zeta}}) \\ g(\theta_0) &= \text{vec}(\boldsymbol{\beta}\boldsymbol{\nu}) \end{aligned}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ is in general a basis for $\mathcal{S}_{\boldsymbol{\xi}}$ and $\boldsymbol{\nu} \in \mathbb{R}^{d\times(h-1)}$. With these associations we next verify that $\boldsymbol{\Delta}_{\boldsymbol{\zeta}} = (\boldsymbol{\nu}^T \otimes \mathbf{I}_p, \mathbf{I}_{h-1} \otimes \boldsymbol{\beta})$ as defined previously in (3.8). Let $\mathbf{B} = [\mathbf{b}_1, ..., \mathbf{b}_d]$, where $\mathbf{b}_j = (b_{1j}, ..., b_{pj})^T \in \mathbb{R}^p$, $j = 1, 2, ..., d$. Denote $\mathbf{C} = [\mathbf{C}_1, ..., \mathbf{C}_{h-1}]$, where $\mathbf{C}_k = (c_{1k}, c_{2k}, ..., , c_{dk})^T \in \mathbb{R}^d$, $k = 1, 2, ..., h-1$. Then,

$$g(\theta) = \text{vec}(\mathbf{BC}) = \text{vec}([\sum_{j=1}^{d} c_{j1}\mathbf{b}_j, \sum_{j=1}^{d} c_{j2}\mathbf{b}_j, ..., \sum_{j=1}^{d} c_{j(h-1)}\mathbf{b}_j]).$$

We have

$$\frac{\partial g(\theta)}{\partial b_{ij}} = (\underbrace{0, ..., 0, \overbrace{c_{j1}}^{\text{the ith element}}, 0, .., 0}_{\text{p elements}} | 0, ..., 0, \overbrace{c_{j2}}^{\text{the ith element}}, 0, .., 0|$$

$$\ldots|0,\ldots,0,\ \overbrace{c_{j(h-1)}}^{\text{the ith element}}\ ,0,..,0)^T$$

Therefore,

$$\frac{\partial g(\theta)}{\partial \operatorname{vec}(\mathbf{B})} = \begin{pmatrix} c_{11} & c_{21} & \cdots & c_{d1} \\ c_{12} & c_{22} & \cdots & c_{d2} \\ \vdots & \vdots & \ddots & \vdots \\ c_{1(h-1)} & c_{2(h-1)} & \cdots & c_{d(h-1)} \end{pmatrix} \otimes \mathbf{I}_p = \mathbf{C}^T \otimes \mathbf{I}_p.$$

We also have

$$\frac{\partial g(\theta)}{\partial c_{jk}} = \operatorname{vec}(\underbrace{0,\ldots,0,\ \overbrace{b_j}^{\text{the kth column}}\ ,0,..,0}_{\text{(h-1) columns}})$$

Therefore,

$$\frac{\partial g(\theta)}{\partial \operatorname{vec}(\mathbf{C})} = \begin{pmatrix} \mathbf{b}_1\mathbf{b}_2\ldots\mathbf{b}_d & & & \\ & \mathbf{b}_1\mathbf{b}_2\ldots\mathbf{b}_d & & \\ & & \ddots & \\ & & & \mathbf{b}_1\mathbf{b}_2\ldots\mathbf{b}_d \end{pmatrix} = \mathbf{I}_p \otimes \mathbf{B}.$$

Hence,

$$\frac{\partial g(\theta)}{\partial \theta} = \left[\mathbf{C}^T \otimes \mathbf{I}_p, \mathbf{I}_{h-1} \otimes \mathbf{B}\right]$$

Since $\mathbf{V} > 0$, conditions 2, including properties p1-p4, and 4 of Proposition 1 are met. Condition 3 is met also since $g(\theta)$ is analytic. See Shapiro (1986) for details about regular points. Since $\mathbf{V} = \mathbf{\Gamma}_{\hat{\zeta}}^{-1}$, based on Conclusion 3 and 4 in Proposition 1, $\operatorname{vec}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\nu}})$ of $H_d(\mathbf{B}, \mathbf{C})$ is asymptotically efficient with

$$\sqrt{n}(\operatorname{vec}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\nu}}) - \operatorname{vec}(\boldsymbol{\beta}\boldsymbol{\nu})) \xrightarrow{\mathcal{D}} \text{Normal}(0, \boldsymbol{\Delta}_{\zeta}(\boldsymbol{\Delta}_{\zeta}^T\mathbf{V}\boldsymbol{\Delta}_{\zeta})^{-}\boldsymbol{\Delta}_{\zeta}),$$

which leads to the conclusion number one. Meanwhile, $n\hat{H} \xrightarrow{\mathcal{D}} \chi^2_k$, where the degrees of freedom $k = p(h-1) - \text{rank}(\boldsymbol{\Delta}_\zeta)$. Since

$$\begin{aligned} \text{rank}(\boldsymbol{\Delta}_\zeta) &= \text{rank}([\boldsymbol{\nu}^T \otimes \mathbf{Q}_\beta, \mathbf{I}_{h-1} \otimes \boldsymbol{\beta}]) \\ &= d(p-d) + d(h-1) \\ &= d(h+p-d-1), \end{aligned}$$

we have $k = (p-d)(h-d-1)$. Thus, conclusion number two is proved. The consistency of $\text{Span}(\hat{\boldsymbol{\beta}})$ in conclusion number three follows directly from the conclusion number one.

3.4 Computation of Optimal IRE

To make Optimal IRE practical, we need solve two issues. The first issue is how to decide $\mathbf{V}_n$, which should be a consistent estimate of $\boldsymbol{\Gamma}_{\hat{\zeta}}^{-1}$. The second one is the minimization of the discrepancy function given $\mathbf{V}_n$. We address these two issues in turn.

Estimation of V

We know $\boldsymbol{\Gamma}_{\hat{\zeta}} = (\mathbf{A}^T \otimes \mathbf{I})\boldsymbol{\Gamma}^*(\mathbf{A} \otimes \mathbf{I})$, where $\boldsymbol{\Gamma}^* = \text{Cov}(\text{vec}(\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{Z}\boldsymbol{\varepsilon}^T)) \in \mathbb{R}^{ph \times ph}$ as shown in Theorem 1. Thus, since $\mathbf{A}$ is a constant matrix, to estimate $\mathbf{V} = \boldsymbol{\Gamma}_{\hat{\zeta}}^{-1}$, we only need plug in the estimate of $\boldsymbol{\Gamma}^*$ that is the sample version of $\text{Cov}(\text{vec}(\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{Z}\boldsymbol{\varepsilon}^T))$, which is easy to obtain in light of the fact $\text{E}[\text{vec}(\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{Z}\boldsymbol{\varepsilon}^T)] = 0$.

Minimization of F_d

The discrepancy function (3.3)

$$F_d(\mathbf{B},\mathbf{C}) \;=\; (\,\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\mathbf{BC}))^T\mathbf{V}_n(\,\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\mathbf{BC})),$$

can be minimized by treating it as a separable nonlinear least squares problem (see Ruhe and Wedin 1980). We have separate sets of parameters, $\mathbf{B}$ and $\mathbf{C}$. Given $\mathbf{B}$, the minimization with respect to $\mathbf{C}$ is straightforward: It is a linear regression of $\mathbf{V}_n^{\frac{1}{2}}\,\mathrm{vec}(\hat{\boldsymbol{\zeta}})$ on $\mathbf{V}_n^{\frac{1}{2}}(\mathbf{I}_{h-1}\otimes\mathbf{B})$ of which the coefficients are $\mathrm{vec}(\mathbf{C})$. On the other hand, consider minimization with respect to one column $\mathbf{b}_k$ of $\mathbf{B}$, given $\mathbf{C}$ and the remaining columns of $\mathbf{B}$ and subject to the length constraint $\|\mathbf{b}_k\| = 1$ and the orthogonality constraint $\mathbf{b}_k^T\mathbf{B}_{(-k)} = 0$, where $\mathbf{B}_{(-k)}$ is the matrix that is left after taking away $\mathbf{b}_k$ from $\mathbf{B}$. For this partial minimization problem, the discrepancy function can be re-expressed as

$$F^*(\mathbf{b}) \;=\; (\boldsymbol{\alpha}_k - (\mathbf{c}_k^T\otimes\mathbf{I}_p)\mathbf{Q}_{\mathbf{B}_{(-k)}}\mathbf{b})^T\mathbf{V}_n(\boldsymbol{\alpha}_k - (\mathbf{c}_k^T\otimes\mathbf{I}_p)\mathbf{Q}_{\mathbf{B}_{(-k)}}\mathbf{b}),$$

where $\boldsymbol{\alpha}_k = \mathrm{vec}(\hat{\boldsymbol{\zeta}} - \mathbf{B}_{(-k)}\mathbf{C}_{(-k)}) \in \mathbb{R}^{p(h-1)}$, c_k is the k-th row of $\mathbf{C}$, and $\mathbf{C}_{(-k)}$ consists of all but the k-th row of $\mathbf{C}$. This is a linear regression problem again.

Outline of Algorithm

We are now in a position to describe an algorithm for the minimization of (3.3). We call it the *alternating least squares method.* See Kiers (2002) for background on alternating least squares optimization algorithms. For a given dimension d, the algorithm searches over $\mathbb{R}^{p\times d}$ for $\mathbf{B}$ which minimizes (3.3). The outline is as follows.

1. Choose the initial $\mathbf{B} \leftarrow (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_d)$. Constant initial starting vectors are often a good choice. Our experience is that the initial values do not generally affect the ultimate result.

2. Calculate

$$\text{vec}(\mathbf{C}) = [(\mathbf{I}_{h-1} \otimes \mathbf{B}^T)\mathbf{V}_n(\mathbf{I}_{h-1} \otimes \mathbf{B})]^{-1}(\mathbf{I}_{h-1} \otimes \mathbf{B}^T)\mathbf{V}_n \,\text{vec}(\hat{\boldsymbol{\zeta}}).$$

   Assign $e_0 \leftarrow F_d(\mathbf{B}, \mathbf{C})$ and $iter \leftarrow 0$.

3. (a) For $k = 1, 2, \ldots, d$:

   - At the current step $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_{k-1}, \mathbf{b}_k, \mathbf{b}_{k+1}, \ldots, \mathbf{b}_d)$. Assign

$$\boldsymbol{\alpha}_k \leftarrow \text{vec}(\hat{\boldsymbol{\zeta}} - \mathbf{B}_{(-k)}\mathbf{C}_{(-k)})$$

     which is a residual vector with $\mathbf{b}_k$ excluded. Find a new $\mathbf{b}_k$ minimizing the function with the constraint that it is orthogonal to $\mathbf{B}_{(-k)}$ and has length 1:

$$\hat{\mathbf{b}}_k = \mathbf{Q}_{\mathbf{B}_{(-k)}}[\mathbf{Q}_{\mathbf{B}_{(-k)}}(c_k^T \otimes \mathbf{I}_p)\mathbf{V}_n(c_k \otimes \mathbf{I}_p)\mathbf{Q}_{\mathbf{B}_{(-k)}}]^{-}(c_k^T \otimes \mathbf{I}_p)\mathbf{Q}_{\mathbf{B}_{(-k)}}\mathbf{V}_n\boldsymbol{\alpha}_k.$$

   - Update

$$\mathbf{B} \leftarrow (\mathbf{b}_1, \ldots, \mathbf{b}_{k-1}, \hat{\mathbf{b}}_k, \mathbf{b}_{k+1}, \ldots, \mathbf{b}_d)$$

$$\mathbf{C} \leftarrow \arg_{\mathbf{C}^*} \min F_d(\mathbf{B}, \mathbf{C}^*)$$

   (b) $e_1 \leftarrow F_d(\mathbf{B}, \mathbf{C})$ and $iter \leftarrow iter + 1$.

4. Return to step 3 until e_1 no longer changes and then assign $\widetilde{\mathbf{B}} \leftarrow \mathbf{B}$ and exit.

At termination, $\widetilde{\mathbf{B}}$ is an estimate of $\boldsymbol{\beta}$. After one iteration of step 3, the algorithm produces a monotonically decreasing series of evaluations and thus is guaranteed to converge because $F_d \geq 0$. When the computing time is an un-negligible issue, we may put an upper bound on the number of the iteration and stop the iteration when the difference between two consecutive evaluations of e_1 is smaller than a pre-specified number.

As we shall see in SIR in Chapter 5 and WCT in Chapter 6, estimated basis directions are ordered by the eigenvalues of the sample kernel matrix. This algorithm will not necessarily produce an analogous ordering. However, we can construct an ordered basis for Span$\{\widetilde{\mathbf{B}}\}$ with respect to the amount by which directions decrease $F_d(\mathbf{B}, \mathbf{C})$. For example, the most important direction is

$$\hat{\mathbf{b}}_1 = \arg_{\mathbf{b}} \min(\operatorname{vec}(\hat{\boldsymbol{\zeta}}) - \operatorname{vec}(\mathbf{bC}))^T \mathbf{V}_n (\operatorname{vec}(\hat{\boldsymbol{\zeta}}) - \operatorname{vec}(\mathbf{bC})),$$

where $\mathbf{C} \in \mathbb{R}^{1\times(h-1)}$, and the minimization is over $\mathbf{b} \in$ Span$\{\widetilde{\mathbf{B}}\}$ with $\|\mathbf{b}\| = 1$. The second direction is

$$\hat{\mathbf{b}}_2 = \arg_{\mathbf{b}} \min(\operatorname{vec}(\hat{\boldsymbol{\zeta}}) - \operatorname{vec}([\hat{\mathbf{b}}_1 \mathbf{b}]\mathbf{C}))^T \mathbf{V}_n (\operatorname{vec}(\hat{\boldsymbol{\zeta}}) - \operatorname{vec}([\hat{\mathbf{b}}_1 \mathbf{b}]\mathbf{C})),$$

where $\mathbf{C} \in \mathbb{R}^{2\times(h-1)}$, and the minimization is over $\mathbf{b} \in$ Span$\{\widetilde{\mathbf{B}}\}$ with $\|\mathbf{b}\| = 1$ and $\mathbf{b}^T \hat{\mathbf{b}}_1 = 0$. And so on.

The alternating least squares method of Optimal IRE utilizes the special features of the objective function, therefore it is more efficient than a general optimization algorithm. In the computation we need to find $\mathbf{V}_n^{-1}$ which is a $p(h-1) \times p(h-1)$ matrix. Usually this is not a big issue. However, then

p and h are large it may bring some computational difficulties. Later on we propose a simplified version of inverse regression estimation, which we call *simple inverse regression estimate* (Simple IRE). However the asymptotic distribution of test statistic in Simple IRE is more complicated than a chi-squared distribution. A simulation study comparing SIR, WCT, Optimal IRE, and Simple IRE is included in Chapter 8.

Chapter 4

Sub-Optimal Inverse Regression Estimation

In Chapter 3, we discussed Optimal IRE, which takes into accounts two important issues: the intrinsic location constraint (cf. (2.6)) and using the inverse of covariance of the limiting distribution as the inner product matrix. In this chapter, we consider a sub-optimal class that does not acknowledge either issue. Two consequences are associated with this negligence: the estimate of a basis of the CS may not be asymptotically efficient, and the test statistic for dimension is not a chi-squared distribution under the null hypothesis, but a more complicated linear combination of chi-squares. Since efficient methods have been discussed previously and in practice we are more concerned with the estimation of dimension, for this sub-optimal class we focus on the asymptotic distributions of test statistics.

Suppose we examine all columns of $\hat{\boldsymbol{\xi}}$ (cf. (2.7)) and we use a positive

definite block diagonal matrix as the inner product matrix in the discrepancy function:

$$\begin{aligned} F_d(\mathbf{B},\mathbf{C}) &= (\,\mathrm{vec}(\hat{\boldsymbol{\xi}}) - \mathrm{vec}(\mathbf{BC}))^T \mathbf{V}_n (\,\mathrm{vec}(\hat{\boldsymbol{\xi}}) - \mathrm{vec}(\mathbf{BC})) \\ &= \sum_y (\hat{\boldsymbol{\xi}}_y - \mathbf{BC}_y)^T \mathbf{V}_{ny} (\hat{\boldsymbol{\xi}}_y - \mathbf{BC}_y) \end{aligned} \tag{4.1}$$

where $\mathbf{C}_y \in \mathbb{R}^d$. The function (4.1) is a restatement of (2.8). We call the methods using (4.1) the *sub-optimal class* in the MDA family. It turns out that SIR is a member of this class. Later on, we will see that *simple inverse regression estimation* (Simple IRE)—another member in this sub-optimal class—can perform better than SIR and WCT. When $\mathbf{V}_n$ converges to a positive definite matrix, the value of $\mathbf{B}$ that minimizes the function still provides a consistent estimate of the CS as we shall see. In this chapter, we address general asymptotic properties of test statistics for any $\mathbf{V}_n > 0$ that converges to $\mathbf{V} > 0$. In subsequent chapters, for different choices of $\mathbf{V}_n$, we investigate minimization algorithms and specific asymptotic distributions of test statistics. Under different assumptions, SIR and WCT use the same block diagonal matrix as $\mathbf{V}_n$, where the blocks only differ by a scalar which makes the minimization reduce to a spectral decomposition. Simple IRE uses another block diagonal matrix which brings a better performance but at same time requires a more complicated algorithm.

4.1 Asymptotic Distribution of $n\hat{F}_d$

To report the asymptotic distribution of $n\hat{F}_d$ of (4.1), we need the $ph\times(p+h)d$ matrix

$$\boldsymbol{\Delta}_{\boldsymbol{\xi}} \equiv (\boldsymbol{\gamma}^T \otimes I_p, I_h \otimes \boldsymbol{\beta}). \tag{4.2}$$

The matrix $\boldsymbol{\Delta}_{\boldsymbol{\xi}}$ is the Jacobian matrix for the discrepancy function of (4.1)

$$\boldsymbol{\Delta} = \left(\frac{\partial \operatorname{vec}(\mathbf{BC})}{\partial \operatorname{vec}(\mathbf{B})}, \frac{\partial \operatorname{vec}(\mathbf{BC})}{\partial \operatorname{vec}(\mathbf{C})}\right)$$

evaluated at $(\boldsymbol{\beta}, \boldsymbol{\gamma})$, where $\mathbf{B} \in \mathbb{R}^{p\times d}$, $\mathbf{C} \in \mathbb{R}^{d\times h}$, and $\boldsymbol{\beta}$ along with $\boldsymbol{\gamma}$ are as defined previously in Section 2.1. Let $\mathbf{V}$ be the limit of $\mathbf{V}_n$. Define

$$\tilde{\boldsymbol{\xi}}_y \equiv \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\xi}}_y = \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}}_{y\bullet} - \bar{\mathbf{X}}_{\bullet\bullet}),$$

and

$$\tilde{\boldsymbol{\xi}} \equiv \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\xi}} = (\tilde{\boldsymbol{\xi}}_1, \ldots, \tilde{\boldsymbol{\xi}}_h), \tag{4.3}$$

and define

$$\boldsymbol{\Gamma}_{\tilde{\boldsymbol{\xi}}} \equiv (\mathbf{D}_{\mathbf{g}}^{-1}\mathbf{Q}_{\mathbf{g}} \otimes \boldsymbol{\Sigma}^{-1})\operatorname{diag}\{\boldsymbol{\Sigma}_{\mathbf{X}|y}\}(\mathbf{Q}_{\mathbf{g}}\mathbf{D}_{\mathbf{g}}^{-1} \otimes \boldsymbol{\Sigma}^{-1}), \tag{4.4}$$

where $\mathbf{D}_{\mathbf{g}}$ is a diagonal matrix with the elements of $\mathbf{g}$ on the diagonal, and $\mathbf{Q}_{\mathbf{g}}$ is the projection onto the orthogonal complement of $\operatorname{Span}(\mathbf{g})$. Finally, letting $\boldsymbol{\Phi} = \mathbf{V}^{\frac{1}{2}}\boldsymbol{\Delta}_{\boldsymbol{\xi}}$ and $\boldsymbol{\Omega} = \mathbf{V}^{\frac{1}{2}}\boldsymbol{\Gamma}_{\tilde{\boldsymbol{\xi}}}\mathbf{V}^{\frac{1}{2}}$, the asymptotic distribution of $n\hat{F}_d$ is given in the following theorem.

Theorem 3. *Assume that the data* $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, *are a simple random sample of* $(\mathbf{X}, Y)$. *Let* $\mathcal{S}_{\boldsymbol{\xi}} = \oplus_{y=1}^{h}\operatorname{Span}\{\boldsymbol{\xi}_y\}$, *let* $d = \dim(\mathcal{S}_{\boldsymbol{\xi}})$ *and let* $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \arg_{\mathbf{B},\mathbf{C}} \min F_d(\mathbf{B}, \mathbf{C})$ *as defined previously in (4.1). Then*

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

2. *as* $n \to \infty$,

$$n\hat{F}_d \xrightarrow{\mathcal{D}} \sum_{i=1}^{ph} \lambda_i \chi_i^2(1)$$

where $\{\chi_i^2(1)\}$ *are independent chi-squared random variables each with 1 degree of freedom and* $\{\lambda_1 \geq \ldots \geq \lambda_{ph}\}$ *are the eigenvalues of* $\mathbf{Q}_{\boldsymbol{\Phi}}\boldsymbol{\Omega}\mathbf{Q}_{\boldsymbol{\Phi}}$.

Similar to Theorem 2, this theorem is quite general, requiring none of the special conditions discussed previously. Also we should notice that Theorem 3 is valid for a general $\mathbf{V}_n$, of which the block diagonal $\mathbf{V}_n$ used in the sub-optimal class is a special case. The value $\hat{\boldsymbol{\beta}}$ of $\mathbf{B}$ that minimizes the discrepancy function $F_d(\mathbf{B}, \mathbf{C})$ always provides a consistent estimate of a basis for $\mathcal{S}_{\boldsymbol{\xi}}$, and this theorem allows us to test hypotheses about its dimension. With some of the special conditions, $\mathcal{S}_{\boldsymbol{\xi}}$ will be a subspace of the CS or CS itself. The proof of Theorem 3 is given in Section 4.3. We next summarize the computations necessary to implement the tests available as a result of Theorem 3.

4.2 Computations

To use Theorem 3 in practice, we need to replace $\mathbf{Q}_{\boldsymbol{\Phi}}\boldsymbol{\Omega}\mathbf{Q}_{\boldsymbol{\Phi}}$ with a consistent estimate under the null hypothesis. Under the hypothesis $d = m$, the $ph \times (p+h)m$ Jacobian matrix $\boldsymbol{\Delta}_{\boldsymbol{\xi}}$ can be estimated consistently by substituting the corresponding estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$: $\hat{\boldsymbol{\Delta}}_{\boldsymbol{\xi}} = (\hat{\boldsymbol{\gamma}}^T \otimes I_p, I_h \otimes \hat{\boldsymbol{\beta}})$. To estimate $\mathbf{V}$ we use $\mathbf{V}_n$. We also can estimate $\boldsymbol{\Gamma}_{\tilde{\boldsymbol{\xi}}}$ with

$$\hat{\boldsymbol{\Gamma}}_{\tilde{\boldsymbol{\xi}}} \equiv (\mathbf{D}_{\hat{\mathbf{g}}}^{-1}\mathbf{Q}_{\hat{\mathbf{g}}} \otimes \hat{\boldsymbol{\Sigma}}^{-1})\mathrm{diag}\{\hat{\boldsymbol{\Sigma}}_{\mathbf{X}|y}\}(\mathbf{Q}_{\hat{\mathbf{g}}}\mathbf{D}_{\hat{\mathbf{g}}}^{-1} \otimes \hat{\boldsymbol{\Sigma}}^{-1}),$$

where $\hat{\mathbf{g}} = \sqrt{\hat{\mathbf{f}}}$, $\hat{\mathbf{f}} = (\frac{n_1}{n}, \ldots, \frac{n_h}{n})^T$, $\hat{\mathbf{\Sigma}}$ is the sample covariance matrix, and $\hat{\mathbf{\Sigma}}_{\mathbf{X}|y}$ is the sample covariance matrix for the yth slice. These estimates are then substituted to yield an estimate of $\mathbf{Q}_{\mathbf{\Phi}}\mathbf{\Omega}\mathbf{Q}_{\mathbf{\Phi}}$ from which sample eigenvalues $\hat{\lambda}_j$ are obtained. The statistic $n\hat{F}_m$ is then compared to the percentage points of the distribution of

$$\sum_{i=1}^{ph} \hat{\lambda}_i \chi_i^2(1)$$

to obtain a p-value. There is a substantial literature on computing tail probabilities of the distribution of a linear combination of chi-squared random variables. See Field (1993) for an introduction. Alternatively, the tail areas can usually be approximated adequately by using Satterthwaite's approximation.

4.3 Proof of Theorem 3

Theorem 3 considers the discrepancy function

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

where $\mathbf{V}_n \in \mathbb{R}^{ph \times ph}$, a positive definite block diagonal matrix. The function also can be written as

$$F_d(\mathbf{B}, \mathbf{C}) = (\,\mathrm{vec}(\tilde{\boldsymbol{\xi}}) - \mathrm{vec}(\widetilde{\mathbf{B}}\mathbf{C}))^T \widetilde{\mathbf{V}}_n (\,\mathrm{vec}(\tilde{\boldsymbol{\xi}}) - \mathrm{vec}(\widetilde{\mathbf{B}}\mathbf{C}))$$

where $\widetilde{\mathbf{B}} = \mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}}\mathbf{B}$, $\widetilde{\mathbf{V}}_n = (I \otimes \mathbf{\Sigma}\hat{\mathbf{\Sigma}}^{-1})\mathbf{V}_n(I \otimes \hat{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma})$, and $\tilde{\boldsymbol{\xi}}$ is defined in (4.3). Since $\widetilde{\mathbf{V}}_n$ also converges to $\mathbf{V}$ as $\mathbf{V}_n$ does, it follows from Lemma 2 in Chapter 3 that the asymptotic distribution of $n\hat{F}_d$ is the same as that of

$n\hat{H}_d$, where

$$H_d(\mathbf{B},\mathbf{C}) = (\,\mathrm{vec}(\tilde{\boldsymbol{\xi}}) - \mathrm{vec}(\mathbf{BC}))^T\mathbf{V}(\,\mathrm{vec}(\tilde{\boldsymbol{\xi}}) - \mathrm{vec}(\mathbf{BC})).$$

This H_d is a version of Shaprio's discrepancy function H, which can now be seen by setting

$$\begin{aligned}
\theta &= \begin{pmatrix} \mathrm{vec}(\mathbf{B}) \\ \mathrm{vec}(\mathbf{C}) \end{pmatrix} \in \mathbb{R}^{d(p+h)} \\
g(\theta) &= \mathrm{vec}(\mathbf{BC}) \in \mathbb{R}^{ph} \\
\boldsymbol{\tau}_n &= \mathrm{vec}(\tilde{\boldsymbol{\xi}}) \\
g(\theta_0) &= \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\gamma}).
\end{aligned}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ is in general a basis for $\mathcal{S}_{\boldsymbol{\xi}}$ and $\boldsymbol{\gamma} \in \mathbb{R}^{d\times h}$. With these associations it is straightforward to verify that $\boldsymbol{\Delta}_{\boldsymbol{\xi}} = (\boldsymbol{\gamma}^T \otimes I_p, I_h \otimes \boldsymbol{\beta})$ as defined previously in (4.2). Since $\mathbf{V} > 0$, conditions 2, including properties p1-p4, and 4 of Proposition 1 are met. Condition 3 is met also since $g(\theta)$ is analytic. See Shapiro (1986) for details about regular points. With this, we have verified all of the conditions of Proposition 1, except for asymptotic normality.

It is easy to see that $\mathrm{E}[\tilde{\boldsymbol{\xi}}] = \boldsymbol{\beta}\boldsymbol{\gamma}$ regardless of $\hat{\mathbf{g}}$. Thus,

$$\begin{aligned}
\mathrm{Cov}(\,\mathrm{vec}(\tilde{\boldsymbol{\xi}})) &= \mathrm{E}[\mathrm{Cov}(\,\mathrm{vec}(\tilde{\boldsymbol{\xi}})|\hat{\mathbf{g}})] \\
&= \frac{1}{n}\mathrm{E}[(\mathbf{D}_{\hat{\mathbf{g}}}^{-1}\mathbf{Q}_{\hat{\mathbf{g}}} \otimes \boldsymbol{\Sigma}^{-1})\mathrm{diag}\{\boldsymbol{\Sigma}_{\mathbf{X}|y}\}(\mathbf{Q}_{\hat{\mathbf{g}}}\mathbf{D}_{\hat{\mathbf{g}}}^{-1} \otimes \boldsymbol{\Sigma}^{-1})] \\
&= \frac{1}{n}(\mathbf{D}_{\mathbf{g}}^{-1}\mathbf{Q}_{\mathbf{g}} \otimes \boldsymbol{\Sigma}^{-1})\mathrm{diag}\{\boldsymbol{\Sigma}_{\mathbf{X}|y}\}(\mathbf{Q}_{\mathbf{g}}\mathbf{D}_{\mathbf{g}}^{-1} \otimes \boldsymbol{\Sigma}^{-1}) + o(\frac{1}{n}).
\end{aligned}$$

Therefore,

$$\sqrt{n}(\,\mathrm{vec}(\tilde{\boldsymbol{\xi}}) - \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\gamma})) \xrightarrow{\mathcal{D}} \mathrm{Normal}(0, \boldsymbol{\Gamma}_{\tilde{\boldsymbol{\xi}}}).$$

It now follows from conclusion number one of Proposition 1 in Section 3.3.1 that the asymptotic distribution of $n\hat{F}_d$ is the same as that of $\|\mathbf{Q}_{\mathbf{\Phi}}\mathbf{V}^{\frac{1}{2}}\mathbf{W}\|^2$ where $\mathbf{W}$ is normal with mean 0 and covariance matrix $\mathbf{\Gamma}_{\tilde{\boldsymbol{\xi}}}$, and $\mathbf{\Phi} = \mathbf{V}^{\frac{1}{2}}\mathbf{\Delta}_{\boldsymbol{\xi}}$ as defined for the statement of Theorem 3. Consequently, $n\hat{F}_d$ is asymptotically distributed as a linear combination of independent chi-squared random variables each with one degree of freedom. The coefficients of the chi-squared variables are the eigenvalues of $\mathbf{Q}_{\mathbf{\Phi}}\mathbf{\Omega}\mathbf{Q}_{\mathbf{\Phi}}$, where $\mathbf{\Omega} = \mathbf{V}^{\frac{1}{2}}\mathbf{\Gamma}_{\tilde{\boldsymbol{\xi}}}\mathbf{V}^{\frac{1}{2}}$ is as defined for the statement of the theorem. Finally, consistency follows from conclusion number 3 of Proposition 1 in combination with Lemma 2 in Chapter 3.

Chapter 5

Sliced Inverse Regression

In this chapter, we review one popular dimension reduction method—sliced inverse regression (SIR). Then, we rederive it using the minimum discrepancy approach. It is easy to see that SIR belongs to the sub-optimal class we discussed in Chapter 4. Based on the asymptotic theory of the minimum discrepancy approach, we can simplify the derivation of the asymptotic distributions of the test statistics used by SIR. Furthermore, we set the stage for developing new methods with better performance in this sub-optimal class.

5.1 Review of SIR

We have seen that if the standardized predictor $\mathbf{Z}$ satisfies the linearity condition:

$$\mathrm{E}[\mathbf{Z}|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{Z}}}\mathbf{Z}] = \mathbf{P}_{\mathcal{S}_{Y|\mathbf{Z}}}\mathbf{Z},$$

then $\mathrm{Span}\{\mathrm{E}[\mathbf{Z}|Y]\} \subseteq \mathcal{S}_{Y|\mathbf{Z}}$. Assuming the linearity and coverage conditions, Li (1991) proposed SIR, which constructs a kernel matrix that is a sample

version of $\mathrm{Cov}(\mathrm{E}[\mathbf{Z}|Y])$. The eigenvalues of the SIR kernel matrix are used to construct a test statistic for dimension. Given the dimension, the eigenvectors corresponding to the largest eigenvalues are used to estimate a basis of the CS. The test for dimension Li proposed is based on normality of the predictors. Bura and Cook (2001b) removed this assumption and proposed a *weighted chi-square test* (WCT) for predictors with any general distribution as long as the linearity condition is satisfied. This WCT will be discussed in detail in Chapter 6.

The SIR procedure is as follows:

1. Standardize $\mathbf{X}$ to $\hat{\mathbf{Z}} = \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}(\mathbf{X} - \bar{\mathbf{X}}_{\bullet\bullet})$.

2. Construct a $p \times p$ kernel matrix

$$\widehat{\mathbf{M}}_{\mathrm{SIR}} = \sum_{y=1}^{h} \hat{f}_y \bar{\mathbf{Z}}_{y\bullet} \bar{\mathbf{Z}}_{y\bullet}^T, \tag{5.1}$$

   where $\bar{\mathbf{Z}}_{y\bullet}$ is the average of $\hat{\mathbf{Z}}$ in the yth slice.

3. Construct a spectral decomposition of $\widehat{\mathbf{M}}_{\mathrm{SIR}}$. Suppose $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, ..., \hat{\boldsymbol{\mu}}_p$ are the eigenvectors of $\widehat{\mathbf{M}}_{\mathrm{SIR}}$ corresponding to its eigenvalues $\hat{\lambda}_1 \geq ... \geq \hat{\lambda}_p \geq 0$.

4. If we assume $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$ $(d < \min\{p, h\})$, then $\hat{\boldsymbol{\beta}}_j = \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \hat{\boldsymbol{\mu}}_j$, $j = 1, 2, ..., d$. is the estimated basis of $\mathcal{S}_{Y|\mathbf{X}}$.

To estimate d, test statistics were proposed in the form of $\Lambda_m = n \sum_{i=m+1}^{p} \hat{\lambda}_i$. Suppose the dimension of the central space is d, under the assumption of normality of $\mathbf{X}$, Λ_d has an asymptotically chi-squared distribution with $(p -$

$d)(h-d-1)$ degrees of freedom. Begin with $m=0$ and compare Λ_m with the corresponding chi-squared distribution. If Λ_m is large, conclude that $d > m$ and increment m by 1 until the test statistic is relatively small. The estimation of d is then the terminal value of m.

5.2 SIR in Minimum Discrepancy Approach

In this section, we rederive SIR as a special case of the minimum discrepancy approach and investigate its asymptotic properties. First, let us consider an objective function:

$$\begin{aligned} F_d(\mathbf{B},\mathbf{C}) &= \sum_{y=1}^{h} \hat{f}_y(\hat{\boldsymbol{\xi}}_y - \mathbf{B}\mathbf{C}_y)^T\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\xi}}_y - \mathbf{B}\mathbf{C}_y) \qquad (5.2) \\ &= \sum_{y=1}^{h}(\sqrt{\hat{f}_y}\bar{\mathbf{Z}}_{y\bullet} - \sqrt{\hat{f}_y}\hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\mathbf{B}\mathbf{C}_y)^T(\sqrt{\hat{f}_y}\bar{\mathbf{Z}}_{y\bullet} - \sqrt{\hat{f}_y}\hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\mathbf{B}\mathbf{C}_y), \end{aligned}$$

where $\bar{\mathbf{Z}}_{y\bullet}$ is the average of $\hat{\mathbf{Z}}$ in the yth slice. Based on Lemma 5 in the Appendix, Span$(\hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\hat{\mathbf{B}})$ is the space spanned by the d eigenvectors corresponding to $\widehat{\mathrm{M}}_{\mathrm{SIR}}$'s d largest eigenvalues, where $\hat{\mathbf{B}}$ is the value of $\mathbf{B}$ that minimize (5.2). Thus, the estimate of $\boldsymbol{\beta}$, $\hat{\mathbf{B}}$, is $\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}[\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, ..., \hat{\boldsymbol{\mu}}_d]$. It is clear that SIR is a sub-optimal class member with $\mathbf{V}_n = \mathrm{diag}\{\hat{f}_y\hat{\boldsymbol{\Sigma}}\} = (\mathbf{D}_{\hat{\mathbf{f}}} \otimes \hat{\boldsymbol{\Sigma}})$. Since the diagonal blocks in $\mathbf{V}_n$ only differ by a scalar, the minimization of (5.2) reduces to a spectral decomposition problem. We can easily see that this *spectral decomposition approach* is a special case of the *minimum discrepancy approach.*

5.3 Test Statistic for Dimensionality

We note that based on Lemma 5 in the Appendix the test statistic Λ_m proposed by Li (1991) is the same as $n\hat{F}_m$ of (5.2). Without special conditions, $n\hat{F}_d$ generally has an asymptotic distribution of a linear combination of independent chi-squares with one degree of freedom based on Theorem 3. Li (1991) showed that under normality of $\mathbf{X}$, it has an asymptotic chi-squared distribution with degrees of freedom $(p-d)(h-d-1)$. Cook (1998) proved that a weaker condition—*marginal covariance condition*—suffices for Λ_d to converge to a random variable with a chi-squared distribution. Here we restate Cook's result followed by a new proof via the minimum discrepancy approach.

Corollary 1. *(Cook 1998, Prop. 11.5) Assume*

1. *The linearity condition:* $\mathrm{E}[\mathbf{Z}|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{Z}}}\mathbf{Z}] = \mathbf{P}_{\mathcal{S}_{Y|\mathbf{Z}}}\mathbf{Z}$.
2. *The marginal covariance condition:* $\mathrm{Cov}[\mathbf{Z}|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{Z}}}\mathbf{Z}] = \mathbf{Q}_{\mathcal{S}_{Y|\mathbf{Z}}}$.
3. *The coverage condition:* $\mathcal{S}_{\boldsymbol{\xi}} = \mathcal{S}_{Y|\mathbf{X}}$.

Suppose $\dim(\mathcal{S}_{Y|\mathbf{X}}) = d$. Λ_m *is the test statistic in SIR for the null hypothesis* $d = m$. *Then,*

$$\Lambda_d \xrightarrow{\mathcal{D}} \chi^2_{(p-d)(h-d-1)}, \quad \text{as } n \to \infty.$$

Meanwhile, the estimate of $\mathcal{S}_{Y|\mathbf{X}}$ *is consistent.*

Proof of Corollary 1

Our proof of Corollary 1 involves a fair amount of algebra. From the proof of Theorem 3 in Section 4.3 we have

$$\mathbf{\Gamma}_{\tilde{\boldsymbol{\xi}}} = (\mathbf{D}_{\mathbf{g}}^{-1}\mathbf{Q}_{\mathbf{g}} \otimes \mathbf{\Sigma}^{-1})\text{diag}\{\mathbf{\Sigma}_{\mathbf{X}|y}\}(\mathbf{Q}_{\mathbf{g}}\mathbf{D}_{\mathbf{g}}^{-1} \otimes \mathbf{\Sigma}^{-1})$$

and

$$\mathbf{\Delta}_{\boldsymbol{\xi}} = (\boldsymbol{\gamma}^T \otimes \mathbf{I}_p, \mathbf{I}_h \otimes \boldsymbol{\beta}).$$

Here $\mathbf{V} = (\mathbf{D}_{\mathbf{f}} \otimes \mathbf{\Sigma})$. The proof involves using the various conditions of the corollary to verifying algebraically that $\mathbf{\Gamma}_{\tilde{\boldsymbol{\xi}}}\mathbf{U}\mathbf{\Gamma}_{\tilde{\boldsymbol{\xi}}}\mathbf{U}\mathbf{\Gamma}_{\tilde{\boldsymbol{\xi}}} = \mathbf{\Gamma}_{\tilde{\boldsymbol{\xi}}}\mathbf{U}\mathbf{\Gamma}_{\tilde{\boldsymbol{\xi}}}$, where $\mathbf{U} = \mathbf{V}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{\Phi}}\mathbf{V}^{\frac{1}{2}}$, $\mathbf{\Phi} = \mathbf{V}^{\frac{1}{2}}\mathbf{\Delta}_{\boldsymbol{\xi}}$. Proposition 1 then implies that the limiting distribution is chi-squared. The conditions of Corollary 1 allow us to use the following lemma in its proof.

Lemma 4. *(Cook 1998, Prop. 10.2) Assume that the (1) linearity condition, (2) marginal covariance condition, and (3) coverage condition are satisfied. Then, for each value y of Y*

$$\text{Span}(\mathbf{I} - \mathbf{\Sigma}_{\mathbf{Z}|y}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}.$$

In another words, $\text{Span}(\mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{\mathbf{X}|y}\mathbf{\Sigma}^{-1}) \subseteq \mathcal{S}_{Y|\mathbf{X}}$.

For notational simplicity, we drop the subscripts on $\mathbf{\Gamma}_{\tilde{\boldsymbol{\xi}}}$ and $\mathbf{\Delta}_{\boldsymbol{\xi}}$ in the rest of the proof. Now,

$$\begin{aligned}\mathbf{\Gamma}\mathbf{U}\mathbf{\Gamma}\mathbf{U}\mathbf{\Gamma} - \mathbf{\Gamma}\mathbf{U}\mathbf{\Gamma} &= \mathbf{\Gamma}\mathbf{V}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{\Phi}}\mathbf{V}^{\frac{1}{2}}\mathbf{\Gamma}\mathbf{V}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{\Phi}}\mathbf{V}^{\frac{1}{2}}\mathbf{\Gamma} - \mathbf{\Gamma}\mathbf{V}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{\Phi}}\mathbf{V}^{\frac{1}{2}}\mathbf{\Gamma} \\ &= \mathbf{\Gamma}\mathbf{V}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{\Phi}}(\mathbf{\Omega} - \mathbf{I}_{ph})\mathbf{Q}_{\mathbf{\Phi}}\mathbf{V}^{\frac{1}{2}}\mathbf{\Gamma},\end{aligned}$$

where $\mathbf{\Omega} = \mathbf{V}^{\frac{1}{2}}\mathbf{\Gamma}\mathbf{V}^{\frac{1}{2}}$, and

$$\mathbf{\Phi} = \mathbf{V}^{\frac{1}{2}}\mathbf{\Delta} = (\mathbf{D}_{\mathbf{g}} \otimes \mathbf{\Sigma}^{\frac{1}{2}})[\boldsymbol{\gamma}^T \otimes \mathbf{I}_p, \mathbf{I}_h \otimes \boldsymbol{\beta}] = (\mathbf{D}_{\mathbf{g}}\boldsymbol{\gamma}^T \otimes \mathbf{\Sigma}^{\frac{1}{2}}, \mathbf{D}_{\mathbf{g}} \otimes \mathbf{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}).$$

Based on the properties of projection number 1 and 4 introduced in Appendix,

$$\begin{aligned}
\mathbf{Q}_{\mathbf{\Phi}} &= \mathbf{Q}_{[\mathbf{D_g}\boldsymbol{\gamma}^T \otimes \mathbf{Q}_{\mathbf{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}}\mathbf{\Sigma}^{\frac{1}{2}}]}\mathbf{Q}_{[\mathbf{D_g}\otimes\mathbf{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}]} \\
&= \mathbf{Q}_{[\mathbf{D_g}\boldsymbol{\gamma}^T \otimes \mathbf{Q}_{\mathbf{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}}\mathbf{\Sigma}^{\frac{1}{2}}]}(\mathbf{I}_h \otimes \mathbf{Q}_{\mathbf{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}}) \\
&= (\mathbf{P}_{\mathbf{D_g}\boldsymbol{\gamma}^T} \otimes \mathbf{Q}_{\mathbf{Q}_{\mathbf{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}}\mathbf{\Sigma}^{\frac{1}{2}}} + \mathbf{Q}_{\mathbf{D_g}\boldsymbol{\gamma}^T} \otimes \mathbf{I}_p)(\mathbf{I}_h \otimes \mathbf{Q}_{\mathbf{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}}) \\
&= \mathbf{Q}_{\mathbf{D_g}\boldsymbol{\gamma}^T} \otimes \mathbf{Q}_{\mathbf{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}}.
\end{aligned}$$

We have

$$\begin{aligned}
\mathbf{\Omega} - \mathbf{I}_{ph} &= (\mathbf{Q_g} \otimes \mathbf{I}_p)\mathrm{diag}\{\mathbf{\Sigma}_{\mathbf{Z}|y}\}(\mathbf{Q_g} \otimes \mathbf{I}_p) - \mathbf{I}_{ph} \\
&= (\mathbf{Q_g} \otimes \mathbf{I}_p)\mathrm{diag}\{\mathbf{\Sigma}_{\mathbf{Z}|y} - \mathbf{I}_p\}(\mathbf{Q_g} \otimes \mathbf{I}_p) - (\mathbf{P_g} \otimes \mathbf{I}_p), \quad (5.3)
\end{aligned}$$

where the diagonal matrix is over the values of Y. By Lemma 4, $\mathrm{Span}(\mathbf{\Sigma}_{\mathbf{Z}|y} - \mathbf{I}_p) \subseteq \mathrm{Span}(\mathbf{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta})$ and consequently

$$\mathrm{Span}(\mathrm{diag}\{\mathbf{\Sigma}_{\mathbf{Z}|y} - \mathbf{I}_p\}) \subseteq \mathrm{Span}(\mathbf{I}_h \otimes \mathbf{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta})$$

and

$$\mathrm{Span}((\mathbf{Q_g} \otimes \mathbf{I}_p)\mathrm{diag}\{\mathbf{\Sigma}_{\mathbf{Z}|y} - \mathbf{I}_p\}) \subseteq \mathrm{Span}(\mathbf{Q_g} \otimes \mathbf{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}) \subseteq \mathrm{Span}(\mathbf{D_g} \otimes \mathbf{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}).$$

Thus the first term of $\mathbf{\Omega} - \mathbf{I}_{ph}$ is in $\mathrm{Span}(\mathbf{\Phi})$. It follows from (5.3) that

$$\mathbf{Q}_{\mathbf{\Phi}}(\mathbf{\Omega} - \mathbf{I}_{ph}) = -\mathbf{Q}_{\mathbf{\Phi}}(\mathbf{P_g} \otimes \mathbf{I}_p) = -\mathbf{Q}_{\mathbf{D_g}\boldsymbol{\gamma}^T}\mathbf{P_g} \otimes \mathbf{Q}_{\mathbf{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}} = -\mathbf{P_g} \otimes \mathbf{Q}_{\mathbf{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}},$$

where the last equality holds because of $\boldsymbol{\gamma}\mathbf{D_g}\mathbf{P_g} = \boldsymbol{\gamma}\mathbf{f}\mathbf{g}^T = 0$. Then

$$\mathbf{\Gamma}\mathbf{V}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{\Phi}}(\mathbf{\Omega} - \mathbf{I}_{ph}) = -(\mathbf{D}_{\mathbf{g}}^{-1}\mathbf{Q_g} \otimes \mathbf{\Sigma}^{-1})\mathrm{diag}\{\mathbf{\Sigma}_{\mathbf{X}|y}\}(\mathbf{Q_g}\mathbf{P_g} \otimes \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{Q}_{\mathbf{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}}) = 0.$$

Therefore, $\mathbf{\Gamma U \Gamma U \Gamma} = \mathbf{\Gamma U \Gamma}$. The degrees of freedom are

$$\begin{aligned}
\text{trace}(\mathbf{U\Gamma}) &= \text{trace}(\mathbf{V}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{\Phi}}\mathbf{V}^{\frac{1}{2}}\mathbf{\Gamma}) \\
&= \text{trace}(\mathbf{Q}_{\mathbf{\Phi}}\mathbf{\Omega}) \\
&= \text{trace}(\mathbf{Q}_{\mathbf{\Phi}}) - \text{trace}(\mathbf{P}_{\mathbf{g}} \otimes \mathbf{Q}_{\mathbf{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}}) \\
&= ph - \text{rank}(\mathbf{\Phi}) - \text{rank}(\mathbf{P}_{\mathbf{g}} \otimes \mathbf{Q}_{\mathbf{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}}) \\
&= ph - \text{rank}(\mathbf{\Delta}) - (p-d),
\end{aligned}$$

where

$$\begin{aligned}
\text{rank}(\mathbf{\Delta}) &= \text{rank}([\boldsymbol{\gamma}^T \otimes \mathbf{I}_p, \mathbf{I}_h \otimes \boldsymbol{\beta}]) \\
&= \text{rank}([\boldsymbol{\gamma}^T \otimes \mathbf{Q}_{\boldsymbol{\beta}}, \mathbf{I}_h \otimes \boldsymbol{\beta}]) \\
&= \text{rank}(\boldsymbol{\gamma}^T \otimes \mathbf{Q}_{\boldsymbol{\beta}}) + \text{rank}(\mathbf{I}_h \otimes \boldsymbol{\beta}) \\
&= d(p-d) + hd \\
&= d(h+p-d).
\end{aligned}$$

Therefore, the degrees of freedom are $(p-d)(h-d-1)$.

Remarks

Given the correct dimension of the CS, the linearity condition and coverage condition can assure consistent estimation of the central space. However, we may need further assumptions for the test statistic to have some well-studied limit distribution. Particularly, the marginal covariance condition suffices for its distribution to converge to a chi-squared distribution. It is interesting that under the same conditions, the test statistic still converges to the same chi-squared distribution letting $\mathbf{V}_n = \text{diag}\{\hat{f}_y \hat{\mathbf{\Sigma}} \, (\hat{\text{E}}[\mathbf{\Sigma}_{\mathbf{X}|Y}])^{-1} \hat{\mathbf{\Sigma}}\}$ as shown in Corollary 2, where $\hat{\text{E}}[\mathbf{\Sigma}_{\mathbf{X}|Y}]$ is the average of sample conditional covariance of $\mathbf{X}$ given Y.

5.4 Variant of SIR

Corollary 2. *Assume that the (1) linearity condition, (2) marginal covariance condition, and (3) coverage condition are satisfied. Suppose* $\dim(\mathcal{S}_{Y|\mathbf{X}}) = d$. *Define a discrepancy function*

$$F_d(\mathbf{B}, \mathbf{C}) = \sum_{y=1}^{h} \hat{f}_y (\hat{\boldsymbol{\xi}}_y - \mathbf{B}\mathbf{C}_y)^T \hat{\boldsymbol{\Sigma}} \left(\hat{\mathrm{E}}[\boldsymbol{\Sigma}_{\mathbf{X}|Y}]\right)^{-1} \hat{\boldsymbol{\Sigma}} (\hat{\boldsymbol{\xi}}_y - \mathbf{B}\mathbf{C}_y), \quad (5.4)$$

where $\hat{\mathrm{E}}[\boldsymbol{\Sigma}_{\mathbf{X}|Y}] = \sum_{y=1}^{h} \frac{n_y}{n} \hat{\boldsymbol{\Sigma}}_{\mathbf{X}|Y=y}$. *Then, the test statistic*

$$n\hat{F}_d \xrightarrow{\mathcal{D}} \chi^2_{(p-d)(h-d-1)}, \text{ as } n \to \infty.$$

Meanwhile, the estimate of $\mathcal{S}_{Y|\mathbf{X}}$ *is consistent.*

We should be aware that the test statistic $n\hat{F}_d$ of (5.4) is always larger than Λ_d in SIR even though both converge to the same chi-squared distribution, since $\boldsymbol{\Sigma} > \mathrm{E}[\boldsymbol{\Sigma}_{\mathbf{X}|Y}]$.

Sketch Proof of Corollary 2

The proof of Corollary 2 is similar to that of Corollary 1. For notational simplicity, we still drop the subscripts from $\boldsymbol{\Gamma}_{\tilde{\boldsymbol{\xi}}}$ and $\boldsymbol{\Delta}_{\boldsymbol{\xi}}$ and let $\mathbf{S}$ denote $\boldsymbol{\Sigma}(\mathrm{E}[\boldsymbol{\Sigma}_{\mathbf{X}|Y}])^{-1}\boldsymbol{\Sigma}$. Here we have $\mathbf{V} = \mathbf{D_f} \otimes \mathbf{S}$. Thus,

$$\boldsymbol{\Phi} = \mathbf{V}^{\frac{1}{2}}\boldsymbol{\Delta} = (\mathbf{D_g}\boldsymbol{\gamma}^T \otimes \mathbf{S}^{\frac{1}{2}}, \mathbf{D_g} \otimes \mathbf{S}^{\frac{1}{2}}\boldsymbol{\beta}).$$

Using the same argument in the proof of Corollary 1 we have $\mathbf{Q}_{\boldsymbol{\Phi}} = \mathbf{Q}_{\mathbf{D_g}\boldsymbol{\gamma}^T} \otimes \mathbf{Q}_{\mathbf{S}^{\frac{1}{2}}\boldsymbol{\beta}}$.

Under the linearity condition, we have $\mathrm{E}[\mathbf{Z}|Y] \in \mathcal{S}_{Y|\mathbf{Z}}$. Thus,

$$\mathrm{Span}(\mathrm{Cov}(\mathrm{E}[\mathbf{Z}|Y])) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$$

and

$$\mathrm{Span}(\boldsymbol{\Sigma}^{-1}\mathrm{Cov}(\mathrm{E}[\mathbf{X}|Y])\boldsymbol{\Sigma}^{-1}) \subseteq \mathcal{S}_{Y|\mathbf{X}}.$$

Since $\boldsymbol{\Sigma} = \mathrm{Cov}(\mathrm{E}[\mathbf{X}|Y]) + \mathrm{E}[\mathrm{Cov}(\mathbf{X}|Y)]$, we have

$$\mathrm{Span}\{\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\,\mathrm{E}[\boldsymbol{\Sigma}_{\mathbf{X}|Y}]\,\boldsymbol{\Sigma}^{-1}\} = \mathrm{Span}\{\boldsymbol{\Sigma}^{-1} - \mathbf{S}^{-1}\} \subseteq \mathcal{S}_{Y|\mathbf{X}}.$$

Meanwhile, from Lemma 4, we know

$$\mathrm{Span}\{\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\mathbf{X}|Y}\boldsymbol{\Sigma}^{-1}\} \subseteq \mathcal{S}_{Y|\mathbf{X}}.$$

Thus, $\mathrm{Span}\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\mathbf{X}|Y}\boldsymbol{\Sigma}^{-1} - \mathbf{S}^{-1}\} \subseteq \mathcal{S}_{Y|\mathbf{X}}$. We have

$$\begin{aligned}
\boldsymbol{\Omega} - \mathbf{I}_{ph} &= (\mathbf{Q}_{\mathbf{g}} \otimes \mathbf{I}_p)\mathrm{diag}\{\mathbf{S}^{\frac{1}{2}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\mathbf{X}|y}\boldsymbol{\Sigma}^{-1}\mathbf{S}^{\frac{1}{2}} - \mathbf{I}_p\}(\mathbf{Q}_{\mathbf{g}} \otimes \mathbf{I}_p) - (\mathbf{P}_{\mathbf{g}} \otimes \mathbf{I}_p) \\
&= (\mathbf{Q}_{\mathbf{g}} \otimes \mathbf{I}_p)\mathrm{diag}\{\mathbf{S}^{\frac{1}{2}}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\mathbf{X}|y}\boldsymbol{\Sigma}^{-1} - \mathbf{S}^{-1})\mathbf{S}^{\frac{1}{2}}\}(\mathbf{Q}_{\mathbf{g}} \otimes \mathbf{I}_p) - (\mathbf{P}_{\mathbf{g}} \otimes \mathbf{I}_p)
\end{aligned}$$

where the diagonal matrix is over the values of Y. The first term above is in $\mathrm{Span}\{\mathbf{I}_h \otimes \mathbf{S}^{\frac{1}{2}}\boldsymbol{\beta}\} \subseteq \mathrm{Span}\{\boldsymbol{\Phi}\}$. Therefore,

$$\begin{aligned}
\mathbf{Q}_{\boldsymbol{\Phi}}(\boldsymbol{\Omega} - \mathbf{I}_{ph}) &= -\mathbf{Q}_{\boldsymbol{\Phi}}(\mathbf{P}_{\mathbf{g}} \otimes \mathbf{I}_p) \\
&= -\mathbf{Q}_{\mathbf{D}_{\mathbf{g}}\boldsymbol{\gamma}^T}\mathbf{P}_{\mathbf{g}} \otimes \mathbf{Q}_{\mathbf{S}^{\frac{1}{2}}\boldsymbol{\beta}} \\
&= -\mathbf{P}_{\mathbf{g}} \otimes \mathbf{Q}_{\mathbf{S}^{\frac{1}{2}}\boldsymbol{\beta}}
\end{aligned}$$

because $\mathbf{g}^T\mathbf{D}_{\mathbf{g}}\boldsymbol{\gamma}^T = \mathbf{f}^T\boldsymbol{\gamma}^T = 0$. Then, $\boldsymbol{\Gamma}\mathbf{V}^{\frac{1}{2}}\mathbf{Q}_{\boldsymbol{\Phi}}(\boldsymbol{\Omega} - \mathbf{I}_{ph}) = 0$. Hence, $\boldsymbol{\Gamma}\mathbf{U}\boldsymbol{\Gamma}\mathbf{U}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}\mathbf{U}\boldsymbol{\Gamma}$. The degrees of freedom can be obtained similarly as in the proof of Corollary 1.

Chapter 6

Weighted Chi-Squared Test

In Chapter 5, we discussed SIR in detail. The SIR test of dimension in Corollary 1 requires the linearity condition, the coverage condition, and the marginal covariance condition. Bura and Cook (2001b) proposed a *weighted chi-squared test* (WCT) which extended the SIR test to more general situations without assuming the marginal covariance and coverage conditions. WCT's estimation procedure and test statistics are the same as in SIR but the asymptotic distribution of the test statistic becomes a linear combination of independent chi-squares. In this chapter, we consider WCT via the minimum discrepancy approach.

As long as the linearity condition is satisfied, SIR always obtains a consistent estimate of $\mathcal{S}_{\xi}$ which may be a proper subset of the central subspace. Generally the test statistic Λ_d converges to a linear combination of independent chi-squares each with degree of freedom 1 as pointed out by Theorem 3 in Chapter 4. WCT summarizes the asymptotic properties in the following

corollary.

Corollary 3. *Assume only the linearity condition. Suppose* $\dim(\mathcal{S}_{\boldsymbol{\xi}}) = d$. Λ_m *is the test statistic in SIR for the hypothesis that* $d = m$. *Then,* Λ_d *converges to a linear combination of independent chi-squared distributions with degree of freedom 1, of which the coefficients are the eigenvalues of* $\mathbf{Q}_{\boldsymbol{\Phi}}\boldsymbol{\Omega}\mathbf{Q}_{\boldsymbol{\Phi}}$, *where* $\mathbf{V} = \mathbf{D_f} \otimes \boldsymbol{\Sigma}$, $\boldsymbol{\Phi} = \mathbf{V}^{\frac{1}{2}}\boldsymbol{\Delta}_{\boldsymbol{\xi}}$, $\boldsymbol{\Omega} = \mathbf{V}^{\frac{1}{2}}\boldsymbol{\Gamma}_{\tilde{\boldsymbol{\xi}}}\mathbf{V}^{\frac{1}{2}}$, $\boldsymbol{\Delta}_{\boldsymbol{\xi}}$ *and* $\boldsymbol{\Gamma}_{\tilde{\boldsymbol{\xi}}}$ *are defined as in (4.2) and (4.4). Meanwhile, the estimate of* $\mathcal{S}_{\boldsymbol{\xi}}$ *is consistent.*

Corollary 3 is a direct result of Theorem 3. The computations introduced in Section 4.2 are generally valid. Here we consider estimating the eigenvalues of $\mathbf{Q}_{\boldsymbol{\Phi}}\boldsymbol{\Omega}\mathbf{Q}_{\boldsymbol{\Phi}}$ in more detail. We have

$$\boldsymbol{\Phi} = \mathbf{V}^{\frac{1}{2}}\boldsymbol{\Delta}_{\boldsymbol{\xi}} = [\mathbf{D_g}\boldsymbol{\gamma}^T \otimes \boldsymbol{\Sigma}^{\frac{1}{2}}, \mathbf{D_g} \otimes \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}]$$

and

$$\mathbf{Q}_{\boldsymbol{\Phi}} = \mathbf{Q}_{\mathbf{D_g}\boldsymbol{\gamma}^T} \otimes \mathbf{Q}_{\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\beta}}$$

as illustrated in the proof of Corollary 1 in Chapter 5. Let $\mathbf{Z} = \boldsymbol{\Sigma}^{\frac{1}{2}}(\mathbf{X} - \mathrm{E}[\mathbf{X}])$ and $g_y = \sqrt{f_y}$ where $f_y = \Pr(Y = y)$ as defined previously. Define

$$\mathbf{T_Z} = [g_1\mathrm{E}[\mathbf{Z}|Y = 1], ..., g_h\mathrm{E}[\mathbf{Z}|Y = h]] = [\mathrm{E}[\mathbf{Z}|Y = 1], ..., \mathrm{E}[\mathbf{Z}|Y = h]]\mathbf{D_g}$$

with singular value decomposition

$$\begin{aligned}
\mathbf{T_Z} &= \boldsymbol{\Gamma}_1 \begin{bmatrix} \mathbf{D} & 0 \\ 0 & 0 \end{bmatrix} \boldsymbol{\Gamma}_2^T \\
&= [\boldsymbol{\Gamma}_{11}, \boldsymbol{\Gamma}_{12}] \begin{bmatrix} \mathbf{D} & 0 \\ 0 & 0 \end{bmatrix} [\boldsymbol{\Gamma}_{21}, \boldsymbol{\Gamma}_{22}]^T \\
&= \boldsymbol{\Gamma}_{11}\mathbf{D}\boldsymbol{\Gamma}_{21}^T
\end{aligned}$$

where $\mathbf{\Gamma}_{11}$ is an orthonormal $p \times d$ matrix, $\mathbf{\Gamma}_{21}$ is an orthonormal $h \times d$ matrix, and $\mathbf{D}$ is a $d \times d$ diagonal matrix with $d = \dim(\mathcal{S}_{\boldsymbol{\xi}})$. We know $\boldsymbol{\beta}\boldsymbol{\gamma} = \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{T}_{\mathbf{Z}}\mathbf{D}_{\mathbf{g}}^{-1}$. Therefore, without loss of generality, let $\boldsymbol{\beta} = \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Gamma}_{11}$ and $\boldsymbol{\gamma} = \mathbf{D}\mathbf{\Gamma}_{21}^T\mathbf{D}_{\mathbf{g}}^{-1}$. Thus,

$$\begin{aligned}
\mathbf{Q}_{\mathbf{\Phi}} &= \mathbf{Q}_{\mathbf{\Gamma}_{21}\mathbf{D}} \otimes \mathbf{Q}_{\mathbf{\Gamma}_{11}} \\
&= \mathbf{Q}_{\mathbf{\Gamma}_{21}} \otimes \mathbf{Q}_{\mathbf{\Gamma}_{11}} \\
&= (\mathbf{\Gamma}_{22}\mathbf{\Gamma}_{22}^T) \otimes (\mathbf{\Gamma}_{12}\mathbf{\Gamma}_{12}^T) \\
&= (\mathbf{\Gamma}_{22} \otimes \mathbf{\Gamma}_{12})(\mathbf{\Gamma}_{22}^T \otimes \mathbf{\Gamma}_{12}^T).
\end{aligned}$$

The eigenvalues of $\mathbf{Q}_{\mathbf{\Phi}}\mathbf{\Omega}\mathbf{Q}_{\mathbf{\Phi}}$ are the same as those of $(\mathbf{\Gamma}_{22}^T \otimes \mathbf{\Gamma}_{12}^T)\mathbf{\Omega}(\mathbf{\Gamma}_{22} \otimes \mathbf{\Gamma}_{12})$, where

$$\mathbf{\Omega} = (\mathbf{Q}_{\mathbf{g}} \otimes \mathbf{I}_p)\mathrm{diag}\{\mathbf{\Sigma}_{\mathbf{Z}|y}\}(\mathbf{Q}_{\mathbf{g}} \otimes \mathbf{I}_p)$$

We can replace $\mathbf{\Gamma}_{12}$, $\mathbf{\Gamma}_{22}$, and $\mathbf{\Omega}$ with their sample estimates to estimate the eigenvalues. Then, we compare the test statistic to the distribution of the linear combination of chi-squares. This is exactly the computation procedures of weighted chi-squared test (Bura and Cook 2001b), where the asymptotic distribution is derived following Eaton and Tyler (1994). Here we derive it through the minimum discrepancy approach.

Unlike SIR, WCT does not require either the marginal covariance condition or the coverage condition. Thus WCT can be used for inference about the dimension of the $\mathcal{S}_{\boldsymbol{\xi}}$ instead of the central subspace. One special case is to test $H_0 : d = 0$. When Y is independent of $\mathbf{X}$, the within-in slice covariance $\mathbf{\Sigma}_{\mathbf{X}|y}$ is same as the overall covariance matrix $\mathbf{\Sigma}$. Thus,

$\mathbf{Q}_{\mathbf{\Phi}}\mathbf{\Omega}\mathbf{Q}_{\mathbf{\Phi}} = (\mathbf{\Gamma}_{22}^T\mathbf{Q}_{\mathbf{g}}\mathbf{\Gamma}_{22} \otimes \mathbf{I}_{p-d})$ is idempotent with a trace $(p-d)(h-1)$. The limit distribution of $n\hat{F}_0$ is a chi-squared distribution with $p(h-1)$ degrees of freedom, which is the same as in SIR. Since the overall sample covariance matrix is a more stable estimate than within-slice sample covariances, we recommend using SIR for testing $d = 0$.

Chapter 7

Simple Inverse Regression Estimation

Theorem 3 in Chapter 4 shows that any positive definite $\mathbf{V}_n$ that converges to a constant matrix $\mathbf{V} > 0$ guarantees a consistent estimate of $\mathcal{S}_{\boldsymbol{\xi}}$ and that the minimum value of the discrepancy function also provides a venue for testing its dimension. SIR and WCT adopt $\mathbf{V}_n = (\mathbf{D}_{\hat{\mathbf{f}}} \otimes \hat{\boldsymbol{\Sigma}})$. In this chapter, we consider another sub-optimal class member that incorporates the variation of within-slice covariances.

As we have seen, both SIR and WCT use the sample covariance $\hat{\boldsymbol{\Sigma}}$ for all slice means regardless the variation of within slice covariances. When $\boldsymbol{\Sigma}_{\mathbf{X}|Y}$ varies considerably, we consider positive definite matrices $\mathbf{V}_{ny} = \hat{f}_y \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{\Sigma}}_{\mathbf{X}|y}^{-1} \hat{\boldsymbol{\Sigma}}$. In another words, we let

$$\mathbf{V} = \text{diag}\{f_y \boldsymbol{\Sigma} \boldsymbol{\Sigma}_{\mathbf{X}|y}^{-1} \boldsymbol{\Sigma}\} = [(\mathbf{D}_{\mathbf{g}}^{-1} \otimes \boldsymbol{\Sigma}^{-1}) \text{diag}\{\boldsymbol{\Sigma}_{\mathbf{X}|y}\} (\mathbf{D}_{\mathbf{g}}^{-1} \otimes \boldsymbol{\Sigma}^{-1})]^{-1}.$$

We expect this $\mathbf{V}$ can speed up the convergence of the test statistics. Therefore, we consider the discrepancy function

$$F_d(\mathbf{B},\mathbf{C}) = \sum_{y=1}^{h} \hat{f}_y(\hat{\boldsymbol{\xi}}_y - \mathbf{B}\mathbf{C}_y)^T \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{\Sigma}}_{\mathbf{X}|y}^{-1} \hat{\boldsymbol{\Sigma}} (\hat{\boldsymbol{\xi}}_y - \mathbf{B}\mathbf{C}_y), \tag{7.1}$$

where $\hat{\boldsymbol{\xi}}_y = \hat{\boldsymbol{\Sigma}}^{-1}(\bar{\mathbf{X}}_{y\bullet} - \bar{\mathbf{X}}_{\bullet\bullet})$ as defined previously. We can estimate $\boldsymbol{\beta}$ by minimizing (7.1) and estimate the dimension of $\mathcal{S}_{\boldsymbol{\xi}}$ by Theorem 3, Chapter 4. We call this method, *simple inverse regression estimation* (Simple IRE).

7.1 Algorithm for Minimization

In SIR and WCT, since $\mathbf{V}_{ny}$'s are the same except for a scalar, we can achieve minimization by a spectral decomposition. When the $\boldsymbol{\Sigma}_{\mathbf{X}|y}$'s are not the same, we have to rely on more complicate numerical methods. Simple IRE is based on the minimization of the discrepancy function

$$\begin{aligned} F_d(\mathbf{B},\mathbf{C}) &= \sum_{y=1}^{h} (\hat{\boldsymbol{\xi}}_y - \mathbf{B}\mathbf{C}_y)^T \mathbf{V}_{ny} (\hat{\boldsymbol{\xi}}_y - \mathbf{B}\mathbf{C}_y) \\ &= \sum_{y=1}^{h} (\mathbf{V}_{ny}^{\frac{1}{2}} \hat{\boldsymbol{\xi}}_y - \mathbf{V}_{ny}^{\frac{1}{2}} \mathbf{B}\mathbf{C}_y)^T (\mathbf{V}_{ny}^{\frac{1}{2}} \hat{\boldsymbol{\xi}}_y - \mathbf{V}_{ny}^{\frac{1}{2}} \mathbf{B}\mathbf{C}_y). \end{aligned}$$

This is a special case of finding the values of $\mathbf{B}$ and $\mathbf{C}$ which minimize a generic discrepancy function of the form

$$H(\mathbf{B},\mathbf{C}) = \sum_{j=1}^{h} (\boldsymbol{\alpha}_j - \mathbf{S}_j \mathbf{B}\mathbf{C}_j)^T (\boldsymbol{\alpha}_j - \mathbf{S}_j \mathbf{B}\mathbf{C}_j), \tag{7.2}$$

where $\mathbf{B} \in \mathbb{R}^{p\times d}$, $\mathbf{C} = (\mathbf{C}_1, \cdots, \mathbf{C}_h) \in \mathbb{R}^{d\times h}$, $\boldsymbol{\alpha}_j \in \mathbb{R}^p$ and $\mathbf{S}_j \in \mathbb{R}^{p\times p}$. The $\mathbf{S}_j$'s are positive definite. All $\boldsymbol{\alpha}_j$ and $\mathbf{S}_j$ are fixed in the minimization

algorithm. The discrepancy function H can be minimized by treating it as a separable nonlinear least squares problem. We have separate sets of parameters, $\mathbf{B}$ and $\mathbf{C}$ in (7.2). Given $\mathbf{B}$, the minimization with respect to $\mathbf{C}$ is straightforward: We only need to solve h independent linear regressions of $\boldsymbol{\alpha}_j$ on $\mathbf{S}_j\mathbf{B}$, $j = 1, \ldots, h$. On the other hand, consider minimizing H with respect to one column $\mathbf{b}_k$ of $\mathbf{B}$, given $\mathbf{C}$ and the remaining columns of $\mathbf{B}$ and subject to the length constraint $\|\mathbf{b}_k\| = 1$ and the orthogonality constraint $\mathbf{b}_k^T\mathbf{B}_{(-k)} = 0$, where $\mathbf{B}_{(-k)}$ is the matrix that is left after taking away $\mathbf{b}_k$ from $\mathbf{B}$. For this partial minimization problem, H can be re-expressed as

$$H^*(\mathbf{b}_k) = \sum_{j=1}^{h} (\boldsymbol{\alpha}_j^{(k)} - c_{jk}\mathbf{S}_j\mathbf{b}_k)^T(\boldsymbol{\alpha}_j^{(k)} - c_{jk}\mathbf{S}_j\mathbf{b}_k),$$

where $\boldsymbol{\alpha}_j^{(k)} = \boldsymbol{\alpha}_j - \mathbf{S}_j\mathbf{B}_{(-k)}\mathbf{C}_{j(-k)}$, c_{jk} is the k-th element of $\mathbf{C}_j$, and $\mathbf{C}_{j(-k)}$ consists of all but the k-th element of $\mathbf{C}_j$. Based on Lemma 6 in Appendix B, the solution to this partial minimization problem is as follows: The argument that minimizes $H^*(\mathbf{b}_k)$ subject to the constraints $\|\mathbf{b}_k\| = 1$ and $\mathbf{b}_k^T\mathbf{B}_{(-k)} = 0$ is

$$\hat{\mathbf{b}}_k = \mathbf{W}_2^{-1}[\mathbf{I} - \mathbf{B}_{(-k)}(\mathbf{B}_{(-k)}^T\mathbf{W}_2^{-1}\mathbf{B}_{(-k)})^-\mathbf{B}_{(-k)}^T\mathbf{W}_2^{-1}]\mathbf{W}_1,$$

where $\mathbf{W}_1 = \sum_{j=1}^{h} c_{jk}\mathbf{S}_j^T\boldsymbol{\alpha}_j^{(k)}$ and $\mathbf{W}_2 = \sum_{j=1}^{h} c_{jk}^2\mathbf{S}_j^T\mathbf{S}_j$.

We now describe an algorithm for the minimization of H in the spirit of the alternating least square method introduced for Optimal IRE in Section 3.4. The outline is as follows.

Outline of Simple IRE

For a given dimension d, the algorithm searches over $\mathbb{R}^{p\times d}$ for $\mathbf{B}$ which minimizes (7.2).

1. Choose the initial $\mathbf{B} \leftarrow (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_d)$. Constant initial starting vectors are often a good choice. Our experience is that the initial values do not generally affect the ultimate result.

2. Calculate

$$\begin{aligned}\mathbf{C} &= \arg_{\mathbf{C}^*} \min H(\mathbf{B}, \mathbf{C}^*) \\ &= ((\mathbf{B}^T\mathbf{S}_1^T\mathbf{S}_1\mathbf{B})^{-1}\mathbf{B}^T\mathbf{S}_1^T\boldsymbol{\alpha}_1, \ldots, (\mathbf{B}^T\mathbf{S}_h^T\mathbf{S}_h\mathbf{B})^{-1}\mathbf{B}^T\mathbf{S}_h^T\boldsymbol{\alpha}_h)\end{aligned}$$

   Assign $e_0 \leftarrow H(\mathbf{B}, \mathbf{C})$ and $iter \leftarrow 0$.

3. (a) For $k = 1, 2, \ldots, d$:

   - At the current step $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_{k-1}, \mathbf{b}_k, \mathbf{b}_{k+1}, \ldots, \mathbf{b}_d)$. Assign

$$\boldsymbol{\alpha}_j^{(k)} \leftarrow \boldsymbol{\alpha}_j - \mathbf{S}_j\mathbf{B}_{(-k)}\mathbf{C}_{j(-k)}$$

     which is a residual vector with $\mathbf{b}_k$ excluded. Find a new $\mathbf{b}_k$ minimizing the function with the constraint that it is orthogonal to $\mathbf{B}_{(-k)}$ and has length 1:

$$\mathbf{b}_k^* = \arg_{\mathbf{b}\perp\mathbf{B}_{(-k)},\ \|\mathbf{b}\|=1} \min \sum_{j=1}^{h} (\boldsymbol{\alpha}_j^{(k)} - c_{jk}\mathbf{S}_j\mathbf{b})^T(\boldsymbol{\alpha}_j^{(k)} - c_{jk}\mathbf{S}_j\mathbf{b}).$$

     This minimization can be achieved using Lemma 6 in Appendix B.

- Update

$$\mathbf{B} \leftarrow (\mathbf{b}_1, \ldots, \mathbf{b}_{k-1}, \mathbf{b}_k^*, \mathbf{b}_{k+1}, \ldots, \mathbf{b}_d)$$

$$\mathbf{C} \leftarrow \arg_{\mathbf{C}^*} \min H(\mathbf{B}, \mathbf{C}^*)$$

(b) $e_1 \leftarrow H(\mathbf{B}, \mathbf{C})$ and $iter \leftarrow iter + 1$.

4. Return to step 3 until e_1 no longer changes and then assign $\widetilde{\mathbf{B}} \leftarrow \mathbf{B}$ and exit.

At termination, $\widetilde{\mathbf{B}}$ is an estimate of $\boldsymbol{\beta}$. After one iteration of step 3, the algorithm produces a monotonically decreasing series of evaluations and thus is guaranteed to converge because $H \geq 0$.

This algorithm will not necessarily produce an analogous ordering as in SIR or WCT. However, as in Optimal IRE, we can construct an ordered basis for Span$\{\widetilde{\mathbf{B}}\}$ with respect to the amount by which directions decrease $H(\mathbf{B}, \mathbf{C})$. For example, the most important direction is

$$\hat{\mathbf{b}}_1 = \arg_{\mathbf{b}} \min \sum_{j=1}^{h} (\mathbf{Q}_{\mathbf{S}_j \mathbf{b}} \, \boldsymbol{\alpha}_j)^T (\mathbf{Q}_{\mathbf{S}_j \mathbf{b}} \, \boldsymbol{\alpha}_j).$$

where the minimization is over $\mathbf{b} \in$ Span$\{\widetilde{\mathbf{B}}\}$ with $\|\mathbf{b}\| = 1$. The second direction is

$$\hat{\mathbf{b}}_2 = \arg_{\mathbf{b}} \min \sum_{j=1}^{h} (\mathbf{Q}_{\mathbf{S}_j [\mathbf{b}, \hat{\mathbf{b}}_1]} \, \boldsymbol{\alpha}_j)^T (\mathbf{Q}_{\mathbf{S}_j [\mathbf{b}, \hat{\mathbf{b}}_1]} \, \boldsymbol{\alpha}_j).$$

where the minimization is over $\mathbf{b} \in$ Span$\{\widetilde{\mathbf{B}}\}$ with $\|\mathbf{b}\| = 1$ and $\mathbf{b}^T \hat{\mathbf{b}}_1 = 0$. And so on.

7.2 Asymptotic Distribution of the Test Statistic

The following corollary provides the asymptotic distribution of the Simple IRE test statistics—$n\hat{F}_d$ of (7.1)—based on Theorem 3.

Corollary 4. *Assume only the linearity condition. Suppose* $\dim(\mathcal{S}_{\boldsymbol{\xi}}) = d$. *The* $n\hat{F}_d$ *of*

$$F_d(\mathbf{B}, \mathbf{C}) = \sum_{y=1}^{h} \hat{f}_y(\hat{\boldsymbol{\xi}}_y - \mathbf{B}\mathbf{C}_y)^T \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{\Sigma}}_{\mathbf{X}|y}^{-1} \hat{\boldsymbol{\Sigma}} (\hat{\boldsymbol{\xi}}_y - \mathbf{B}\mathbf{C}_y)$$

converges to a linear combination of independent chi-squared distributions with degree of freedom 1, of which the coefficients are the eigenvalues of $\mathbf{Q}_{\boldsymbol{\Phi}} \boldsymbol{\Omega} \mathbf{Q}_{\boldsymbol{\Phi}}$, *where* $\mathbf{V} = \operatorname{diag}\{f_y \boldsymbol{\Sigma} \boldsymbol{\Sigma}_{\mathbf{X}|y}^{-1} \boldsymbol{\Sigma}\}$, $\boldsymbol{\Phi} = \mathbf{V}^{\frac{1}{2}} \boldsymbol{\Delta}_{\boldsymbol{\xi}}$, $\boldsymbol{\Omega} = \mathbf{V}^{\frac{1}{2}} \boldsymbol{\Gamma}_{\tilde{\boldsymbol{\xi}}} \mathbf{V}^{\frac{1}{2}}$, $\boldsymbol{\Delta}_{\boldsymbol{\xi}}$ *and* $\boldsymbol{\Gamma}_{\tilde{\boldsymbol{\xi}}}$ *are defined as in (4.2) and (4.4). Meanwhile, the estimate of* $\mathcal{S}_{\boldsymbol{\xi}}$ *is consistent.*

Corollary 4 is quite general which does not require any special condition on the distribution of $(\mathbf{X}, Y)$. If we we add more assumptions by assuming linearity condition, marginal covariance condition, and coverage condition, then the asymptotic distribution is simplified to a chi-squared distribution with degrees of freedom $(p-d)(h-d-1)$, which is the same as in SIR. Corollary 5 states this result.

Corollary 5. *Assume*

1. *The linearity condition:* $\mathrm{E}[\mathbf{Z}|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{Z}}}\mathbf{Z}] = \mathbf{P}_{\mathcal{S}_{Y|\mathbf{Z}}}\mathbf{Z}$.

2. *The marginal covariance condition:* $\mathrm{Cov}[\mathbf{Z}|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{Z}}}\mathbf{Z}] = \mathbf{Q}_{\mathcal{S}_{Y|\mathbf{Z}}}$.

3. The coverage condition: $\mathcal{S}_{\boldsymbol{\xi}} = \mathcal{S}_{Y|\mathbf{X}}$.

Suppose $\dim(\mathcal{S}_{Y|\mathbf{X}}) = d$. *Then, the test statistic of (7.1)*

$$n\hat{F}_d \xrightarrow{\mathcal{D}} \chi^2_{(p-d)(h-d-1)}, \quad \text{as } n \to \infty.$$

Meanwhile, the estimate of $\mathcal{S}_{Y|\mathbf{X}}$ *is consistent.*

Proof of Corollary 5

Simple IRE adopts $\mathbf{V} = \text{diag}\{f_y \boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}_{\mathbf{X}|y}\boldsymbol{\Sigma}\}$. For notational simplicity, we drop the subscripts from $\boldsymbol{\Gamma}_{\tilde{\boldsymbol{\xi}}}$ and $\boldsymbol{\Delta}_{\boldsymbol{\xi}}$, and we let $\mathbf{S} = \text{diag}\{\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}_{\mathbf{X}|y}\boldsymbol{\Sigma}\} \in \mathbb{R}^{ph\times ph}$, where the diagonal matrix is over the values of Y. Therefore,

$$\begin{aligned}
\mathbf{V} &= \mathbf{S}(\mathbf{D_f} \otimes \mathbf{I}) = (\mathbf{D_f} \otimes \mathbf{I})\mathbf{S} = (\mathbf{D_g} \otimes \mathbf{I})\mathbf{S}(\mathbf{D_g} \otimes \mathbf{I}) \\
\boldsymbol{\Gamma} &= (\mathbf{D_g^{-1}Q_g} \otimes \mathbf{I})\mathbf{S}^{-1}(\mathbf{Q_g D_g^{-1}} \otimes \mathbf{I}), \\
\boldsymbol{\Phi} &= \mathbf{V}^{\frac{1}{2}}\boldsymbol{\Delta} = \mathbf{S}^{\frac{1}{2}}(\mathbf{D_g}\boldsymbol{\gamma}^T \otimes \mathbf{I}, \mathbf{D_g} \otimes \boldsymbol{\beta}), \\
\boldsymbol{\Omega} &= \mathbf{V}^{\frac{1}{2}}\boldsymbol{\Gamma}\mathbf{V}^{\frac{1}{2}} = \mathbf{S}^{\frac{1}{2}}(\mathbf{Q_g} \otimes \mathbf{I})\mathbf{S}^{-1}(\mathbf{Q_g} \otimes \mathbf{I})\mathbf{S}^{\frac{1}{2}}.
\end{aligned}$$

By Lemma 4 in Section 5.3,

$$\text{Span}(\mathbf{S}^{-1} - \mathbf{I} \otimes \boldsymbol{\Sigma}^{-1}) \subseteq \text{Span}(\mathbf{I} \otimes \boldsymbol{\beta}),$$

therefore,

$$\text{Span}(\mathbf{S}^{\frac{1}{2}}(\mathbf{S}^{-1} - \mathbf{I} \otimes \boldsymbol{\Sigma}^{-1})) \subseteq \text{Span}(\boldsymbol{\Phi}).$$

Let $\mathbf{S}^{-1} - \mathbf{I} \otimes \boldsymbol{\Sigma}^{-1} = (\mathbf{I} \otimes \boldsymbol{\beta})\mathbf{A}$, where $\mathbf{A} \in \mathbb{R}^{ph\times ph}$, then

$$\text{Span}(\mathbf{S}^{\frac{1}{2}}(\mathbf{Q_g} \otimes \mathbf{I})(\mathbf{S}^{-1} - \mathbf{I} \otimes \boldsymbol{\Sigma}^{-1})) \subseteq \text{Span}(\mathbf{S}^{\frac{1}{2}}(\mathbf{Q_g} \otimes \mathbf{I})(\mathbf{I} \otimes \boldsymbol{\beta}))$$

$$\subseteq \text{Span}(\mathbf{S}^{\frac{1}{2}}(\mathbf{I} \otimes \boldsymbol{\beta})) \subseteq \text{Span}(\boldsymbol{\Phi}).$$

By Theorem 3, to reach desired result, we only need to show that $\mathbf{Q}_\Phi\mathbf{\Omega}\mathbf{Q}_\Phi$ is idempotent. First, we consider

$$\begin{aligned}
\mathbf{Q}_\Phi\mathbf{\Omega} &= \mathbf{Q}_\Phi\mathbf{S}^{\frac{1}{2}}(\mathbf{Q}_\mathbf{g}\otimes\mathbf{I})\mathbf{S}^{-1}(\mathbf{Q}_\mathbf{g}\otimes\mathbf{I})\mathbf{S}^{\frac{1}{2}} \\
&= \mathbf{Q}_\Phi\mathbf{S}^{\frac{1}{2}}(\mathbf{Q}_\mathbf{g}\otimes\mathbf{I})(\mathbf{S}^{-1}-\mathbf{I}\otimes\mathbf{\Sigma}^{-1})(\mathbf{Q}_\mathbf{g}\otimes\mathbf{I})\mathbf{S}^{\frac{1}{2}}+\mathbf{Q}_\Phi\mathbf{S}^{\frac{1}{2}}(\mathbf{Q}_\mathbf{g}\otimes\mathbf{\Sigma}^{-1})\mathbf{S}^{\frac{1}{2}} \\
&= \mathbf{Q}_\Phi\mathbf{S}^{\frac{1}{2}}(\mathbf{Q}_\mathbf{g}\otimes\mathbf{\Sigma}^{-1})\mathbf{S}^{\frac{1}{2}} \\
&= \mathbf{Q}_\Phi\mathbf{S}^{\frac{1}{2}}(\mathbf{I}\otimes\mathbf{\Sigma}^{-1})(\mathbf{Q}_\mathbf{g}\otimes\mathbf{I})\mathbf{S}^{\frac{1}{2}} \\
&= \mathbf{Q}_\Phi\mathbf{S}^{\frac{1}{2}}(\mathbf{I}\otimes\mathbf{\Sigma}^{-1}-\mathbf{S}^{-1})(\mathbf{Q}_\mathbf{g}\otimes\mathbf{I})\mathbf{S}^{\frac{1}{2}}+\mathbf{Q}_\Phi\mathbf{S}^{-\frac{1}{2}}(\mathbf{Q}_\mathbf{g}\otimes\mathbf{I})\mathbf{S}^{\frac{1}{2}} \\
&= \mathbf{Q}_\Phi\mathbf{S}^{-\frac{1}{2}}(\mathbf{Q}_\mathbf{g}\otimes\mathbf{I})\mathbf{S}^{\frac{1}{2}} \\
&= \mathbf{Q}_\Phi-\mathbf{Q}_\Phi\mathbf{S}^{-\frac{1}{2}}(\mathbf{P}_\mathbf{g}\otimes\mathbf{I})\mathbf{S}^{\frac{1}{2}}.
\end{aligned}$$

Thus,

$$\mathbf{Q}_\Phi\mathbf{\Omega}\mathbf{Q}_\Phi=\mathbf{Q}_\Phi-\mathbf{Q}_\Phi\mathbf{S}^{-\frac{1}{2}}(\mathbf{P}_\mathbf{g}\otimes\mathbf{I})\mathbf{S}^{\frac{1}{2}}\mathbf{Q}_\Phi=\mathbf{Q}_\Phi-\mathbf{Q}_\Phi\mathbf{S}^{-\frac{1}{2}}(\mathbf{P}_\mathbf{g}\otimes\mathbf{Q}_\beta)\mathbf{S}^{\frac{1}{2}}\mathbf{Q}_\Phi$$

because of $(\mathbf{P}_\mathbf{g}\otimes\mathbf{P}_\beta)\mathbf{S}^{\frac{1}{2}}\mathbf{Q}_\Phi=0$. Since $[\mathbf{S}^{-\frac{1}{2}}(\mathbf{P}_\mathbf{g}\otimes\mathbf{Q}_\beta)\mathbf{S}^{\frac{1}{2}}]^T\mathbf{\Phi}=0$, thus

$$\mathbf{Q}_\Phi\mathbf{\Omega}\mathbf{Q}_\Phi=\mathbf{Q}_\Phi-\mathbf{S}^{-\frac{1}{2}}(\mathbf{P}_\mathbf{g}\otimes\mathbf{Q}_\beta)\mathbf{S}^{\frac{1}{2}}\mathbf{Q}_\Phi.$$

Therefore,

$$\mathbf{Q}_\Phi\mathbf{\Omega}\mathbf{Q}_\Phi\mathbf{\Omega}\mathbf{Q}_\Phi=\mathbf{Q}_\Phi\mathbf{\Omega}\mathbf{Q}_\Phi-\mathbf{Q}_\Phi\mathbf{\Omega}\mathbf{S}^{-\frac{1}{2}}(\mathbf{P}_\mathbf{g}\otimes\mathbf{Q}_\beta)\mathbf{S}^{\frac{1}{2}}\mathbf{Q}_\Phi=\mathbf{Q}_\Phi\mathbf{\Omega}\mathbf{Q}_\Phi$$

because of $\mathbf{\Omega}\mathbf{S}^{-\frac{1}{2}}(\mathbf{P}_\mathbf{g}\otimes\mathbf{Q}_\beta)=0$. $\mathbf{Q}_\Phi\mathbf{\Omega}\mathbf{Q}_\Phi$ is indeed idempotent. The degrees of freedom are

$$\begin{aligned}
\text{trace}(\mathbf{Q}_\Phi\mathbf{\Omega}\mathbf{Q}_\Phi) &= \text{trace}(\mathbf{Q}_\Phi)-\text{trace}(\mathbf{S}^{-\frac{1}{2}}(\mathbf{P}_\mathbf{g}\otimes\mathbf{Q}_\beta)\mathbf{S}^{\frac{1}{2}}) \\
&= \text{trace}(\mathbf{Q}_\Phi)-\text{trace}(\mathbf{P}_\mathbf{g}\otimes\mathbf{Q}_\beta) \\
&= \text{rank}(\mathbf{Q}_\Phi)-\text{rank}(\mathbf{P}_\mathbf{g}\otimes\mathbf{Q}_\beta) \\
&= ph-\text{rank}(\mathbf{\Delta})-(p-d) \\
&= (p-d)(h-d-1).
\end{aligned}$$

Chapter 8

Comparison of SIR, WCT, Simple IRE, and Optimal IRE

Four methods for sufficient dimension reduction were discussed in previous chapters: SIR, WCT, Simple IRE, and Optimal IRE. All of them are members of the MDA family that considers discrepancy functions in the form of

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

where $\mathbf{V}_n > 0$ converges to $\mathbf{V} > 0$ in probability (cf. Section 2.2). Table 8.1 summarizes the choices of $(\mathbf{M}_n, \mathbf{V})$, asymptotic distributions of test statistics for dimension, and required conditions on the predictor distribution. This table is not exhaustive. Refer to corresponding chapters for details.

A simulation study considers three models in this chapter, where all three conditions are satisfied. Thus, theoretically, test statistics of all four meth-

Table 8.1: Summary of Dimension Reduction Methods

	$\mathbf{M}_n$	$\mathbf{V}$	Asymp. Dist.	Conditions
SIR	$\mathbf{I}_h$	$\mathbf{D_f} \otimes \mathbf{\Sigma}$	$\chi^2_{(p-d)(h-d-1)}$	(1)(2)(3)*
WCT	$\mathbf{I}_h$	$\mathbf{D_f} \otimes \mathbf{\Sigma}$	$\sum_{i=1}^{ph} \lambda_i \chi^2_i(1)$	(1)
Simple IRE	$\mathbf{I}_h$	$\text{diag}\{f_y \mathbf{\Sigma} \mathbf{\Sigma}^{-1}_{\mathbf{X}\|Y} \mathbf{\Sigma}\}$	$\sum_{i=1}^{ph} \lambda_i \chi^2_i(1)$	(1)
Optimal IRE	$\mathbf{D}_{\hat{\mathbf{f}}}\mathbf{A}$	$\mathbf{\Gamma}^{-1}_{\hat{\boldsymbol{\zeta}}}$	$\chi^2_{(p-d)(h-d-1)}$	(1)

Note: (1) linearity condition; (2) coverage condition;

(3) marginal covariance condition.

ods should converge to their nominal asymptotic distributions. We compare performances of SIR, WCT, Optimal IRE, and Simple IRE as in Corollary 4 by examining p-values of their test statistics for dimension when the null hypothesis is true. The closer the empirical distribution is to a uniform distribution, the faster the convergence.

8.1 Model A

Model A is a 1-D exponential model. The predictor $\mathbf{X} = (X_1, X_2, \ldots, X_5)^T$ is 5-dimensional. The response Y is generated according to the model

$$Y = \exp[-(X_1 + X_2 + X_3)] + 0.5\epsilon,$$

where X_1, X_2, X_3, V are i.i.d. $t_{(3)}$, and

$$\begin{aligned} X_4 &= 4(X_1 + X_2 + X_3)W_1 + V, \\ X_5 &= 4(X_1 + X_2 + X_3)W_2 - V. \end{aligned}$$

The variables ϵ, W_1,W_2 are i.i.d. standard normal and they are independent of X_1, X_2, X_3, and V. We ran 1000 simulations at each of the sample sizes 200, 400, and 800. For each simulation, we tested the hypothesis $d = 1$ with slice numbers $h = 4$, 6, and 8. The results are in Table 8.2. We also plotted p-values against the uniform quantiles in Figure 8.1. If the actual distribution of test statistic is close to the asymptotic distribution, the points should lie near a straight line. The test statistic of SIR converges relatively slowly to its limiting distribution. For WCT, Simple IRE, and Optimal IRE, the empirical levels converge to the nominal levels much faster. Since Simple IRE and Optimal IRE use estimates of within-slice covariance matrices, they need a larger sample size to reach an agreement similar to that of WCT. We also notice that the impact of the number of slices h is relatively small.

8.2 Model B

Here, we consider a 2-D model. The response Y is generated by

$$Y = \frac{X_1}{0.5 + (X_2 + 1.5)^2} + \epsilon,$$

where $\mathbf{X} = (X_1, \ldots, X_5)^T$ is 5-dimensional multivariate normal with zero mean and identity covariance matrix, ϵ is a standard normal independent of $\mathbf{X}$. This model is as same as model (6.3) in Li (1991). We ran 1000 simulations with sample sizes 200, 400, and 800. Table 8.3 presents the empirical levels for testing $H_0 : d = 2$. It is clear that the test statistics for $H_0 : d = 2$ converge to their asymptotic distributions for all four methods. In this model, we assume normality of the predictor $\mathbf{X}$, which is the ideal

Table 8.2: Estimated level in percent of nominal 1 and 5 percent tests for Model A: $H_0 : d = 1$ vs $H_1 : d > 1$. The nominal simulation standard errors are 0.3 for 1 percent and 0.7 for 5 percent.

	SIR		WCT		Simple IRE		Optimal IRE	
n	1	5	1	5	1	5	1	5
				$h = 4$				
200	0.2	1.7	0.6	5.0	0.7	7.0	2.1	7.8
400	0.3	1.7	0.6	4.7	1.2	4.4	2.1	7.2
800	0.1	1.5	0.5	5.4	0.7	6.0	1.6	8.0
				$h = 6$				
200	0.2	2.3	0.8	4.6	1.8	6.8	1.6	7.7
400	0	1.7	0.4	4.1	0.9	6.1	0.9	6.1
800	0.5	1.7	0.9	3.9	0.9	4.7	1.3	6.2
				$h = 8$				
200	0.5	2.2	0.8	4.9	2.8	10.8	1.6	8.3
400	0.1	2.2	0.3	4.8	1.0	6.6	1.1	6.5
800	0.4	2.5	0.8	4.5	0.9	5.3	1.6	6.2

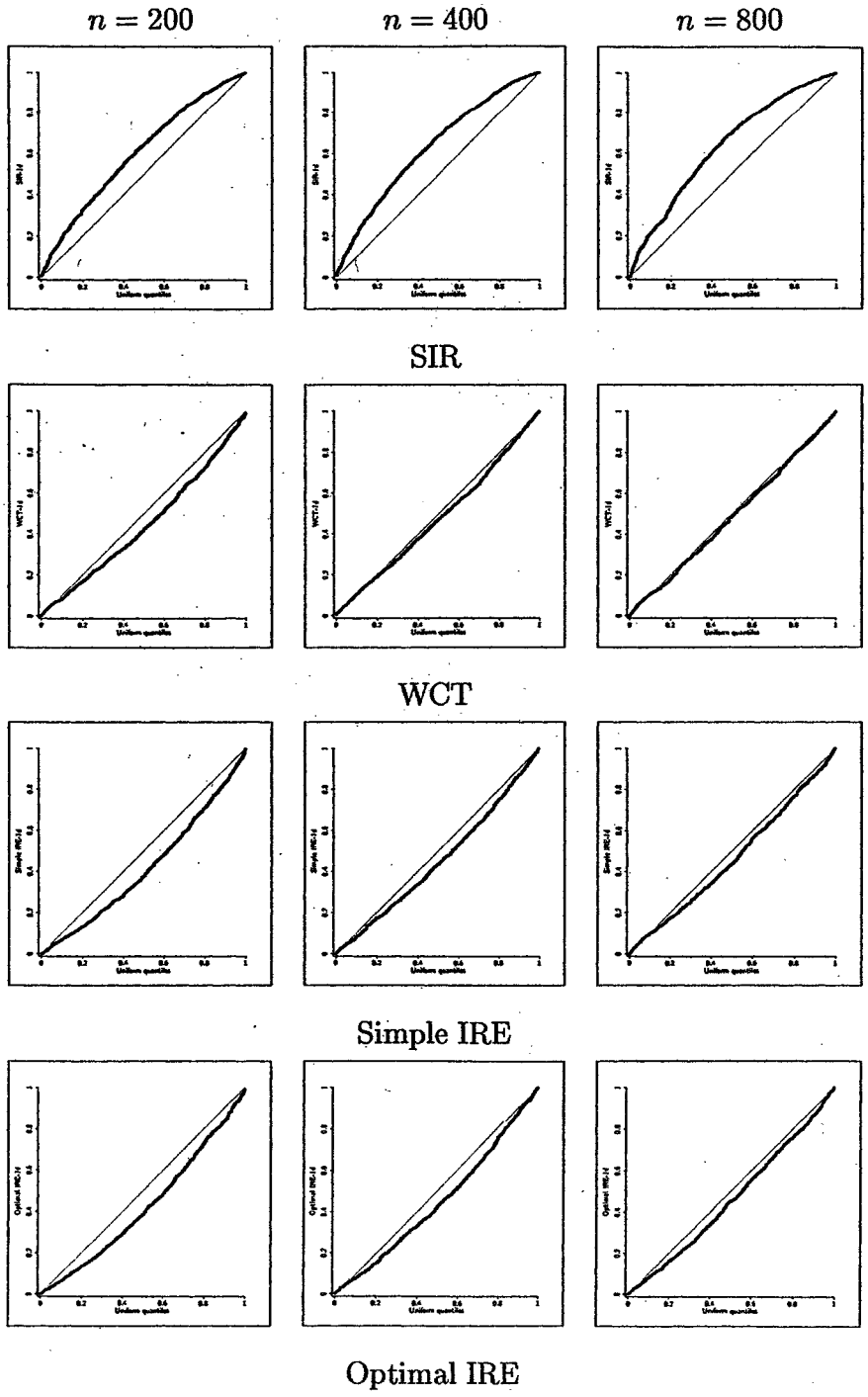


Figure 8.1: Model A: uniform quantile plot of p-values for testing $d = 1$ with $h = 6$.

situation for SIR. The new methods Simple IRE and Optimal IRE also do well.

Table 8.3: Estimated level in percent of nominal 1 and 5 percent tests for Model B. The nominal simulation standard errors are 0.3 for 1 percent and 0.7 for 5 percent.

	SIR		WCT		Simple IRE		Optimal IRE	
n	1	5	1	5	1	5	1	5
				$h = 4$				
200	0.1	1.0	0.2	1.1	0.3	1.2	0.2	1.1
400	0.4	1.7	0.5	2.1	0.3	2.4	0.3	2.2
800	0.6	3.8	0.7	4.0	0.7	3.7	0.7	3.7
				$h = 6$				
200	0	0.7	0	0.7	0.1	1.8	0	0.7
400	0.1	1.9	0.2	2.2	0.5	2.9	0.5	2.2
800	0.7	4.1	0.7	4.3	1.0	4.6	1.0	3.9
				$h = 8$				
200	0	0.7	0	1.0	0.4	2.3	0	0.7
400	0.5	2.4	0.6	2.4	0.8	4.0	0.1	2.3
800	0.8	3.8	0.9	4.0	1.3	4.5	0.9	3.7

8.3 Model C

We consider another 2-D model. The response Y is generated according to

$$Y = (4 + X_1)(2 + X_2 + X_3) + 0.5\epsilon,$$

where ϵ is a standard normal and

$$\begin{aligned} X_1 &= W_1 \\ X_2 &= V_1 + W_2/2, \\ X_3 &= -V_1 + W_2/2, \\ X_4 &= V_2 + V_3, \\ X_5 &= V_2 - V_3. \end{aligned}$$

All V_i's and W_j's are independent. The variables V_1, V_2, V_4 are i.i.d. $t_{(4)}$, $V_3 \sim t_{(3)}$, $V_5 \sim t_{(5)}$, and W_1 and W_2 are i.i.d. gamma(.25) random variables. Model C is the same as model (24) of Velilla (1998) and (19) of Bura and Cook (2001b). We ran 1000 simulations at each of the sample sizes 200, 800, and 3200. Table 8.4, Part A, gives the empirical power for testing $H_0 : d = 1$. Part B presents the empirical level for testing $H_0 : d = 2$. We can see from Table 8.4 that while all test statistics for $d = 2$ gradually converge, the two IRE methods converge faster than SIR or WCT. Figure 8.2 demonstrates IRE methods' advantage over the other two methods with respect to the convergence speed. Meanwhile, at each sample size, the IRE methods have better power than the other two to detect $d > 1$.

From this simulation study, we notice that with reasonable sample size, Optimal IRE and Simple IRE have better performance in terms of conver-

gence speed and power. Theoretically, Optimal IRE is the optimal method. However, empirical experience shows that Simple IRE is more stable and can serve as a viable alternative.

Table 8.4: Estimated power or level in percent of nominal 1 and 5 percent tests for Model C.

	SIR		WCT		Simple IRE		Optimal IRE	
n	1	5	1	5	1	5	1	5
Part A: estimated power for $H_0 : d = 1$ *vs* $H_1 : d > 1$								
$h = 4$								
200	0.4	3.1	1.0	6.5	7.3	21.2	4.3	15.5
800	1.7	7.0	3.6	15.4	41.4	64.9	30.9	55.7
3200	32.6	64.5	49.7	80.4	99.5	100	99.4	99.9
$h = 6$								
200	1.7	6.7	3.2	11.1	12.7	28.7	3.5	14.2
800	11.1	23.9	16.2	34.7	45.0	66.5	29.9	51.7
3200	86.9	94.7	92.3	97.7	99.9	100	99.6	100
$h = 8$								
200	3.7	8.7	3.9	12.7	18.3	36.1	3.3	14.0
800	20.2	36.8	24.5	45.4	46.4	67.3	27.5	50.0
3200	93.6	97.6	96.3	98.8	99.8	100	99.4	99.9

	SIR		WCT		Simple IRE		Optimal IRE	
n	1	5	1	5	1	5	1	5

Part B: actual level for $H_0 : d = 2$ *vs* $H_1 : d > 2$

$h = 4$

200	0	0.1	0.1	0.4	0.2	1.7	0	1.7
800	0	0.5	0	1.5	0.6	3.6	1.1	4.6
3200	0.2	2.0	0.4	3.9	0.8	4.4	1.1	5.6

$h = 6$

200	0	0.2	0.1	1.3	0.2	2.5	0.2	1.4
800	0.1	1.2	0.1	2.4	0.6	4.0	0.7	3.9
3200	0.4	4.1	0.6	5.0	0.7	5.1	0.8	5.4

$h = 8$

200	0.2	0.8	0.2	1.0	0.9	3.5	0.4	0.9
800	0.2	1.5	0.4	3.2	0.9	4.9	0.8	3.8
3200	0.6	3.6	0.7	3.4	0.6	3.9	0.6	4.4

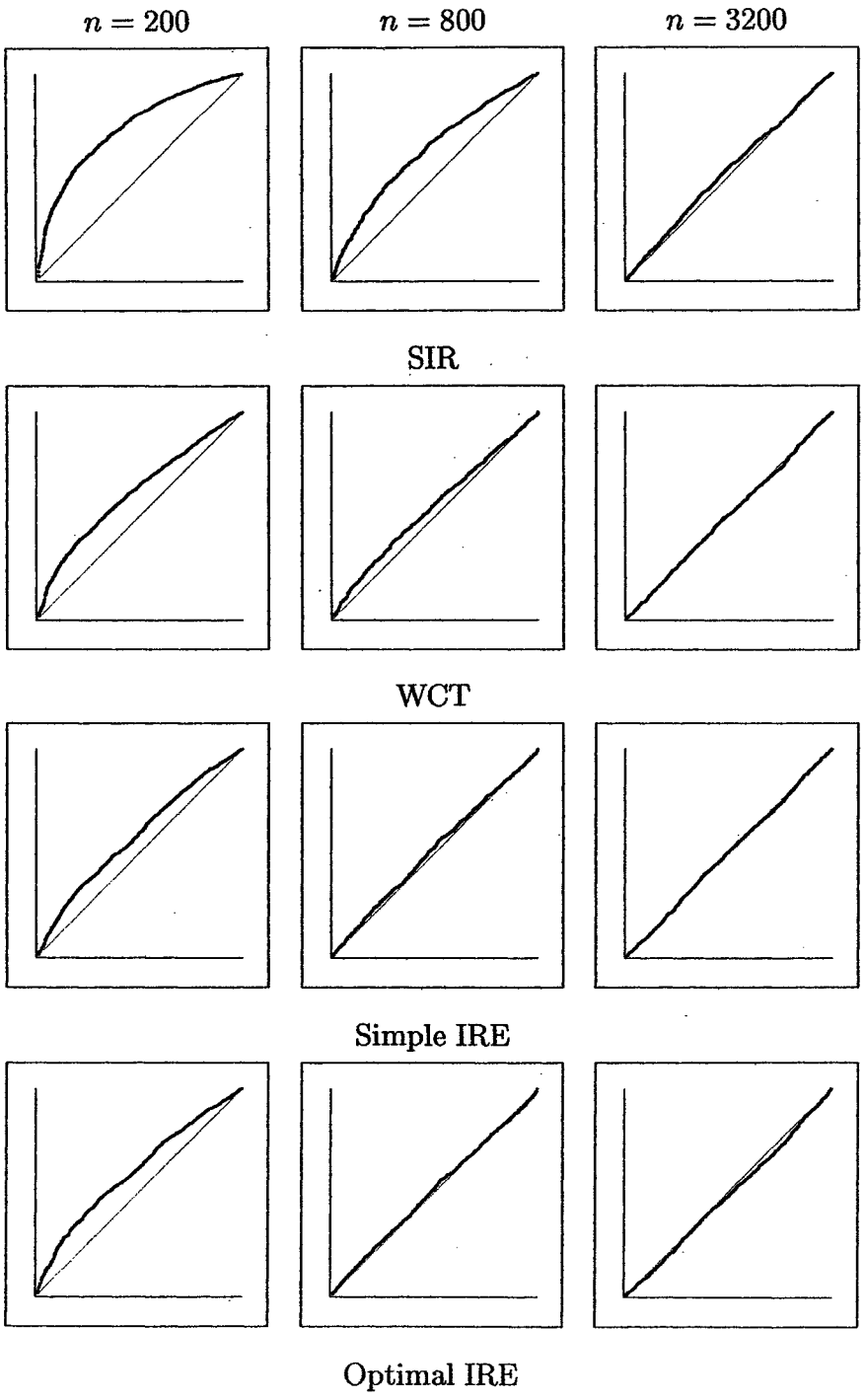


Figure 8.2: Model C: uniform quantile plot of p-values for testing $d = 2$ with $h = 6$.

Part II

Dimension Reduction for Regression Across Multiple Subpopulations

Chapter 9

Sufficient Partial Dimension Reduction

Most methods for estimating the CS are limited to regressions with continuous or many-valued predictors because it is in such cases that a lower dimensional projection $\mathbf{P}_{\mathcal{S}}\mathbf{X}$ of the predictors might provide an effective parsimonious summary. Chiaromonte, Cook, and Li (2002; herein after CCL) removed this limitation by extending the concept of sufficient dimension reduction to include multiple subpopulations identified by a random qualitative predictor W. For example, W might indicate the species or gender of an individual in the population. CCL dealt with the presence of W by developing the idea of a partial dimension reduction subspace, defined as any subspace $\mathcal{S}$ that satisfies the conditional independence statement

$$Y \perp\!\!\!\perp \mathbf{X} | (\mathbf{P}_{\mathcal{S}}\mathbf{X}, W).$$

If the intersection of all partial dimension reduction subspaces is itself a partial dimension reduction subspace it is called the *partial central subspace* (PCS) and denoted as $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$. The PCS is a population meta-parameter like the CS. In particular, if the PCS is known then the regression can again be limited to new sufficient predictors expressed as linear combinations of the original ones: $\boldsymbol{\beta}^T\mathbf{X} = (\boldsymbol{\beta}_1^T\mathbf{X} \ldots \boldsymbol{\beta}_d^T\mathbf{X})^T$, where now the columns of the matrix $\boldsymbol{\beta} = (\boldsymbol{\beta}_1 \ldots \boldsymbol{\beta}_d)$ form a basis for $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$ and $d = \dim(\mathcal{S}^{(W)}_{Y|\mathbf{X}})$.

After describing the foundations of partial dimension reduction, CCL developed methodology called *partial SIR*, which is based on SIR as the name implies. In addition to the usual SIR assumptions, partial SIR requires that the conditional predictor variances $\mathrm{Cov}(\mathbf{X}|W)$ be homogeneous across subpopulations so that $\mathrm{Cov}(\mathbf{X}|W)$ is a constant matrix. Subsequent experience with partial SIR has shown that this *homogeneous covariance condition* is a restriction that should not be neglected in practice. While partial SIR can be an effective method for pursuing sufficient dimension reduction in the presence of a qualitative predictor when the homogeneous covariance condition holds, it can also be quite misleading when the condition fails.

In this chapter, we start with reviewing the methodology of partial SIR along with one illustration. Then in Section 9.3 we develop a dimension reduction method for heterogenous subpopulations–general partial SIR (GP.SIR)–via the minimum discrepancy approach. The algorithm of GP.SIR is similar to that of Simple IRE in Section 7.1. One illustration is presented to demonstrate the potential advantage of this new method. We leave the discussion of

asymptotic properties of GP.SIR to Chapter 10. Extensive simulation study will be presented in Chapter 11.

9.1 Partial SIR

For notational simplicity, we will follow CCL and use $(\mathbf{X}_w, Y_w)$ to indicate a generic pair distributed like $(\mathbf{X}, Y)|(W = w)$ so, for example, $\mathcal{S}_{Y_w|\mathbf{X}_w}$ is the central subspace given $W = w$ and $\mathbf{Z}_w = \boldsymbol{\Sigma}_w^{-\frac{1}{2}}(\mathbf{X}_w - \mathrm{E}[\mathbf{X}_w])$, where $\boldsymbol{\Sigma}_w = \mathrm{Cov}(\mathbf{X}_w) > 0$. The PCS is constructed so that the predictors $\boldsymbol{\beta}^T\mathbf{X}$ are sufficient for every subpopulation, but they might not be necessary for any single subpopulation. In other words, $\mathrm{Span}(\boldsymbol{\beta})$ is a dimension reduction subspace for every subpopulation $Y_w \perp\!\!\!\perp \mathbf{X}_w|\boldsymbol{\beta}^T\mathbf{X}_w$, but it may not be central for any of them. Nevertheless, CCL showed that there is a close connection between the *conditional central subspaces* $\mathcal{S}_{Y_w|\mathbf{X}_w}$ and the partial central subspace:

$$\mathcal{S}_{Y|\mathbf{X}}^{(W)} = \oplus_{w=1}^{K} \mathcal{S}_{Y_w|\mathbf{X}_w}. \tag{9.1}$$

This identity, which does not require any conditions except for the existence of the conditional central subspaces, suggests that $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$ can be estimated by combining dimension reduction across subpopulations. CCL then used (9.1) to develop SIR-type methodology, called *partial SIR*, for inference about the PCS by imposing the condition that the subpopulation covariance matrices are constant, $\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{\mathrm{pool}}$ for all w. Under this *homogeneous covariance condition*, and assuming essentially that the linearity and coverage conditions hold within each subpopulation, they based their partial SIR estimate of $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$

on the implied identity

$$\mathcal{S}_{Y|\mathbf{X}}^{(W)} = \mathbf{\Sigma}_{\text{pool}}^{-\frac{1}{2}} \text{Span}(\text{Cov}(\text{E}[\mathbf{Z}_W|Y_W])). \tag{9.2}$$

In particular, a spectral analysis of a sample version of

$$\text{Cov}(\text{E}[\mathbf{Z}_W|Y_W]) = \text{Cov}(\text{E}[\mathbf{Z}|Y,W])$$

can be used to infer about $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$ in the same way that SIR is used to infer about $\mathcal{S}_{Y|\mathbf{X}}$ by using a spectral analysis of a sample version of $\text{Cov}(\text{E}[\mathbf{Z}|Y])$.

9.2 Lean Body Mass Regression

To help fix ideas and highlight the main issue considered here, we revisit one of the regressions discussed by CCL, which has been seen in Section 1.2. For $n = 202$ athletes at the Australian Institute of Sport, consider the regression of lean body mass L on $p = 5$ continuous or many-values predictors, the logarithms of height, weight, red cell count, white cell count and hemoglobin, represented by $\mathbf{X}$ and gender indicated by $W = m$ or f. While SIR is not directly helpful because of the presence of the qualitative predictor W, this is the kind of regression for which partial SIR was designed. In addition to the usually mild conditions needed for SIR, partial SIR requires that the covariance matrix of $\mathbf{X}$ be the same for males and females, $\mathbf{\Sigma}_m = \mathbf{\Sigma}_f$. Partial SIR can be quite effective for dimension reduction across multiple subpopulations when this homogeneous covariance condition is reasonable, but experience has shown that it can produce misleading results when the condition fails. While departures from the homogeneity condition do not alter the logic of partial SIR, CCL found that they do introduce scaling issues

that affect spectral decompositions.

A p-value of about 0.48 was obtained from the test of $\Sigma_m = \Sigma_f$ described by Anderson (1984, Ch. 10). Thus we felt comfortable assuming homogeneous covariances. Using partial SIR as proposed by CCL we inferred that $\dim(\mathcal{S}^{(W)}_{Y|\mathbf{X}}) = 1$. Table 9.1 shows the test results of partial SIR. A plot of L

Table 9.1: Lean Body Regression

	Partial SIR		
NH: d = m	Test Statistic	D.F.	p-value
0	173.46	30	0
1	22.264	20	0.326

versus the estimated sufficient predictor $\hat{\boldsymbol{\beta}}^T\mathbf{X}$ is shown in Figure 9.1 which is a duplicate of Figure 1.2 for ease of reference. The ordinary least squares fits are shown for males and females as visual aids. The interpretation of the plot is that while males and females have different regressions they both depend on one and the same linear combination of the predictors $\mathbf{X}$. This plot can now be used to guide the remaining analysis, depending on the specific application context.

Without the homogeneous covariance condition, the previous work does not tell us how to apply partial SIR in this regression. However, our results have shown that it is possible to construct useful estimates of $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$ without homogeneous covariances by abandoning the pursuit of SIR-style spectral

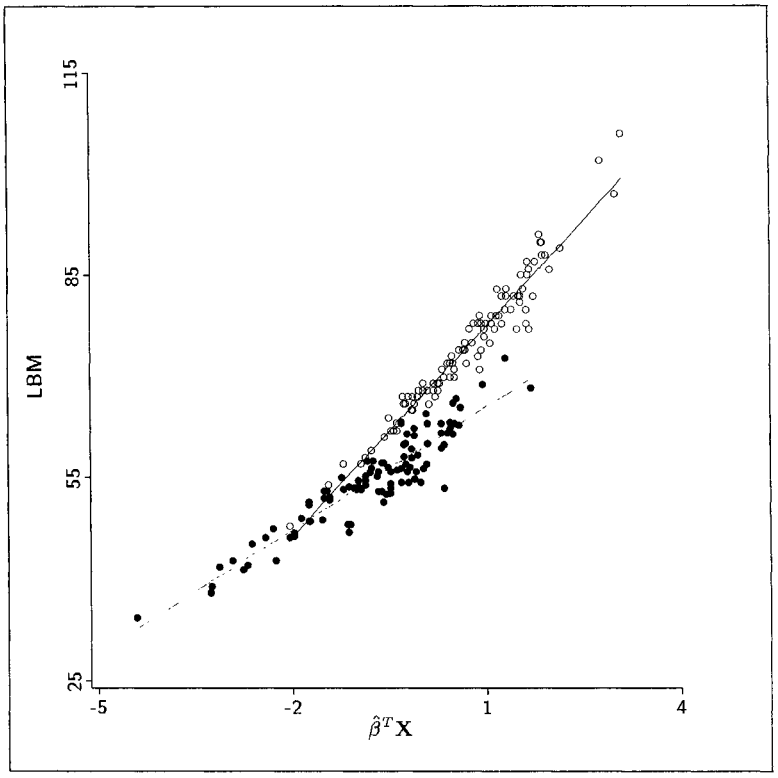


Figure 9.1: Summary plot from application of partial SIR to the lean body mass regression. ○ males. ● females.

decompositions and basing estimation instead on a nonlinear least squares objective function. We describe the new method of estimation, called *general partial SIR*, in the next section and show that it reduces to SIR in the absence of subpopulations and to partial SIR when multiple subpopulations are present and the homogeneous covariance condition holds. In effect, general partial SIR allows the logic of partial SIR to be applied under the same conditions as SIR, without the need for the additional condition of homogeneous covariances.

9.3 General Partial SIR

In this section, we develop dimension reduction methods for regression across multiple subpopulations removing the limitation of the homogenous covariance condition in partial SIR. Instead of constructing a kernel matrix, we proceed via the minimum discrepancy approach by optimization an objective function.

We consider a regression $Y \in \mathbb{R}^1$ on $\mathbf{X} \in \mathbb{R}^p$ across multiple subpopulations indicated by a random variable W with support $\{1, 2, ..., K\}$. Assuming that the linearity and coverage conditions hold for subpopulation w,

$$\mathcal{S}_{Y_w|\mathbf{X}_w} = \oplus_{y=1}^{h_w} \text{Span}\{\boldsymbol{\xi}_{wy}\},$$

where

$$\boldsymbol{\xi}_{wy} = \boldsymbol{\Sigma}_w^{-1}(\text{E}[\mathbf{X}_w|Y_w = y] - \text{E}[\mathbf{X}_w]),$$

h_w is the number of slices in subpopulation w and for consistency with previous notation we let $h = \sum_w h_w$. It follows from (9.1) that

$$\mathcal{S}_{Y|\mathbf{X}}^{(W)} = \oplus_{w=1}^{K} \oplus_{y=1}^{h_w} \text{Span}\{\boldsymbol{\xi}_{wy}\}. \tag{9.3}$$

Recalling that the columns of the $p \times d$ matrix $\boldsymbol{\beta}$ form a basis for $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$, (9.3) implies that for each $\boldsymbol{\xi}_{wy}$ we can find a vector $\boldsymbol{\gamma}_{wy}$ so that

$$\boldsymbol{\xi}_{wy} = \boldsymbol{\beta}\boldsymbol{\gamma}_{wy}.$$

This relation suggests that a basis for $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$ might be estimated by minimizing an average discrepancy between a sample version of $\boldsymbol{\xi}_{wy}$ and the estimate of $\boldsymbol{\beta}\boldsymbol{\gamma}_{wy}$. To develop this idea, we first establish notation to express a sample

version of $\boldsymbol{\xi}_{wy}$ in terms of moment estimates of its components. For later use, we let $\boldsymbol{\gamma}_w = (\boldsymbol{\gamma}_{w1}, \ldots, \boldsymbol{\gamma}_{wh_w})$ and define the $d \times h$ matrix

$$\boldsymbol{\gamma} \equiv (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K). \tag{9.4}$$

Suppose we have a random sample of size n for $(\mathbf{X}, Y, W)$ from the total population. There are n_w points in subpopulation w, among which n_{wy} points have $Y_w = y$. Let $p_w = \Pr(W = w)$ and let $\hat{p}_w = n_w/n$ be the observed fraction for subpopulation w. Similarly, let $f_{wy} = \Pr(Y_w = y)$ and let $\hat{f}_{wy} = n_{wy}/n_w$ denote the corresponding observed fraction. Using notation often associated with an analysis of variance, let $\mathbf{X}_{wyi}$ denote the i-th observation on $\mathbf{X}$ in slice y of subpopulation w, let $\bar{\mathbf{X}}_{w\bullet\bullet}$ be the average in subpopulation w,

$$\bar{\mathbf{X}}_{w\bullet\bullet} = \frac{1}{n_w} \sum_{y=1}^{h_w} \sum_{i=1}^{n_{wy}} \mathbf{X}_{wyi},$$

and let $\bar{\mathbf{X}}_{wy\bullet}$ be the average of n_{wy} points in slice y of subpopulation w. Letting $\hat{\boldsymbol{\Sigma}}_w$ denote the sample covariance of $\mathbf{X}$ in subpopulation w, a sample version of $\boldsymbol{\xi}_{wy}$ can now be represented as

$$\hat{\boldsymbol{\xi}}_{wy} = \hat{\boldsymbol{\Sigma}}_w^{-1}(\bar{\mathbf{X}}_{wy\bullet} - \bar{\mathbf{X}}_{w\bullet\bullet}) = \hat{\boldsymbol{\Sigma}}_w^{-\frac{1}{2}} \bar{\mathbf{Z}}_{wy\bullet},$$

where $\bar{\mathbf{Z}}_{wy\bullet} = \hat{\boldsymbol{\Sigma}}_w^{-\frac{1}{2}}(\bar{\mathbf{X}}_{wy\bullet} - \bar{\mathbf{X}}_{w\bullet\bullet})$. Assuming that the dimension d of $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$ is known, we propose to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_{wy}$ by minimizing the nonlinear weighted least squares discrepancy function

$$F_d(\mathbf{B}, \mathbf{C}) \equiv \sum_{w=1}^{K} \hat{p}_w \sum_{y=1}^{h_w} \hat{f}_{wy} (\hat{\boldsymbol{\xi}}_{wy} - \mathbf{B}\mathbf{C}_{wy})^T \hat{\boldsymbol{\Sigma}}_w (\hat{\boldsymbol{\xi}}_{wy} - \mathbf{B}\mathbf{C}_{wy}) \tag{9.5}$$

so that

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \arg_{\mathbf{B},\mathbf{C}} \min F_d(\mathbf{B}, \mathbf{C}) \tag{9.6}$$

where the minimization is over $\mathbf{C} \in \mathbb{R}^{d \times h}$ with columns $\mathbf{C}_{wy}$, and $\mathbf{B} \in \mathbb{R}^{p \times d}$. We call this method *general partial SIR* (GP.SIR).

The discrepancy function $F_d(\mathbf{B}, \mathbf{C})$ converges almost surely to its population version

$$\begin{aligned} \tilde{F}_d(\mathbf{B}, \mathbf{C}) &\equiv \sum_{w=1}^{K} p_w \sum_{y=1}^{h_w} f_{wy} (\boldsymbol{\xi}_{wy} - \mathbf{B}\mathbf{C}_{wy})^T \boldsymbol{\Sigma}_w (\boldsymbol{\xi}_{wy} - \mathbf{B}\mathbf{C}_{wy}) \\ &= \mathrm{E}(\boldsymbol{\xi}_{WY} - \mathbf{B}\mathbf{C}_{WY})^T \boldsymbol{\Sigma}_W (\boldsymbol{\xi}_{WY} - \mathbf{B}\mathbf{C}_{WY}). \end{aligned}$$

Under the linearity and coverage conditions, $(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \arg\min \tilde{F}_d(\mathbf{B}, \mathbf{C})$ so GP.SIR provides a Fisher consistent estimate of a basis for $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$. The minimizers of $\tilde{F}_d$ are not unique, but that is not a problem since any basis for $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$ will suffice. In the algorithm described later in Section 9.4 we handle the uniqueness issue by imposing orthogonality and length constraints on the columns of $\hat{\boldsymbol{\beta}}$.

The GP.SIR estimate of $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$ reduces to the partial SIR estimate when the pooled covariance matrix

$$\hat{\boldsymbol{\Sigma}}_{\text{pool}} = \sum_{w=1}^{K} \hat{p}_w \hat{\boldsymbol{\Sigma}}_w$$

is used in place of $\hat{\boldsymbol{\Sigma}}_w$ in $\hat{\boldsymbol{\xi}}_{wy}$ and in the inner product of the discrepancy function. To show this, we replace $\hat{\boldsymbol{\Sigma}}_w$ with $\hat{\boldsymbol{\Sigma}}_{\text{pool}}$ in (9.6). Then after a little algebra we find that $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}_{wy})$ can be obtained from the value of $(\dot{\mathbf{B}}, \mathbf{C}_{wy})$ that minimizes

$$\sum_{w=1}^{K} \sum_{y=1}^{h_w} \frac{n_{wy}}{n} (\bar{\mathbf{Z}}_{wy\bullet} - \dot{\mathbf{B}}\mathbf{C}_{wy})^T (\bar{\mathbf{Z}}_{wy\bullet} - \dot{\mathbf{B}}\mathbf{C}_{wy}),$$

where now $\bar{\mathbf{Z}}_{wy\bullet} = \hat{\mathbf{\Sigma}}_{\text{pool}}^{-\frac{1}{2}}[\bar{\mathbf{X}}_{wy\bullet} - \bar{\mathbf{X}}_{w\bullet\bullet}]$ and $\dot{\mathbf{B}} = \hat{\mathbf{\Sigma}}_{\text{pool}}^{\frac{1}{2}}\mathbf{B}$. After minimizing over $\mathbf{C}_{wy}$ for a fixed $\dot{\mathbf{B}}$, we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \hat{\mathbf{\Sigma}}_{\text{pool}}^{-\frac{1}{2}} \arg_{\dot{\mathbf{B}}} \min \sum_{w,y} \frac{n_{wy}}{n} \|\bar{\mathbf{Z}}_{wy\bullet} - \dot{\mathbf{B}}(\dot{\mathbf{B}}^T\dot{\mathbf{B}})^{-1}\dot{\mathbf{B}}^T\bar{\mathbf{Z}}_{wy\bullet}\|^2 \\
&= \hat{\mathbf{\Sigma}}_{\text{pool}}^{-\frac{1}{2}} \arg_{\dot{\mathbf{B}}} \min \sum_{w,y} \frac{n_{wy}}{n} \bar{\mathbf{Z}}_{wy\bullet}^T \mathbf{Q}_{\dot{\mathbf{B}}} \bar{\mathbf{Z}}_{wy\bullet} \\
&= \hat{\mathbf{\Sigma}}_{\text{pool}}^{-\frac{1}{2}} \arg_{\dot{\mathbf{B}}} \min \operatorname{trace}\Big(\sum_{w,y} \frac{n_{wy}}{n} \bar{\mathbf{Z}}_{wy\bullet}\bar{\mathbf{Z}}_{wy\bullet}^T \mathbf{Q}_{\dot{\mathbf{B}}}\Big) \\
&= \hat{\mathbf{\Sigma}}_{\text{pool}}^{-\frac{1}{2}} \arg_{\dot{\mathbf{B}}} \min \operatorname{trace}(\widehat{\mathbf{M}}_{\text{PSIR}}\mathbf{Q}_{\dot{\mathbf{B}}}),
\end{aligned}$$

where $\mathbf{Q}_{(\cdot)} = I - \mathbf{P}_{(\cdot)}$ and

$$\widehat{\mathbf{M}}_{\text{PSIR}} = \sum_{w=1}^{h} \hat{p}_w \sum_{y=1}^{h_w} \hat{f}_{wy} \bar{\mathbf{Z}}_{wy\bullet}\bar{\mathbf{Z}}_{wy\bullet}^T \tag{9.7}$$

is the pooled sample covariance matrix of the slice means, which is the estimate of $\text{Cov}(\text{E}[\mathbf{Z}_W|Y_W])$ (cf. (9.2)) used by CCL. Thus, by Lemma 5 in Appendix B, $\hat{\boldsymbol{\beta}} = \hat{\mathbf{\Sigma}}_{\text{pool}}^{-\frac{1}{2}}[\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_d]$, where $\hat{\boldsymbol{\mu}}_i$'s are the eigenvectors of $\widehat{\mathbf{M}}_{\text{PSIR}}$ that correspond to its d largest eigenvalues. It follows immediately from this result that GP.SIR reduces to SIR, $\widehat{\mathbf{M}}_{\text{SIR}} = \widehat{\mathbf{M}}_{\text{PSIR}}$, when there is only one subpopulation ($K = 1$).

There are two essential tasks left to develop GP.SIR as methodology. The first is to describe a numerical algorithm for the minimization in (9.6). The other is to find an appropriate test statistic for dimensionality and its asymptotic distribution. The algorithm is discussed in Section 9.4. Inference is addressed in Chapter 10.

9.4 Algorithm for GP.SIR

Like many other dimension reduction methods, SIR and partial SIR adopt a spectral approach based on finding a consistently estimable kernel matrix – $\mathrm{Cov}(\mathrm{E}[\mathbf{Z}|Y])$ or $\mathrm{Cov}(\mathrm{E}[\mathbf{Z}|Y,W])$ – that spans either $\mathcal{S}_{Y|\mathbf{X}}$ or $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$. The eigenvectors of the sample kernel matrix ($\widehat{\mathrm{M}}_{\mathrm{SIR}}$ or $\widehat{\mathrm{M}}_{\mathrm{PSIR}}$) corresponding to its eigenvalues that are inferred to be nonzero in the population form the estimate of the target subspace. However, when we have heterogenous subpopulations, the minimization of the discrepancy function (9.6) no longer reduces to a spectral decomposition problem, which is a generalization that allows us to get around the limitations of the previous approach.

From (9.6), GP.SIR is based on the minimization of the discrepancy function

$$\begin{aligned} F_d(\mathbf{B},\mathbf{C}) &= \sum_{w,y} \frac{n_{wy}}{n}(\hat{\boldsymbol{\xi}}_{wy} - \mathbf{B}\mathbf{C}_{wy})^T \hat{\boldsymbol{\Sigma}}_w (\hat{\boldsymbol{\xi}}_{wy} - \mathbf{B}\mathbf{C}_{wy}) \\ &= \sum_{w,y} (\hat{\mathbf{V}}_{wy}\hat{\boldsymbol{\xi}}_{wy} - \hat{\mathbf{V}}_{wy}\mathbf{B}\mathbf{C}_{wy})^T (\hat{\mathbf{V}}_{wy}\hat{\boldsymbol{\xi}}_{wy} - \hat{\mathbf{V}}_{wy}\mathbf{B}\mathbf{C}_{wy}), \end{aligned}$$

where $\hat{\mathbf{V}}_{wy} = \sqrt{n_{wy}/n}\hat{\boldsymbol{\Sigma}}_w^{1/2}$. This is a special case of finding the values of $\mathbf{B}$ and $\mathbf{C}$ which minimize a generic discrepancy function of the form

$$H(\mathbf{B},\mathbf{C}) = \sum_{j=1}^{h} (\boldsymbol{\alpha}_j - \mathbf{S}_j\mathbf{B}\mathbf{C}_j)^T (\boldsymbol{\alpha}_j - \mathbf{S}_j\mathbf{B}\mathbf{C}_j),$$

where $\mathbf{B} \in \mathbb{R}^{p\times d}$, $\mathbf{C} = (\mathbf{C}_1, \cdots, \mathbf{C}_h) \in \mathbb{R}^{d\times h}$, $\boldsymbol{\alpha}_j \in \mathbb{R}^p$ and $\mathbf{S}_j \in \mathbb{R}^{p\times p}$. The $\mathbf{S}_j$'s are positive definite. All $\boldsymbol{\alpha}_j$ and $\mathbf{S}_j$ are fixed in the minimization algorithm. This general discrepancy function is exactly the same as in the discussion of algorithm of Simple IRE in Section 7.1. We use alternate least

squares method again for the minimization. There is one thing we should notice: It seems best to avoid using the eigenvectors from partial SIR as starting vectors for $\mathbf{B}$ since they can result in the algorithm being trapped at a local minimum when the subpopulation covariance matrices are quite different. Our experience is that constant initial values generally do well.

9.5 Illustration

We present a simple example to compare partial SIR and GP.SIR. We generated $\mathbf{X}$ from two multivariate normal populations with mean 0 and covariance matrices $\mathbf{\Sigma}_1 = \text{diag}\{9,9,1,1,1\}$ and $\mathbf{\Sigma}_2 = \text{diag}\{1,1,9,9,9\}$. Each sample has 400 data points. The response Y was generated in both populations according to the model

$$Y = X_2 + X_4 + \exp[X_1 + X_3] + 0.2\,\epsilon,$$

where ϵ is an independent standard normal variate. The PCS is 2-dimensional with true directions $\boldsymbol{\beta}_1 = (0,1,0,1,0)^T$ and $\boldsymbol{\beta}_2 = (1,0,1,0,0)^T$. We applied partial SIR to this data with 4 slices in each population. Partial SIR concludes $d > 3$ at 0.001 significance level based on its test statistics, and the first two important directions are approximately

$$\mathbf{b}_{p1} = (0.59, 0.37, 0.62, 0.36, 0.01)^T$$

and

$$\mathbf{b}_{p2} = (0.59, 0.34, -0.56, -0.46, 0.05)^T.$$

The first direction falls close to $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$: $\mathbf{b}_{p1}^T\mathbf{X}$ has a multiple correlation of 0.999 with $\boldsymbol{\beta}_1^T\mathbf{X}$ and $\boldsymbol{\beta}_2^T\mathbf{X}$. But the second direction is nearly orthogonal

to $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$. The multiple correlation between $\mathbf{b}_{p2}^T\mathbf{X}$ and $(\boldsymbol{\beta}_1^T\mathbf{X}, \boldsymbol{\beta}_2^T\mathbf{X})$ is 0.095. It seems that partial SIR confuses the covariance difference as part of the partial central space. Partial SIR considers

$$\boldsymbol{\Sigma}_{\text{pool}}^{-1}\mathrm{E}[\bar{\mathbf{X}}_{wy\bullet} - \bar{\mathbf{X}}_{w\bullet\bullet}] = \boldsymbol{\Sigma}_{\text{pool}}^{-1}\boldsymbol{\Sigma}_w\boldsymbol{\Sigma}_w^{-1}\mathrm{E}[\bar{\mathbf{X}}_{wy\bullet} - \bar{\mathbf{X}}_{w\bullet\bullet}] = \boldsymbol{\Sigma}_{\text{pool}}^{-1}\boldsymbol{\Sigma}_w\boldsymbol{\beta}\boldsymbol{\gamma}_{wy}.$$

By construction, in this example, $\boldsymbol{\Sigma}_{\text{pool}} = 5\mathbf{I}$. Therefore, the partial SIR discrepancy function will estimate the space $\oplus_{y,w}\text{Span}\{\boldsymbol{\Sigma}_w\boldsymbol{\beta}\boldsymbol{\gamma}_{wy}\}$, which is 4 dimensional.

On the other hand, given $d = 2$, GP.SIR estimates the directions as

$$\mathbf{b}_{gp1} = (0.69, 0.23, 0.675, 0.12, 0.01)^T$$

and

$$\mathbf{b}_{gp2} = (-0.21, 0.62, -0.13, 0.74, -0.05)^T.$$

The multiple correlation with $\beta_1^T\mathbf{X}$ and $\beta_2^T\mathbf{X}$ is 0.997 for $\mathbf{b}_{gp1}^T\mathbf{X}$ and is 0.993 for $\mathbf{b}_{gp2}^T\mathbf{X}$. Thus, the GP.SIR estimates are much closer to the true space than partial SIR estimates.

There is one key issue remaining, which is how to infer about the dimension of the PCS when using GP.SIR. Here we can also benefit from the minimum discrepancy approach because the minimum value can be used to construct a test statistic for dimensionality. The construction of test statistics and inference are the topics of the next chapter.

Chapter 10

Inference about Partial Dimension Reduction

We saw at the end of Section 9.3 that when there is only a single subpopulation ($K = 1$) and $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$ is known, minimization of the discrepancy function $F_d(\mathbf{B}, \mathbf{C})$ (9.5) results in the SIR estimate of $\mathcal{S}_{Y|\mathbf{X}}$. In this case the minimum value $\hat{F}_d$ of F_d is

$$\hat{F}_d = \sum_{j=d+1}^{p} \hat{\lambda}_j$$

where $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p$ are the eigenvalues of $\widehat{\mathbf{M}}_{\text{SIR}}$ defined in (5.1). The usual SIR test statistic for testing $d = m$ versus $d > m$, where $m < p$, is simply $n\hat{F}_m$, with relatively large values resulting in rejection.

Assuming that $\mathbf{X}$ has a multivariate normal distribution and implicitly assuming the coverage condition, Li (1991) proved that the null distribution of $n\hat{F}_d$ is asymptotically chi-squared with $(p-d)(h-d-1)$ degrees of freedom.

Bura and Cook (2001b) proved that $n\hat{F}_d$ has the same asymptotic distribution under the *coverage* and *linearity conditions* plus the *marginal covariance condition* $\mathrm{Cov}(\mathbf{Z}|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{Z}}}\mathbf{Z}) = \mathbf{Q}_{\mathcal{S}_{Y|\mathbf{Z}}}$ where $\mathbf{Q}_{\mathcal{S}_{Y|\mathbf{Z}}} = \mathbf{I}_p - \mathbf{P}_{\mathcal{S}_{Y|\mathbf{Z}}}$ projects onto the orthogonal complement of $\mathcal{S}_{Y|\mathbf{Z}}$. This condition is equivalent to requiring that $\mathrm{Cov}(\mathbf{Z}|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{Z}}}\mathbf{Z})$ be a nonrandom matrix and, like the linearity condition, it applies to the marginal distribution of $\mathbf{Z}$. Normality of $\mathbf{X}$ implies the linearity and marginal covariance conditions, but not the coverage condition. Bura and Cook (2001b) also proved that in general $n\hat{F}_d$ is asymptotically distributed as a weighted sum of independent chi-squared random variables and showed how to construct consistent estimates of the weights for use in practice (cf. Chapter 6).

There are parallel results for partial SIR. The partial SIR statistic proposed by CCL for the hypothesis $d = m$ versus $d > m$ is again proportional to the minimum value of F_m:

$$\hat{F}_m = \sum_{j=m+1}^{p} \hat{\alpha}_j$$

where $\hat{\alpha}_1 \geq \ldots \geq \hat{\alpha}_p$ are the eigenvalues of $\widehat{\mathbf{M}}_{\mathrm{PSIR}}$ defined in (9.7). CCL showed in effect that if subpopulation coverage and linearity hold, if the subpopulation covariance matrices are homogeneous ($\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{\mathrm{pool}}$) and if $\mathrm{Cov}(\mathbf{Z}_w|\mathbf{P}_{\mathcal{S}^{(w)}_{Y|\mathbf{Z}}}\mathbf{Z}_w)$ is a nonrandom matrix given w, then in partial SIR applications, $n\hat{F}_d$ is asymptotically distributed as a chi-squared random variable with $(p-d)(h-d-K)$ degrees of freedom. Although not emphasized in the main part of their article, they also showed in an appendix that without these conditions $n\hat{F}_d$ is distributed asymptotically as a linear combination of

independent chi-squared random variables. We will revisit the asymptotics of partial SIR in Section 10.4.

In SIR and partial SIR applications, d is often estimated by hypothesis testing using the statistic $n\hat{F}_m$: Starting with $m = 0$, test the hypothesis $d = m$ versus $d > m$. If the test is rejected, increment m by one and test again, stopping with the first nonsignificant result. This type of procedure is fairly common for estimating the dimension of a subspace (see, for example, Rao 1965, p. 472).

Since $n\hat{F}_m$ is a generalized version of the test statistics for SIR and partial SIR, we propose to use it to test the hypothesis $d = m$ versus $d > m$ in GP.SIR. This requires the asymptotic distribution of $n\hat{F}_d$ or perhaps a nonparametric alternative. Here we follow the asymptotic route.

10.1 Asymptotic Distribution of the Test Statistic in GP.SIR

A little setup is necessary before we can report the asymptotic distribution of $n\hat{F}_d$ in GP.SIR. Conditioning on subpopulation w for the time being, define the random variable J_{wy} to equal 1 if $Y_w = y$ and 0 otherwise. Given w, $\mathrm{E}(J_{wy}) = \Pr(Y_w = y) = f_{wy}$. Define

$$\varepsilon_{wy} = J_{wy} - f_{wy} - \mathbf{Z}_w^T \mathrm{E}[\mathbf{Z}_w J_{wy}]$$

to be the population residuals from the ordinary least square fit of J_{wy} on $\mathbf{Z}_w$, still conditioning on w. Let $\boldsymbol{\varepsilon}_w = (\varepsilon_{w1}, \ldots, \varepsilon_{wh_w})^T$ denote the $h_w \times 1$ vector of residuals, one for each slice, for a typical observation from subpopulation w, and let $\mathbf{f}_w = (f_{w1}, f_{w2}, \ldots, f_{wh_w})^T$, where $f_{wy} = \Pr(Y_w = y)$, and let $\mathbf{D}_{\mathbf{f}_w} \equiv \text{diag}\{f_{wy}\}$ be the $h_w \times h_w$ diagonal matrix with the elements of $\mathbf{f}_w$ on the diagonal. With this notation we can now define the following $ph_w \times ph_w$ covariance matrix for subpopulation w:

$$\boldsymbol{\Omega}_w = \text{Cov}(\mathbf{D}_{\mathbf{f}_w}^{-1/2} \boldsymbol{\varepsilon}_w \otimes \mathbf{Z}_w). \tag{10.1}$$

We then arrange these covariance matrices in a $ph \times ph$ block diagonal matrix $\boldsymbol{\Omega} \equiv \text{diag}\{\boldsymbol{\Omega}_w\}$, which is one component that we need to describe the asymptotic distribution of $n\hat{F}_d$.

We also need the $ph \times ph$ block diagonal matrix

$$\mathbf{V} \equiv \text{diag}\{p_w \mathbf{D}_{\mathbf{f}_w}^{-1} \otimes \boldsymbol{\Sigma}_w\} \tag{10.2}$$

and the $ph \times (p+h)d$ matrix

$$\boldsymbol{\Delta} \equiv (\boldsymbol{\nu}^T \otimes \mathbf{I}_p, \mathbf{I}_h \otimes \boldsymbol{\beta}). \tag{10.3}$$

The matrix $\boldsymbol{\Delta}$ is the Jacobian matrix for a vectorized version of the discrepancy function:

$$\boldsymbol{\Delta} = \left(\frac{\partial \,\text{vec}(\mathbf{BC})}{\partial \,\text{vec}(\mathbf{B})}, \frac{\partial \,\text{vec}(\mathbf{BC})}{\partial \,\text{vec}(\mathbf{C})} \right).$$

evaluated at $(\boldsymbol{\beta}, \boldsymbol{\nu})$, where $\mathbf{B} \in \mathbb{R}^{p \times d}$, $\mathbf{C} \in \mathbb{R}^{d \times h}$, $\boldsymbol{\beta}$ is a basis for $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$ as defined previously, $\boldsymbol{\nu} = \boldsymbol{\gamma}\text{diag}\{\mathbf{D}_{\mathbf{f}_w}\}$, and $\boldsymbol{\gamma}$ is as defined in (9.4). Notice that the definitions of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\nu}$ are different from those in Part I. Here

$\mathbf{V}$ is the inner product matrix for the same vectorized version of F_d. Finally, letting $\mathbf{\Phi} = \mathbf{V}^{\frac{1}{2}}\mathbf{\Delta}$, the asymptotic distribution of $n\hat{F}_d$ is given in the following theorem.

Theorem 4. *Assume that the data* $(\mathbf{X}_i, Y_i, W_i)$, $i = 1, \ldots, n$, *are a simple random sample of* $(\mathbf{X}, Y, W)$. *Let* $\mathcal{S}_\xi = \oplus_{w=1}^{K} \oplus_{y=1}^{h_w} \text{Span}\{\boldsymbol{\xi}_{wy}\}$, *let* $d = \dim(\mathcal{S}_{\boldsymbol{\xi}})$ *and let* $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \arg_{\mathbf{B},\mathbf{C}} \min F_d(\mathbf{B}, \mathbf{C})$ *where* F_d *as defined previously in (9.5):*

$$F_d(\mathbf{B}, \mathbf{C}) = \sum_{w=1}^{K} \hat{p}_w \sum_{y=1}^{h_w} \hat{f}_{wy} (\hat{\boldsymbol{\xi}}_{wy} - \mathbf{B}\mathbf{C}_{wy})^T \hat{\mathbf{\Sigma}}_w (\hat{\boldsymbol{\xi}}_{wy} - \mathbf{B}\mathbf{C}_{wy})$$

Then

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

2. *as* $n \to \infty$,

$$n\hat{F}_d \xrightarrow{\mathcal{D}} \sum_{i=1}^{ph} \lambda_i \chi_i^2(1)$$

where $\{\chi_i^2(1)\}$ *are independent chi-squared random variables each with 1 degree of freedom and* $\{\lambda_1 \geq \ldots \geq \lambda_{ph}\}$ *are the eigenvalues of* $\mathbf{Q}_{\mathbf{\Phi}}\mathbf{\Omega}\mathbf{Q}_{\mathbf{\Phi}}$.

This theorem is quite general, requiring none of the special conditions discussed previously. The value $\hat{\boldsymbol{\beta}}$ of $\mathbf{B}$ that minimizes the discrepancy function $F_d(\mathbf{B}, \mathbf{C})$ always provides a consistent estimate of a basis for $\mathcal{S}_{\boldsymbol{\xi}}$, and this theorem allows us to test hypothesis about its dimension. However, without some of the special conditions, $\mathcal{S}_{\boldsymbol{\xi}}$ might not be a useful population parameter and therefore tests on its dimension might not be of interest.

If the linearity condition holds within subpopulations then $\mathcal{S}_{\boldsymbol{\xi}} \subseteq \mathcal{S}_{Y|\mathbf{X}}^{(W)}$. The subspace spanned by $\hat{\boldsymbol{\beta}}$ is still a consistent estimate of $\mathcal{S}_{\boldsymbol{\xi}}$, which is now a

subspace of the PCS. In this case we are able to use Theorem 4 to infer about a possibly proper subset of the PCS. If the linearity and coverage conditions both hold, then we are back to the main line introduced at the beginning of Section 9.3. In this case, as previously pointed out in (9.3), $\mathcal{S}_{\boldsymbol{\xi}} = \mathcal{S}^{(W)}_{Y|\mathbf{X}}$, and we can use Theorem 4 to infer about the full PCS.

As similar to that of Theorem 3, the proof of Theorem 4 hinges on Proposition 1 by Shapiro (1986) on the asymptotics of overparameterized structural models (cf. Section 3.3.1).

Proof of Theorem 4

To use Shapiro's results we first write our discrepancy function F_d defined at (9.5) in the form of the general discrepancy function of Proposition 1.

Using the definitions of $\boldsymbol{\xi}_{wy}$ and $\hat{\boldsymbol{\xi}}_{wy}$ established at the beginning of Section 9.3, define $\boldsymbol{\zeta}_{wy} = f_{wy}\boldsymbol{\xi}_{wy}$ with corresponding sample version $\hat{\boldsymbol{\zeta}}_{wy} = \hat{f}_{wy}\hat{\boldsymbol{\xi}}_{wy}$. Define also, $\boldsymbol{\zeta}_w = (\boldsymbol{\zeta}_{w1}, \ldots, \boldsymbol{\zeta}_{wh_w})$ and

$$\boldsymbol{\zeta} \;=\; (\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_K) \tag{10.4}$$

with corresponding sample versions $\hat{\boldsymbol{\zeta}}_w$ and $\hat{\boldsymbol{\zeta}}$. Then for fixed dimension d the GP.SIR discrepancy function can be written as

$$F_d(\mathbf{B}, \mathbf{C}) = (\,\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\mathbf{BC}))^T \hat{\mathbf{V}} (\,\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \mathrm{vec}(\mathbf{BC})) \tag{10.5}$$

where $\hat{\mathbf{V}} = \mathrm{diag}\{\hat{p}_w \mathbf{D}^{-1}_{\hat{\mathbf{f}}_w} \otimes \hat{\boldsymbol{\Sigma}}_w\}$. The argument $\mathbf{C}$ used here corresponds to the argument $\mathbf{C}$ in (9.5) times $\mathrm{diag}\{\mathbf{D}_{\mathbf{f}_w}\}$. The same relationship holds between

$\boldsymbol{\nu}$ and $\boldsymbol{\gamma}$ mentioned following (10.3), $\boldsymbol{\nu} = \boldsymbol{\gamma}\text{diag}\{\mathbf{D}_{\mathbf{f}_w}\}$. The argument $\mathbf{B}$ that stands for a basis of $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$ is the same in both versions. That F_d is a version of Shaprio's discrepancy function H which can be seen by setting

$$\begin{aligned} \theta &= \begin{pmatrix} \text{vec}(\mathbf{B}) \\ \text{vec}(\mathbf{C}) \end{pmatrix} \in \mathbb{R}^{d(p+h)} \\ g(\theta) &= \text{vec}(\mathbf{BC}) \in \mathbb{R}^{ph} \\ \boldsymbol{\tau}_n &= \text{vec}(\hat{\boldsymbol{\zeta}}) \\ g(\theta_0) &= \text{vec}(\boldsymbol{\beta}\boldsymbol{\nu}) \end{aligned}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ is in general a basis for $\mathcal{S}_{\boldsymbol{\xi}}$ and $\boldsymbol{\nu} \in \mathbb{R}^{d\times h}$. With these associations it is straightforward to verify that $\boldsymbol{\Delta} = (\boldsymbol{\nu}^T \otimes \mathbf{I}_p, \mathbf{I}_h \otimes \boldsymbol{\beta})$ as defined previously in (10.3).

We next address the issue that the inner product matrix in Proposition 1 is assumed to be known, while the inner product matrix in F_d is estimated. Since $\hat{\mathbf{V}}$ converges to $\mathbf{V}$ in probability, where $\mathbf{V}$ is as defined in (10.2), it follows from Lemma 2 in Section 3.3.1 that the asymptotic distribution of $n\hat{F}_d$ is the same whether we use $\mathbf{V}$ or $\hat{\mathbf{V}}$ as the inner product matrix. Thus we now replace $\hat{\mathbf{V}}$ with $\mathbf{V}$ in F_d. Since $\mathbf{V} > 0$, conditions 2, including properties p1-p4, and 4 of Proposition 1 are met. Condition 3 is met also since $g(\theta)$ is analytic. See Shapiro (1986) for details about regular points. With this, we have verified all of the conditions of Proposition 1, except for asymptotic normality.

The strategy to showing asymptotic normality is to decompose $\sqrt{n}(\text{vec}(\hat{\boldsymbol{\zeta}}) - \text{vec}(\boldsymbol{\beta}\boldsymbol{\nu}))$ as a summation of i.i.d. observations plus a remainder converging

to 0 in probability. Then, by the central limit theorem, we obtain the conclusion. This is the same strategy we used in Section 3.1. Now let us focus on a generic wth population. For notational simplicity, we drop w from the subscripts and will restore it when we reach the conclusion. The subscript y still denotes a slice in the subpopulation.

With w suppressed, recall that $\bar{\mathbf{X}}_{y\bullet}$ is the average of n_y observations in the yth slice and $\bar{\mathbf{X}}_{\bullet\bullet}$ is the grand average of all n observations. Letting $\boldsymbol{\mu}_y = \mathrm{E}[\bar{\mathbf{X}}_{y\bullet}]$ and $\boldsymbol{\mu} = \mathrm{E}[\bar{\mathbf{X}}_{\bullet\bullet}]$, consider

$$\begin{aligned}
& \sqrt{n}(\hat{\boldsymbol{\zeta}}_y - \boldsymbol{\zeta}_y) \\
= \; & \sqrt{n}\hat{f}_y\hat{\boldsymbol{\Sigma}}^{-1}(\bar{\mathbf{X}}_{y\bullet} - \bar{\mathbf{X}}_{\bullet\bullet}) - \sqrt{n}f_y\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_y - \boldsymbol{\mu}) \\
= \; & \sqrt{n}(\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1})f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu}) + \sqrt{n}\boldsymbol{\Sigma}^{-1}[\hat{f}_y(\bar{\mathbf{X}}_{y\bullet} - \bar{\mathbf{X}}_{\bullet\bullet}) - f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu})] \\
& + \sqrt{n}(\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1})[\hat{f}_y(\bar{\mathbf{X}}_{y\bullet} - \bar{\mathbf{X}}_{\bullet\bullet}) - f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu})] \\
= \; & \sqrt{n}(\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1})f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu}) + \sqrt{n}\boldsymbol{\Sigma}^{-1}[\hat{f}_y(\bar{\mathbf{X}}_{y\bullet} - \bar{\mathbf{X}}_{\bullet\bullet}) - f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu})] \\
& + O_p(n^{-\frac{1}{2}}). \qquad (10.6)
\end{aligned}$$

Define h random variables J_y such that J_y equals 1 if the point in the y-th slice and 0 otherwise, $y = 1, 2, \ldots, h$. Then, $\mathrm{E}[J_y] = f_y$. Let J_{yj} denote the value of J_y for the jth observation, $j = 1, 2, \ldots, n$. By Lemma 1 in Section 3.3.1, we have

$$\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1} = -n^{-1}\boldsymbol{\Sigma}^{-\frac{1}{2}}\sum_{j=1}^{n}(\mathbf{Z}_j\mathbf{Z}_j^T - \mathbf{I})\boldsymbol{\Sigma}^{-\frac{1}{2}} + O_p(n^{-1}).$$

Therefore, the first term in (10.6) can be simplified as

$$\begin{aligned}
& \sqrt{n}(\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}) f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu}) \\
= & -n^{-\frac{1}{2}} \boldsymbol{\Sigma}^{-\frac{1}{2}} \sum_{j=1}^{n} (\mathbf{Z}_j \mathbf{Z}_j^T - \mathbf{I}) \boldsymbol{\Sigma}^{-\frac{1}{2}} f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu}) + O_p(n^{-\frac{1}{2}}) \\
= & -n^{-\frac{1}{2}} \boldsymbol{\Sigma}^{-\frac{1}{2}} \sum_{j=1}^{n} (\mathbf{Z}_j \mathbf{Z}_j^T - \mathbf{I}) \mathrm{E}[\mathbf{Z} J_y] + O_p(n^{-\frac{1}{2}}). \qquad (10.7)
\end{aligned}$$

Meanwhile, by definition of J_{yj}, we have

$$\begin{aligned}
\hat{f}_y(\bar{\mathbf{X}}_{y\bullet} - \bar{\mathbf{X}}_{\bullet\bullet}) &= \frac{1}{n} \sum_{j=1}^{n} [(\mathbf{X}_j - \bar{\mathbf{X}}_{\bullet\bullet}) J_{yj}] \\
&= \frac{1}{n} \sum_{j=1}^{n} [(\mathbf{X}_j - \bar{\mathbf{X}}_{\bullet\bullet})(J_{yj} - \mathrm{E}[J_y])] \\
&= \frac{1}{n} \sum_{j=1}^{n} [(\mathbf{X}_j - \boldsymbol{\mu})(J_{yj} - \mathrm{E}[J_y])] - \frac{1}{n} \sum_{j=1}^{n} [(\bar{\mathbf{X}}_{\bullet\bullet} - \boldsymbol{\mu})(J_{yj} - E[J_y])] \\
&= \frac{1}{n} \sum_{j=1}^{n} [(\mathbf{X}_j - \boldsymbol{\mu})(J_{yj} - \mathrm{E}[J_y])] - \frac{1}{n} (\bar{\mathbf{X}}_{\bullet\bullet} - \boldsymbol{\mu}) \sum_{j=1}^{n} (J_{yj} - \mathrm{E}[J_y]) \\
&= \frac{1}{n} \sum_{j=1}^{n} [(\mathbf{X}_j - \boldsymbol{\mu})(J_{yj} - \mathrm{E}[J_y])] + O_p(n^{-1}).
\end{aligned}$$

Therefore, the second term in (10.6) can be simplified as

$$\begin{aligned}
& \sqrt{n} \boldsymbol{\Sigma}^{-1} [\hat{f}_y(\bar{\mathbf{X}}_{y\bullet} - \bar{\mathbf{X}}_{\bullet\bullet}) - f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu})] \\
= & \; n^{-\frac{1}{2}} \boldsymbol{\Sigma}^{-\frac{1}{2}} \sum_{j=1}^{n} [\boldsymbol{\Sigma}^{-\frac{1}{2}} (\mathbf{X}_j - \boldsymbol{\mu})(J_{yj} - \mathrm{E}[J_y])] - \sqrt{n} \boldsymbol{\Sigma}^{-1} f_y(\boldsymbol{\mu}_y - \boldsymbol{\mu}) + O_p(n^{-\frac{1}{2}}) \\
= & \; n^{-\frac{1}{2}} \boldsymbol{\Sigma}^{-\frac{1}{2}} \sum_{j=1}^{n} [\mathbf{Z}_j (J_{yj} - \mathrm{E}[J_y])] - \sqrt{n} \boldsymbol{\Sigma}^{-\frac{1}{2}} \mathrm{E}[\mathbf{Z} J_y] + O_p(n^{-\frac{1}{2}}) \\
= & \; n^{-\frac{1}{2}} \boldsymbol{\Sigma}^{-\frac{1}{2}} \sum_{j=1}^{n} [\mathbf{Z}_j (J_{yj} - \mathrm{E}[J_y]) - \mathrm{E}[\mathbf{Z} J_y]] + O_p(n^{-\frac{1}{2}}). \qquad (10.8)
\end{aligned}$$

We plug (10.7) and (10.8) into (10.6) and obtain

$$\begin{aligned}
& \sqrt{n}(\hat{\boldsymbol{\zeta}}_y - \boldsymbol{\zeta}_y) \\
= \; & n^{-\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}} \sum_{j=1}^{n} [\mathbf{Z}_j(J_{yj} - \mathrm{E}[J_y]) - \mathrm{E}[\mathbf{Z}J_y] - (\mathbf{Z}_j\mathbf{Z}_j^T - \mathbf{I})\mathrm{E}[\mathbf{Z}J_y]] + O_p(n^{-\frac{1}{2}}) \\
= \; & n^{-\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}} \sum_{j=1}^{n} [\mathbf{Z}_j(J_{yj} - \mathrm{E}[J_y] - \mathbf{Z}_j^T\mathrm{E}[\mathbf{Z}J_y])] + O_p(n^{-\frac{1}{2}}) \\
= \; & n^{-\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}} \sum_{j=1}^{n} [\mathbf{Z}_j\varepsilon_{yj}] + O_p(n^{-\frac{1}{2}}),
\end{aligned}$$

where $\varepsilon_{yj} = J_{yj} - \mathrm{E}[J_y] - \mathbf{Z}_j^T\mathrm{E}[\mathbf{Z}J_y]$ is the population ordinary least square residual. Denote $\boldsymbol{\epsilon}_j = [\varepsilon_{1j}, \cdots, \varepsilon_{hj}]^T$ as the jth value for the random vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_h)^T$. We have

$$\sqrt{n}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}) = n^{-\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}} \sum_{j=1}^{n} \mathbf{Z}_j\boldsymbol{\epsilon}_j^T + O_p(n^{-\frac{1}{2}})$$

and

$$\sqrt{n}(\,\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \,\mathrm{vec}(\boldsymbol{\zeta})) = n^{-\frac{1}{2}} \sum_{j=1}^{n} \mathrm{vec}(\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{Z}_j\boldsymbol{\epsilon}_j^T) + O_p(n^{-\frac{1}{2}}),$$

where $(\mathbf{Z}_j, \boldsymbol{\epsilon}_j)$ are i.i.d. random vectors.

Restoring w in subscripts we have

$$\sqrt{n_w}(\,\mathrm{vec}(\hat{\boldsymbol{\zeta}}_w) - \,\mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\nu}_w)) \to \mathrm{Normal}(0, \boldsymbol{\Omega}_w^*),$$

where

$$\boldsymbol{\Omega}_w^* \;=\; \mathrm{Cov}(\,\mathrm{vec}(\boldsymbol{\Sigma}_w^{-\frac{1}{2}}\mathbf{Z}_w\boldsymbol{\varepsilon}_w^T)). \tag{10.9}$$

By Slutsky's Theorem, we reach the conclusion that

$$\sqrt{n}(\,\mathrm{vec}(\hat{\boldsymbol{\zeta}}) - \,\mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\nu})) \to \mathrm{Normal}(0, \boldsymbol{\Gamma}).$$

where $\mathbf{\Gamma} = \text{diag}\{\frac{1}{p_w}\mathbf{\Omega}_w^*\}$.

It now follows from Proposition 1 that the asymptotic distribution of $n\hat{F}_d$ is the same as that of $\|\mathbf{Q}_{\mathbf{\Phi}}\mathbf{V}^{1/2}\mathbf{W}\|^2$ where $\mathbf{W}$ is normal with mean 0 and covariance matrix $\mathbf{\Gamma}$ and $\mathbf{\Phi} = \mathbf{V}^{1/2}\mathbf{\Delta}$ as defined for the statement of Theorem 4. Consequently, $n\hat{F}_d$ is asymptotically distributed as a linear combination of independent chi-squared random variables each with one degree of freedom. The coefficients of the chi-squared variables are the eigenvalues of

$$\mathbf{Q}_{\mathbf{\Phi}}\mathbf{V}^{1/2}\mathbf{\Gamma}\mathbf{V}^{1/2}\mathbf{Q}_{\mathbf{\Phi}} = \mathbf{Q}_{\mathbf{\Phi}}\mathbf{\Omega}\mathbf{Q}_{\mathbf{\Phi}}$$

where $\mathbf{\Omega} = \mathbf{V}^{1/2}\mathbf{\Gamma}\mathbf{V}^{1/2}$ is as defined for the statement of the Theorem. Finally, consistency follows from conclusion number 3 of Proposition 1 in combination with Lemma 2.

10.2 Computation of GP.SIR

We next summarize the computations necessary to implement the tests available as a result of Theorem 4.

To use Theorem 4 in practice, we need to replace $\mathbf{Q}_{\mathbf{\Phi}}\mathbf{\Omega}\mathbf{Q}_{\mathbf{\Phi}}$ with a consistent estimate under the null hypothesis. Under the hypothesis $d = m$, the $ph \times (p+h)m$ Jacobian matrix $\mathbf{\Delta}$ can be estimated consistently by substituting the corresponding estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$: $\hat{\mathbf{\Delta}} = (\hat{\boldsymbol{\nu}}^T \otimes \mathbf{I}_p, \mathbf{I}_h \otimes \hat{\boldsymbol{\beta}})$. The remaining unknowns are moments that do not depend on the hypothesis and can be estimated consistently by substituting the usual sample versions. To

estimate $\mathbf{V}$ we use $\hat{\mathbf{V}} = \text{diag}\{\hat{p}_w \mathbf{D}_{\hat{\mathbf{f}}_w}^{-1} \otimes \hat{\mathbf{\Sigma}}_w\}$. Because $\boldsymbol{\varepsilon}_w$ contains residuals from the population ordinary least square fit of J_{wy} on $\mathbf{Z}_w$ within subpopulation w, it is uncorrelated with $\mathbf{Z}_w$. Consequently,

$$\mathbf{\Omega}_w = (\mathbf{D}_{\mathbf{f}_w}^{-1/2} \otimes \mathbf{I}_p)\mathrm{E}(\boldsymbol{\varepsilon}_w \boldsymbol{\varepsilon}_w^T \otimes \mathbf{Z}_w \mathbf{Z}_w^T)(\mathbf{D}_{\mathbf{f}_w}^{-1/2} \otimes \mathbf{I}_p)$$

which suggests the estimate

$$\hat{\mathbf{\Omega}}_w = (\mathbf{D}_{\hat{\mathbf{f}}_w}^{-1/2} \otimes \mathbf{I}_p) \left(\frac{1}{n_w} \sum_{j=1}^{n_w} (\hat{\boldsymbol{\varepsilon}}_{wj} \hat{\boldsymbol{\varepsilon}}_{wj}^T \otimes \hat{\mathbf{Z}}_{wj} \hat{\mathbf{Z}}_{wj}^T) \right) (\mathbf{D}_{\hat{\mathbf{f}}_w}^{-1/2} \otimes \mathbf{I}_p)$$

where $\hat{\mathbf{Z}}_w$ is the sample version of $\mathbf{Z}_w$ and $\hat{\boldsymbol{\varepsilon}}_w$ contains residuals from the sample ordinary least square fit of J_{wy} on $\hat{\mathbf{Z}}_w$. These estimates are then substituted to yield an estimate of $\mathbf{Q}_{\mathbf{\Phi}} \mathbf{\Omega} \mathbf{Q}_{\mathbf{\Phi}}$ from which sample eigenvalues $\hat{\lambda}_j$ are obtained. The statistic $n\hat{F}_m$ is then compared to the percentage points of the distribution of

$$\sum_{i=1}^{ph} \hat{\lambda}_i \chi_i^2(1)$$

to obtain a p-value.

10.3 GP.SIR with Known Population Covariances

In this section we present two corollaries to Theorem 4 that describe the limiting distribution of the test statistic $n\hat{F}_d$ under various additional conditions. The first corollary deals with regressions in which the subpopulation covariance matrices $\mathbf{\Sigma}_w$ are known and used in place of the corresponding

estimates in the discrepancy function F_d defined at (9.5). The second corollary takes a step further. With additional marginal covariance condition, the test statistic has an asymptotic chi-squared distribution.

Corollary 6. *Assume that the subpopulation covariance* $\mathbf{\Sigma}_w$ *are known and used in the discrepancy function to compute* $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \arg_{\mathbf{B},\mathbf{C}} \min F_d(\mathbf{B}, \mathbf{C})$. *Then, as* $n \to \infty$,

$$n\hat{F}_d \xrightarrow{\mathcal{D}} \sum_{i=1}^{ph} \lambda_i \chi_i^2(1).$$

Here $\{\lambda_1 \geq \ldots \geq \lambda_{ph}\}$ *are the eigenvalues of* $\mathbf{Q}_{\mathbf{\Phi}}\mathbf{\Omega}\mathbf{Q}_{\mathbf{\Phi}}$, *where* $\mathbf{Q}_{\mathbf{\Phi}}$ *is as defined in Theorem 4, but* $\mathbf{\Omega} = \text{diag}\{\mathbf{\Omega}_w\}$ *with*

$$\mathbf{\Omega}_w = (\mathbf{Q}_{\mathbf{g}_w} \otimes \mathbf{I}_p)\text{diag}\{\mathbf{\Sigma}_{\mathbf{Z}_w|y}\}(\mathbf{Q}_{\mathbf{g}_w} \otimes \mathbf{I}_p)$$

where the block diagonal matrix is over the values of y *in subpopulation* w, *and* $\mathbf{g}_w$ *is the* $h_w \times 1$ *vector with elements* $\sqrt{f_{wy}}$, $y = 1, \ldots, h_w$.

Corollary 6 shows that the asymptotic distribution of $n\hat{F}_d$ simplifies somewhat if the subpopulation covariance matrices are known, but it still involves a linear combination of chi-squared random variables. However, if we add the linearity, coverage and marginal covariance conditions then the asymptotic distribution simplifies to a chi-squared as described in Corollary 7.

Proof of Corollary 6

Define the $p \times h_w$ matrix

$$\bar{\mathcal{X}}_w \equiv [\bar{\mathbf{X}}_{w1\bullet}, \cdots, \bar{\mathbf{X}}_{wh_w\bullet}].$$

Then using the known subpopulation covariance matrices $\mathbf{\Sigma}_w$, $\hat{\boldsymbol{\zeta}}_w$ in (10.5) now can be written as

$$\hat{\boldsymbol{\zeta}}_w = \mathbf{\Sigma}_w^{-1} \bar{\mathcal{X}}_w \mathbf{D}_{\hat{\mathbf{g}}_w} \mathbf{Q}_{\hat{\mathbf{g}}_w} \mathbf{D}_{\hat{\mathbf{g}}_w}$$

where $\hat{\mathbf{g}}_w$ is the $h_w \times 1$ vector with elements $\sqrt{\hat{f}_{wy}}$, $y = 1, \ldots, h_w$. Estimation in this case is based on minimizing the discrepancy function (10.5) using the $\hat{\boldsymbol{\zeta}}_w$ with inner product matrix $\hat{\mathbf{V}} = \text{diag}\{\hat{p}_w \mathbf{D}_{\hat{\mathbf{f}}_w}^{-1} \otimes \mathbf{\Sigma}_w\}$. Define

$$G_d(\mathbf{B}, \mathbf{C}) = \sum_w (\text{ vec}(\tilde{\boldsymbol{\zeta}}_w) - \text{ vec}(\mathbf{BC}_w))^T \mathbf{V}_w (\text{ vec}(\tilde{\boldsymbol{\zeta}}_w) - \text{ vec}(\mathbf{BC}_w)),$$

where $\tilde{\boldsymbol{\zeta}}_w = \hat{\boldsymbol{\zeta}}_w \mathbf{D}_{\hat{\mathbf{f}}_w}^{-1} \mathbf{D}_{\mathbf{f}_w} = [f_{w1} \mathbf{\Sigma}_w^{-1} (\bar{\mathbf{X}}_{w1\bullet} - \bar{\mathbf{X}}_{w\bullet\bullet}), \ldots, f_{wh_w} \mathbf{\Sigma}_w^{-1} (\bar{\mathbf{X}}_{wh_w\bullet} - \bar{\mathbf{X}}_{w\bullet\bullet})]$, and $\mathbf{V}_w = p_w \mathbf{D}_{\mathbf{f}_w}^{-1} \otimes \mathbf{\Sigma}_w$. By Lemma 2 in Section 3.3.1, the asymptotic distribution of $n\hat{F}_d$ is the same as that of $n\hat{G}_d$. It is easy to see that as $n_w \to \infty$, $\sqrt{n_w}(\text{ vec}(\tilde{\boldsymbol{\zeta}}_w) - \text{ vec}(\boldsymbol{\beta}\boldsymbol{\nu}_w))$ converges to a normal random vector with mean zero and some covariance matrix $\mathbf{\Phi}_w$.

To find $\mathbf{\Phi}_w$, first note that $\text{E}(\tilde{\boldsymbol{\zeta}}_w | \hat{\mathbf{f}}_w)$ is constant and thus

$$\begin{aligned}
& \text{Cov}(\text{ vec}(\tilde{\boldsymbol{\zeta}}_w)) \\
= \; & \text{E}[\text{Cov}(\text{ vec}(\tilde{\boldsymbol{\zeta}}_w) | \hat{\mathbf{f}}_w)] + \text{Cov}(\text{E}[\text{ vec}(\tilde{\boldsymbol{\zeta}}_w) | \hat{\mathbf{f}}_w]) \\
= \; & \text{E}[(\mathbf{D}_{\mathbf{f}_w} \mathbf{D}_{\hat{\mathbf{g}}_w}^{-1} \mathbf{Q}_{\hat{\mathbf{g}}_w} \mathbf{D}_{\hat{\mathbf{g}}_w} \otimes \mathbf{\Sigma}_w^{-1}) \text{diag}\{\frac{1}{n_{wy}} \mathbf{\Sigma}_{\mathbf{X}_w|y}\} (\mathbf{D}_{\hat{\mathbf{g}}_w} \mathbf{Q}_{\hat{\mathbf{g}}_w} \mathbf{D}_{\hat{\mathbf{g}}_w}^{-1} \mathbf{D}_{\mathbf{f}_w} \otimes \mathbf{\Sigma}_w^{-1})] \\
= \; & \frac{1}{n_w} (\mathbf{D}_{\mathbf{g}_w} \mathbf{Q}_{\mathbf{g}_w} \otimes \mathbf{\Sigma}_w^{-1}) \text{diag}\{\mathbf{\Sigma}_{\mathbf{X}_w|y}\} (\mathbf{Q}_{\mathbf{g}_w} \mathbf{D}_{\mathbf{g}_w} \otimes \mathbf{\Sigma}_w^{-1}) + o(\frac{1}{n_w})
\end{aligned}$$

where $\mathbf{\Sigma}_{\mathbf{X}_w|y}$ is the conditional covariance of $\mathbf{X}_w$ given $Y_w = y$ in the wth population. Therefore, we have

$$\mathbf{\Phi}_w = (\mathbf{D}_{\mathbf{g}_w} \mathbf{Q}_{\mathbf{g}_w} \otimes \mathbf{\Sigma}_w^{-1}) \text{diag}\{\mathbf{\Sigma}_{\mathbf{X}_w|y}\} (\mathbf{Q}_{\mathbf{g}_w} \mathbf{D}_{\mathbf{g}_w} \otimes \mathbf{\Sigma}_w^{-1}).$$

Consequently, with $\tilde{\boldsymbol{\zeta}} = (\tilde{\boldsymbol{\zeta}}_1, \ldots, \tilde{\boldsymbol{\zeta}}_K)$,

$$\sqrt{n}(\operatorname{vec}(\tilde{\boldsymbol{\zeta}}) - \operatorname{vec}(\boldsymbol{\beta\nu})) \xrightarrow{\mathcal{D}} \text{Normal}(0, \boldsymbol{\Gamma})$$

where $\boldsymbol{\Gamma} = \operatorname{diag}\{\boldsymbol{\Gamma}_w\}$ and $\boldsymbol{\Gamma}_w = \boldsymbol{\Phi}_w / p_w$.

It now follows from Proposition 1 that $n\hat{F}_d$ is asymptotically distributed as a linear combination of independent chi-squared random variables each with one degree of freedom. The coefficients of the chi-squared variables are the eigenvalues of

$$\mathbf{Q}_{\boldsymbol{\Phi}} \mathbf{V}^{1/2} \boldsymbol{\Gamma} \mathbf{V}^{1/2} \mathbf{Q}_{\boldsymbol{\Phi}}.$$

Thus, $\boldsymbol{\Omega} = \operatorname{diag}\{\boldsymbol{\Omega}_w\}$ with

$$\begin{aligned}
\boldsymbol{\Omega}_w &= (\sqrt{p_w}\mathbf{D}_{\mathbf{g}_w}^{-1} \otimes \boldsymbol{\Sigma}_w^{1/2}) \boldsymbol{\Phi}_w / p_w (\sqrt{p_w}\mathbf{D}_{\mathbf{g}_w}^{-1} \otimes \boldsymbol{\Sigma}_w^{1/2}) \\
&= (\mathbf{Q}_{\mathbf{g}_w} \otimes \boldsymbol{\Sigma}_w^{-1/2}) \operatorname{diag}\{\boldsymbol{\Sigma}_{\mathbf{X}_w|y}\} (\mathbf{Q}_{\mathbf{g}_w} \otimes \boldsymbol{\Sigma}_w^{-1/2}) \\
&= (\mathbf{Q}_{\mathbf{g}_w} \otimes \mathbf{I}_p) \operatorname{diag}\{\boldsymbol{\Sigma}_{\mathbf{Z}_w|y}\} (\mathbf{Q}_{\mathbf{g}_w} \otimes \mathbf{I}_p),
\end{aligned}$$

which is the desired conclusion. □

Corollary 7. *Assume that*

1. *the linearity condition is satisfied within each subpopulation,*
2. *for each subpopulation,* $\operatorname{Cov}(\mathbf{Z}_w | \mathbf{P}_{\mathcal{S}_{Y_w|\mathbf{Z}_w}} \mathbf{Z}_w) = \mathbf{Q}_{\mathcal{S}_{Y_w|\mathbf{Z}_w}}$,
3. *the coverage condition holds,* $\dim(\mathcal{S}_{Y|\mathbf{X}}^{(W)}) = \dim(\mathcal{S}_{\boldsymbol{\xi}}) = d$,
4. *the subpopulation covariances* $\boldsymbol{\Sigma}_w$ *are known and used in the discrepancy function* F_d *to produce the estimate* $\operatorname{Span}(\hat{\boldsymbol{\beta}})$ *of* $\mathcal{S}_{Y|\mathbf{X}}^{(W)}$.

Then $n\hat{F}_d$ has an asymptotic chi-squared distribution with $(p-d)(h-d-K)$ degrees of freedom.

The premise leading to the asymptotic distribution of Corollary 7 is similar to that for the asymptotic distribution of the test statistic for partial SIR (CCL, prop. 4.2). Both use conditions 1-3 of Corollary 7. Partial SIR then adds the homogeneous covariance condition and estimates the common value by pooling. In contrast, condition 4 stipulates that the subpopulation covariances are different but known. Both ways lead to the same asymptotic chi-squared distribution. The asymptotic distribution for partial SIR can also be derived by using the minimum discrepancy approach as shown in Section 10.4.

Proof of Corollary 7

Our proof of Corollary 7 involves a fair amount of algebra. From the proof of Corollary 6 we have $\mathbf{\Gamma}$ and $\mathbf{V}$; $\mathbf{\Delta}$ is as given in (10.3). The rest of the proof involves using the various conditions of the corollary to verifying algebraically that $\mathbf{\Gamma U \Gamma U \Gamma} = \mathbf{\Gamma U \Gamma}$, where

$$\mathbf{U} = \mathbf{V} - \mathbf{V\Delta}(\mathbf{\Delta}^T\mathbf{V\Delta})^-\mathbf{\Delta}^T\mathbf{V} = \mathbf{V}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{\Phi}}\mathbf{V}^{\frac{1}{2}}.$$

Proposition 1 then implies that the limiting distribution is chi-squared.

Now,

$$\begin{aligned}\mathbf{\Gamma U \Gamma U \Gamma} - \mathbf{\Gamma U \Gamma} &= \mathbf{\Gamma V}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{\Phi}}\mathbf{V}^{\frac{1}{2}}\mathbf{\Gamma V}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{\Phi}}\mathbf{V}^{\frac{1}{2}}\mathbf{\Gamma} - \mathbf{\Gamma V}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{\Phi}}\mathbf{V}^{\frac{1}{2}}\mathbf{\Gamma} \\ &= \mathbf{\Gamma V}^{\frac{1}{2}}\mathbf{Q}_{\mathbf{\Phi}}(\mathbf{\Omega} - \mathbf{I}_{ph})\mathbf{Q}_{\mathbf{\Phi}}\mathbf{V}^{\frac{1}{2}}\mathbf{\Gamma},\end{aligned}$$

where $\mathbf{\Omega} = \mathbf{V}^{\frac{1}{2}}\mathbf{\Gamma}\mathbf{V}^{\frac{1}{2}}$ as indicated at the end of the proof of Theorem 4, and

$$\begin{aligned}\mathbf{\Phi} &= \mathbf{V}^{\frac{1}{2}}\mathbf{\Delta} = \mathrm{diag}\{\sqrt{p_w}\mathbf{D}_{\mathbf{g}_w}^{-1} \otimes \mathbf{\Sigma}_w^{\frac{1}{2}}\}[\boldsymbol{\nu}^T \otimes \mathbf{I}_p, \mathbf{I}_h \otimes \boldsymbol{\beta}] \\ &= \begin{bmatrix} \sqrt{p_1}\mathbf{D}_{\mathbf{g}_1}^{-1}\boldsymbol{\nu}_1^T \otimes \mathbf{\Sigma}_1^{\frac{1}{2}} & \sqrt{p_1}\mathbf{D}_{\mathbf{g}_1}^{-1} \otimes \mathbf{\Sigma}_1^{\frac{1}{2}}\boldsymbol{\beta} \\ \vdots & \ddots \\ \sqrt{p_K}\mathbf{D}_{\mathbf{g}_K}^{-1}\boldsymbol{\nu}_K^T \otimes \mathbf{\Sigma}_K^{\frac{1}{2}} & \sqrt{p_K}\mathbf{D}_{\mathbf{g}_K}^{-1} \otimes \mathbf{\Sigma}_K^{\frac{1}{2}}\boldsymbol{\beta} \end{bmatrix}.\end{aligned}$$

where $\boldsymbol{\nu}_w = \boldsymbol{\gamma}_w\mathbf{D}_{\mathbf{f}_w}$, $w = 1, \ldots, K$.

Considering a generic w-th subpopulation, we have from Corollary 6

$$\begin{aligned}\mathbf{\Omega}_w - \mathbf{I}_{ph_w} &= (\mathbf{Q}_{\mathbf{g}_w} \otimes \mathbf{I}_p)\mathrm{diag}\{\mathbf{\Sigma}_{\mathbf{Z}_w|y}\}(\mathbf{Q}_{\mathbf{g}_w} \otimes \mathbf{I}_p) - \mathbf{I}_{ph_w} \\ &= (\mathbf{Q}_{\mathbf{g}_w} \otimes \mathbf{I}_p)\mathrm{diag}\{\mathbf{\Sigma}_{\mathbf{Z}_w|y} - \mathbf{I}_p\}(\mathbf{Q}_{\mathbf{g}_w} \otimes \mathbf{I}_p) - (\mathbf{P}_{\mathbf{g}_w} \otimes \mathbf{I}_p) \quad (10.10)\end{aligned}$$

where the diagonal matrix is over the values of Y_w. The conditions of Corollary 2 allow us to use Lemma 4 in Section 5.3, such that

$$\mathrm{Span}(\mathbf{\Sigma}_{\mathbf{Z}_w|y} - \mathbf{I}_p) \subseteq \mathrm{Span}(\mathbf{\Sigma}_w^{\frac{1}{2}}\boldsymbol{\beta}) = \mathcal{S}_{Y_w|\mathbf{Z}_w}.$$

Consequently, we have

$$\mathrm{Span}(\mathrm{diag}\{\mathbf{\Sigma}_{\mathbf{Z}_w|y} - \mathbf{I}_p\}) \subseteq \mathrm{Span}(\mathbf{I}_{h_w} \otimes \mathbf{\Sigma}_w^{\frac{1}{2}}\boldsymbol{\beta})$$

and

$$\mathrm{Span}((\mathbf{Q}_{\mathbf{g}_w} \otimes \mathbf{I}_p)\mathrm{diag}\{\mathbf{\Sigma}_{\mathbf{Z}_w|y} - \mathbf{I}_p\}) \subseteq \mathrm{Span}(\mathbf{Q}_{\mathbf{g}_w} \otimes \mathbf{\Sigma}_w^{\frac{1}{2}}\boldsymbol{\beta}) \subseteq \mathrm{Span}(\mathbf{D}_{\mathbf{g}_w}^{-1} \otimes \mathbf{\Sigma}_w^{\frac{1}{2}}\boldsymbol{\beta}).$$

It follows from (10.10) that the first term of $\mathbf{\Omega} - \mathbf{I}_{ph} = \mathrm{diag}\{\mathbf{\Omega}_w - \mathbf{I}_{ph_w}\}$ is in $\mathrm{Span}(\mathbf{\Phi})$ and therefore

$$\mathbf{Q}_{\mathbf{\Phi}}(\mathbf{\Omega} - \mathbf{I}_{ph}) = -\mathbf{Q}_{\mathbf{\Phi}}\mathrm{diag}\{\mathbf{P}_{\mathbf{g}_w} \otimes \mathbf{I}_p\}$$

where $\text{diag}\{\mathbf{P}_{\mathbf{g}_w} \otimes \mathbf{I}_p\}$ is over w. For the projection operator $\mathbf{Q}_{\mathbf{\Phi}}$, we can consider the complement orthogonal projection of any matrix which spans the same space as $\mathbf{\Phi}$ does. We notice that

$$\mathbf{Q}_{\mathbf{\Phi}} = \mathbf{Q}_{\mathbf{\Phi}_1} = \mathbf{Q}_{\mathbf{\Phi}_{11}}\mathbf{Q}_{\mathbf{\Phi}_{12}} = \mathbf{Q}_{\mathbf{\Phi}_{12}}\mathbf{Q}_{\mathbf{\Phi}_{11}},$$

where

$$\mathbf{\Phi}_1 = [\mathbf{\Phi}_{11}, \mathbf{\Phi}_{12}] = \begin{bmatrix} \sqrt{p_1}\mathbf{D}_{\mathbf{g}_1}^{-1}\boldsymbol{\nu}_1^T \otimes \mathbf{Q}_{\mathbf{\Sigma}_1^{\frac{1}{2}}\boldsymbol{\beta}}\mathbf{\Sigma}_1^{\frac{1}{2}} & \mathbf{I}_{h_1} \otimes \mathbf{\Sigma}_1^{\frac{1}{2}}\boldsymbol{\beta} \\ \vdots & \ddots \\ \sqrt{p_K}\mathbf{D}_{\mathbf{g}_K}^{-1}\boldsymbol{\nu}_K^T \otimes \mathbf{Q}_{\mathbf{\Sigma}_K^{\frac{1}{2}}\boldsymbol{\beta}}\mathbf{\Sigma}_K^{\frac{1}{2}} & \mathbf{I}_{h_K} \otimes \mathbf{\Sigma}_K^{\frac{1}{2}}\boldsymbol{\beta} \end{bmatrix}.$$

Here $\mathbf{\Phi}_{11}$ and $\mathbf{\Phi}_{12}$ are implicitly defined. We also have

$$\begin{aligned} (\sqrt{p_w}\mathbf{D}_{\mathbf{g}_w}^{-1}\boldsymbol{\nu}_w^T \otimes \mathbf{Q}_{\mathbf{\Sigma}_w^{\frac{1}{2}}\boldsymbol{\beta}}\mathbf{\Sigma}_w^{\frac{1}{2}})^T(\mathbf{P}_{\mathbf{g}_w} \otimes \mathbf{I}_p) &= \sqrt{p_w}\boldsymbol{\nu}_w\mathbf{D}_{\mathbf{g}_w}^{-1}\mathbf{P}_{\mathbf{g}_w} \otimes \mathbf{\Sigma}_w^{\frac{1}{2}}\mathbf{Q}_{\mathbf{\Sigma}_w^{\frac{1}{2}}\boldsymbol{\beta}} \\ &= \sqrt{p_w}\boldsymbol{\nu}_w\mathbf{1}\mathbf{g}_w^T \otimes \mathbf{\Sigma}_w^{\frac{1}{2}}\mathbf{Q}_{\mathbf{\Sigma}_w^{\frac{1}{2}}\boldsymbol{\beta}} \\ &= 0 \end{aligned}$$

because $\boldsymbol{\nu}_w\mathbf{1} = \boldsymbol{\gamma}_w\mathbf{f}_w = 0$. Thus, $\mathbf{\Phi}_{11}^T\text{diag}\{\mathbf{P}_{\mathbf{g}_w} \otimes \mathbf{I}_p\} = 0$ and

$$\begin{aligned} \mathbf{Q}_{\mathbf{\Phi}}(\mathbf{\Omega} - \mathbf{I}_{ph}) &= -\mathbf{Q}_{\mathbf{\Phi}}\text{diag}\{\mathbf{P}_{\mathbf{g}_w} \otimes \mathbf{I}_p\} = -\mathbf{Q}_{\mathbf{\Phi}_{12}}\mathbf{Q}_{\mathbf{\Phi}_{11}}\text{diag}\{\mathbf{P}_{\mathbf{g}_w} \otimes \mathbf{I}_p\} \\ &= -\mathbf{Q}_{\mathbf{\Phi}_{12}}\text{diag}\{\mathbf{P}_{\mathbf{g}_w} \otimes \mathbf{I}_p\} = -\text{diag}\{\mathbf{P}_{\mathbf{g}_w} \otimes \mathbf{Q}_{\mathbf{\Sigma}_w^{\frac{1}{2}}\boldsymbol{\beta}}\}. \end{aligned}$$

Therefore,

$$
\begin{aligned}
& \boldsymbol{\Gamma}\mathbf{V}^{\frac{1}{2}}\mathbf{Q}_{\boldsymbol{\Phi}}(\boldsymbol{\Omega}-\mathbf{I}_{ph}) \\
= \; & -\boldsymbol{\Gamma}\mathbf{V}^{\frac{1}{2}}\mathrm{diag}\{\mathbf{P}_{\mathbf{g}_w}\otimes\mathbf{Q}_{\boldsymbol{\Sigma}_w^{\frac{1}{2}}\boldsymbol{\beta}}\} \\
= \; & -\mathrm{diag}\{\frac{1}{p_w}(\mathbf{D}_{\mathbf{g}_w}\mathbf{Q}_{\mathbf{g}_w}\otimes\boldsymbol{\Sigma}_w^{-1})\mathrm{diag}\{\boldsymbol{\Sigma}_{\mathbf{X}_w|Y}\}(\mathbf{Q}_{\mathbf{g}_w}\mathbf{D}_{\mathbf{g}_w}\otimes\boldsymbol{\Sigma}_w^{-1})\} \\
& \mathrm{diag}\{\sqrt{p_w}\mathbf{D}_{\mathbf{g}_w}^{-1}\otimes\boldsymbol{\Sigma}_w^{\frac{1}{2}}\}\mathrm{diag}\{\mathbf{P}_{\mathbf{g}_w}\otimes\mathbf{Q}_{\boldsymbol{\Sigma}_w^{\frac{1}{2}}\boldsymbol{\beta}}\} \\
= \; & -\mathrm{diag}\{\frac{1}{\sqrt{p_w}}(\mathbf{D}_{\mathbf{g}_w}\mathbf{Q}_{\mathbf{g}_w}\otimes\boldsymbol{\Sigma}_w^{-1})\mathrm{diag}\{\boldsymbol{\Sigma}_{\mathbf{X}_w|Y}\}(\mathbf{Q}_{\mathbf{g}_w}\mathbf{P}_{\mathbf{g}_w}\otimes\boldsymbol{\Sigma}_w^{-\frac{1}{2}}\mathbf{Q}_{\boldsymbol{\Sigma}_w^{\frac{1}{2}}\boldsymbol{\beta}})\} \\
= \; & 0.
\end{aligned}
$$

Therefore, $\boldsymbol{\Gamma}\mathbf{U}\boldsymbol{\Gamma}\mathbf{U}\boldsymbol{\Gamma}=\boldsymbol{\Gamma}\mathbf{U}\boldsymbol{\Gamma}$. The degrees of freedom are

$$
\begin{aligned}
\mathrm{trace}(\mathbf{U}\boldsymbol{\Gamma}) &= \mathrm{trace}(\mathbf{V}^{\frac{1}{2}}\mathbf{Q}_{\boldsymbol{\Phi}}\mathbf{V}^{\frac{1}{2}}\boldsymbol{\Gamma}) \\
&= \mathrm{trace}(\mathbf{Q}_{\boldsymbol{\Phi}}\boldsymbol{\Omega}) \\
&= \mathrm{trace}(\mathbf{Q}_{\boldsymbol{\Phi}}) - \mathrm{trace}(\mathrm{diag}\{\mathbf{P}_{\mathbf{g}_w}\otimes\mathbf{Q}_{\boldsymbol{\Sigma}_w^{\frac{1}{2}}\boldsymbol{\beta}}\}) \\
&= ph - \mathrm{rank}(\boldsymbol{\Phi}) - \mathrm{rank}(\mathrm{diag}\{\mathbf{P}_{\mathbf{g}_w}\otimes\mathbf{Q}_{\boldsymbol{\Sigma}_w^{\frac{1}{2}}\boldsymbol{\beta}}\}) \\
&= ph - \mathrm{rank}(\boldsymbol{\Delta}) - K(p-d),
\end{aligned}
$$

where

$$
\begin{aligned}
\mathrm{rank}(\boldsymbol{\Delta}) &= \mathrm{rank}([\boldsymbol{\nu}^T\otimes\mathbf{I}_p, \mathbf{I}_h\otimes\boldsymbol{\beta}]) \\
&= \mathrm{rank}([\boldsymbol{\nu}^T\otimes\mathbf{Q}_{\boldsymbol{\beta}}, \mathbf{I}_h\otimes\boldsymbol{\beta}]) \\
&= \mathrm{rank}(\boldsymbol{\nu}^T\otimes\mathbf{Q}_{\boldsymbol{\beta}}) + \mathrm{rank}(\mathbf{I}_h\otimes\boldsymbol{\beta}) \\
&= d(p-d) + hd \\
&= d(h+p-d).
\end{aligned}
$$

Therefore, the degrees of freedom are $(p-d)(h-d-K)$.

10.4 Partial SIR Revisited

We have seen that partial SIR is a special case of GP.SIR when all population share the common covariance matrix. CCL proved that replacing condition number 4 in Corollary 7 with homogeneity of subpopulation covariances, the test statistic has the same asymptotic chi-squared distribution. Up to this moment, we find this a very natural by-product. We restate their result and give the justification in the minimum discrepancy approach.

Corollary 8. *(CCL 2002 Proposition 4.2) Assume that*

1. *the linearity condition is satisfied within each subpopulation,*
2. *for each subpopulation,* $\mathrm{Cov}(\mathbf{Z}_w|\mathbf{P}_{\mathcal{S}_{Y_w|\mathbf{Z}_w}}\mathbf{Z}_w) = \mathbf{Q}_{\mathcal{S}_{Y_w|\mathbf{Z}_w}}$,
3. *the coverage condition holds,* $\dim(\mathcal{S}^{(W)}_{Y|\mathbf{X}}) = \dim(\mathcal{S}_{\boldsymbol{\xi}}) = d$,
4. *the subpopulation covariances* $\boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}_{pool}$. *The estimate of the common covariance*

$$\hat{\boldsymbol{\Sigma}}_{pool} = \sum_{w=1}^{K} \hat{p}_w \hat{\boldsymbol{\Sigma}}_w$$

is used in the discrepancy function F_d *to produce the estimate* $\mathrm{Span}(\hat{\boldsymbol{\beta}})$ *of* $\mathcal{S}^{(W)}_{Y|\mathbf{X}}$.

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

Let us re-examine the discrepancy function of partial SIR

$$
\begin{aligned}
F_d(\mathbf{B},\mathbf{C}) &= \sum_{w=1}^{K}\hat{p}_w\sum_{y=1}^{h_w}\hat{f}_{wy}(\hat{\boldsymbol{\xi}}_{wy}-\mathbf{B}\mathbf{C}_{wy})^T\hat{\boldsymbol{\Sigma}}_{pool}(\hat{\boldsymbol{\xi}}_{wy}-\mathbf{B}\mathbf{C}_{wy}) \qquad (10.11)\\
&= \sum_{w=1}^{K}\hat{p}_w\sum_{y=1}^{h_w}\hat{f}_{wy}(\tilde{\boldsymbol{\xi}}_{wy}-\boldsymbol{\Sigma}_{pool}^{-1}\hat{\boldsymbol{\Sigma}}_{pool}\mathbf{B}\mathbf{C}_{wy})^T\boldsymbol{\Sigma}_{pool}\hat{\boldsymbol{\Sigma}}_{pool}^{-1}\boldsymbol{\Sigma}_{pool}\\
&\quad (\tilde{\boldsymbol{\xi}}_{wy}-\boldsymbol{\Sigma}_{pool}^{-1}\hat{\boldsymbol{\Sigma}}_{pool}\mathbf{B}\mathbf{C}_{wy}),
\end{aligned}
$$

where $\tilde{\boldsymbol{\xi}}_{wy} = \boldsymbol{\Sigma}_{pool}^{-1}\hat{\boldsymbol{\Sigma}}_{pool}\boldsymbol{\xi}_{wy}$. Therefore, by Lemma 2 in Section 3.3.1, the asymptotic distribution of the minimum value of (10.11) is the same as that of

$$
\sum_{w=1}^{K}\hat{p}_w\sum_{y=1}^{h_w}\hat{f}_{wy}(\tilde{\boldsymbol{\xi}}_{wy}-\mathbf{B}\mathbf{C}_{wy})^T\boldsymbol{\Sigma}_{pool}(\tilde{\boldsymbol{\xi}}_{wy}-\mathbf{B}\mathbf{C}_{wy})
$$

that is the objective function when all population covariance $\boldsymbol{\Sigma}_w$ are known. Then the result follows from Corollary 7.

Chapter 11

Comparison of Partial SIR and General Partial SIR

In this chapter, we consider three versions of the test statistic $n\hat{F}_d$. The three versions used were the tests for partial SIR, GP.SIR with known subpopulation covariances $\mathbf{\Sigma}_w$ as in Corollary 6, and GP.SIR with estimated subpopulation covariances as in Theorem 4. Table 11.1 summarizes their main features.

11.1 Simulation Results

In this section, we present results from a simulation study to investigate the actual level of nominal 1 and 5 percent tests. Here we present results based on two models. The predictor $\mathbf{X}$ comes from two populations. For each model, we consider two scenarios: normal populations with equal and unequal covariances. We also present simulation with non-normal cases. We

Table 11.1: Summary of partial SIR, GP.SIR with Known and Unknown $\boldsymbol{\Sigma}_w$.

Methods	$\hat{\boldsymbol{\xi}}_{wy}$	Matrix	Asymp. Dist.	Conditions
Partial SIR	$\hat{\boldsymbol{\Sigma}}_{\text{pool}}^{-1}(\bar{\mathbf{X}}_{wy\bullet} - \bar{\mathbf{X}}_{w\bullet\bullet})$	$\hat{p}_w \hat{f}_{wy} \hat{\boldsymbol{\Sigma}}_{\text{pool}}$	$\chi^2_{(p-d)(h-d-1)}$	(1)(2)(3)(4)*
GP.SIR, $\boldsymbol{\Sigma}_w$	$\boldsymbol{\Sigma}_w^{-1}(\bar{\mathbf{X}}_{wy\bullet} - \bar{\mathbf{X}}_{w\bullet\bullet})$	$\hat{p}_w \hat{f}_{wy} \boldsymbol{\Sigma}_w$	$\sum_{i=1}^{ph} \lambda_i \chi_i^2(1)$	(1)
GP.SIR, $\hat{\boldsymbol{\Sigma}}_w$	$\hat{\boldsymbol{\Sigma}}_w^{-1}(\bar{\mathbf{X}}_{wy\bullet} - \bar{\mathbf{X}}_{w\bullet\bullet})$	$\hat{p}_w \hat{f}_{wy} \hat{\boldsymbol{\Sigma}}_w$	$\sum_{i=1}^{ph} \lambda_i \chi_i^2(1)$	(1)

Note: (1) linearity condition; (2) coverage condition;

(3) marginal covariance condition;

(4) constant subpopulation covariance condition.

ran 1000 simulations for each sample size and ran tests with different slice numbers. Note that the nominal standard errors are 0.3 for 1 percent and 0.7 for 5 percent.

Simulation One

For the first model, the response Y is generated as

$$Y = \exp[-(X_1 + X_2 + 2X_3)] + 0.5\epsilon, \tag{11.1}$$

where $\mathbf{X} = (X_1, X_2, \ldots, X_5)^T$ is sampled from one of two normal populations indicated by W, $\mathbf{X}|W \sim \text{Normal}(0, \boldsymbol{\Sigma}_w)$. For each simulation with sample size n we generated half of the sample from each population, and set the slice number within each population to $\frac{h}{2}$.

The results in Table 11.2 are from simulations with $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}_5$. The

results confirm the asymptotic results and, as expected, confirm that larger sample sizes are needed for the actual level of the GP.SIR test to get usefully close to the nominal level. Large sample sizes tend to be needed to compensate for the additional variability caused by the estimation of the eigenvalue in its large sample distribution.

The results in Table 11.3 are from simulations with $\mathbf{\Sigma}_1 = \mathbf{I}_5$ and

$$\mathbf{\Sigma}_2 = \begin{pmatrix} 1.418 & 0.089 & -0.963 & 0.538 & 0.922 \\ 0.089 & 1.128 & -0.206 & 0.342 & 0.310 \\ -0.963 & -0.206 & 0.853 & -0.270 & -0.659 \\ 0.538 & 0.342 & -0.270 & 0.407 & 0.417 \\ 0.922 & 0.310 & -0.659 & 0.417 & 0.656 \end{pmatrix}.$$

Obviously, $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, the difference being easily detected by the test in Anderson (1984, Ch. 10) in samples of size 100. The very high estimated levels for partial SIR indicate far too many rejections. These results support our previous observation that partial SIR tends to confuse difference between the subpopulation covariances with the PCS. The results for GP.SIR are qualitatively similar to those in Table 11.2.

Table 11.2: Estimated level in percent of nominal 1 and 5 percent tests based on three versions of the statistic $n\hat{F}_d$ with $d = 1$ for model (11.1) with $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}$.

	Partial SIR		GP.SIR, $\mathbf{\Sigma}_w$		GP.SIR, $\hat{\mathbf{\Sigma}}_w$	
n	1	5	1	5	1	5
			$h = 6$			
100	0.8	5.3	0.5	4.2	4.1	12.4
200	0.9	5.5	1.0	6.1	1.5	9.9
400	1.7	5.7	1.5	5.4	1.7	7.5
800	1.5	5.1	1.5	5.6	1.4	6.0
			$h = 8$			
100	0.4	4.0	0.5	4.1	2.7	12.3
200	0.8	4.4	0.9	4.8	1.8	7.1
400	1.1	4.7	1.1	4.8	0.9	6.5
800	0.5	4.6	0.9	4.9	0.8	5.7
			$h = 10$			
100	0.9	4.7	0.7	4.5	3.7	13.9
200	0.8	4.0	0.9	3.9	1.2	7.1
400	0.8	5.3	0.7	5.0	0.9	5.9
800	0.7	4.3	1.0	4.5	1.1	4.6

Table 11.3: Estimated level in percent of nominal 1 and 5 percent tests based on three versions of the statistic $n\hat{F}_d$ with $d = 1$ for model (11.1) with $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$.

	Partial SIR		GP.SIR, $\boldsymbol{\Sigma}_w$		GP.SIR, $\hat{\boldsymbol{\Sigma}}_w$	
n	1	5	1	5	1	5
			$h = 6$			
100	54.4	80.5	0.5	4.7	4.7	12.2
200	98.4	99.5	1.0	4.9	2.3	8.9
400	100	100	0.5	3.8	1.0	4.8
800	100	100	1.0	5.7	1.6	5.3
			$h = 8$			
100	42.4	72.3	0.7	5.3	3.4	14.0
200	96.2	99.3	0.8	4.8	2.2	9.9
400	100	100	0.9	5.9	1.7	6.7
800	100	100	0.9	5.1	1.5	5.5
			$h = 10$			
100	33.7	62.4	0.5	4.2	2.7	13.9
200	93.2	98.3	1.4	5.0	2.2	8.6
400	100	100	0.6	4.6	1.4	6.1
800	100	100	0.8	4.5	1.1	5.4

Simulation Two

The second model we used has structural dimension 2,

$$Y = \frac{X_1}{0.5 + (X_2 + 1.5)^2} + 0.5\epsilon. \tag{11.2}$$

Here ϵ is a standard normal random variable independent of the predictors $\mathbf{X}$ and W. Table 11.4 shows the estimated levels with predictor distribution as in Table 11.2 that has $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. All three methods work well as expected. However, all three tests tend to be conservative for the smaller sample sizes, with the estimated levels for GP.SIR being perhaps slightly better than those for the other two tests. Table 11.5 shows simulations from (11.2) with predictor distribution the same as that used in Table 11.3, where $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$. Again the partial SIR test doesn't do well because of its propensity to overestimate the PCS. The two GP.SIR tests do reasonable well, and as in Table 11.4 the version with estimated covariances tends to do the best.

Non-normal Cases

For non-normal cases we ran tests only with slice number $h = 6$. Table 11.6 gives simulation results for GP.SIR from using non-normal predictors in the version of model (11.1) with $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$. For the first two columns headed $\chi^2(5)$ and $\chi^2(18)$, we used

$$\mathbf{X}_w \sim \frac{\text{Normal}(0, \mathbf{\Sigma}_w)}{\sqrt{\chi^2(D)/D}},$$

where the normal vector and chi-squared variable are independent and the degrees of freedom $D = 5$ in the first column and $D = 18$ in the second. The

Table 11.4: Estimated level in percent of nominal 1 and 5 percent tests based on three versions of the statistic $n\hat{F}_d$ with $d = 2$ for model (11.2) with $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}$.

	Partial SIR		GP.SIR, $\mathbf{\Sigma}_w$		GP.SIR, $\hat{\mathbf{\Sigma}}_w$	
n	1	5	1	5	1	5
			$h = 6$			
100	0.1	1.2	0.1	1.3	0.2	2.7
200	0.1	1.6	0.1	1.4	0.4	3.3
400	0.4	3.2	0.4	3.2	0.9	4.2
800	1.1	4.3	1.1	4.2	1.2	5.0
			$h = 8$			
100	0.0	1.5	0.1	1.2	0.5	3.1
200	0.0	2.5	0.0	1.8	0.7	4.0
400	0.2	3.3	0.3	4.0	0.7	4.4
800	0.3	4.4	0.3	3.9	0.3	4.1
			$h = 10$			
100	0.1	1.1	0.0	0.9	0.3	2.9
200	0.1	2.1	0.3	1.9	0.3	4.0
400	0.6	3.8	0.6	4.2	0.9	4.9
800	0.6	4.0	0.7	4.1	0.5	5.2

Table 11.5: Estimated level in percent of nominal 1 and 5 percent tests based on three versions of the statistic $n\hat{F}_d$ with $d = 2$ for model (11.2) with $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$.

	Partial SIR		GP.SIR, $\boldsymbol{\Sigma}_w$		GP.SIR, $\hat{\boldsymbol{\Sigma}}_w$	
n	1	5	1	5	1	5
			$h = 6$			
100	4.2	13.5	0.2	1.3	1.0	6.1
200	22.5	45.4	0.5	3.3	1.2	4.7
400	74.4	89.0	0.4	3.4	0.9	4.7
800	99.3	99.8	0.9	5.0	0.9	5.1
			$h = 8$			
100	3.7	14.2	0.1	1.3	0.7	5.0
200	30.9	56.9	0.6	3.4	0.8	5.4
400	88.2	96.4	0.8	4.3	0.8	5.7
800	100	100	0.6	5.2	1.2	5.6
			$h = 10$			
100	4.7	15.8	0.2	1.4	1.0	5.2
200	38.3	63.2	1.0	4.3	1.6	6.2
400	92.6	98.3	0.6	3.6	0.7	4.6
800	100	100	0.8	4.7	0.6	5.8

last column labeled as "Uniform" has

$$\mathbf{X}_w = U\frac{\boldsymbol{\Sigma}_w^{\frac{1}{2}}\mathbf{Z}_w}{\|\mathbf{Z}_w\|},$$

where $\mathbf{Z}_w$ are independent Normal$(0, \mathbf{I}_5)$ and U is an independent Uniform$(0, 1)$. The results in Table 11.6 are similar to the corresponding results for GP.SIR in Tables 11.3 and 11.5, but a somewhat larger sample size might be needed to obtain like agreement between the nominal and estimated levels.

Table 11.6: Estimated level in percent of nominal 1 and 5 percent test based on $n\hat{F}_d$ in GP.SIR with $\hat{\boldsymbol{\Sigma}}_w$ for model (11.1).

	$\chi^2(5)$		$\chi^2(18)$		Uniform	
n	1	5	1	5	1	5
100	4.9	13.8	3.6	12.9	5.1	13.5
200	3.3	11.1	1.9	7.5	2.2	7.9
400	1.4	8.8	0.8	6.9	1.8	6.0
800	1.4	6.8	1.2	7.2	1.4	5.9

Similarly, Table 11.7 gives simulation results for GP.SIR from using non-normal predictors in the version of model (11.2) with $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. It is interesting that GP.SIR has much lower actual levels than nominal ones in "Uniform" case, which is associated with the intrinsic operation characteristic of inverse regression estimation. It is usually difficult for SIR-type estimator to detect quadratic terms in symmetric case. However, it can do better when the distribution of the predictor has heavier tails as in $\chi^2(5)$ or $\chi^2(18)$.

Table 11.7: Estimated level in percent of nominal 1 and 5 percent test based on $n\hat{F}_d$ in GP.SIR with $\hat{\mathbf{\Sigma}}_w$ for model (11.2).

	$\chi^2(5)$		$\chi^2(18)$		Uniform	
n	1	5	1	5	1	5
100	2.1	8.3	1.2	6.2	0.7	2.6
200	1.6	6.6	1.4	6.5	0.4	1.6
400	1.1	5.9	0.7	4.3	0.1	1.3
800	0.9	5.8	0.3	5.2	0.4	2.3

From this simulation study, it is clear that for heterogenous subpopulations we definitely prefer general partial SIR over partial SIR. Since in homogeneous subpopulation, general partial SIR also does well, we recommend using GP.SIR unless there is strong evidence supporting constant subpopulation covariances.

11.2 Horse Mussels

In this section we consider a data set on New Zealand horse mussels. The response is the mussel's muscle mass, M. The $p = 4$ predictors in $\mathbf{X}$ are the shell length, shell height H and the logarithms of shell mass and shell width. The data, which were analyzed by Bura and Cook (2001b) in a different dimension reduction context, consist of observations on 172 mussels distributed across 5 collection sites represented by 58, 37, 37, 34, and 6

cases. We exclude the site with only 6 mussels, leaving $K = 4$ levels of the site indicator variable W. Table 11.8 shows the test statistics $n\hat{F}_m$, degrees of freedom or the trace of the sample version of $\mathbf{Q}_{\Phi}\mathbf{\Omega}\mathbf{Q}_{\Phi}$, p-values for partial SIR and GP.SIR when $h = 8$ so each slice has around 14 observations. The p-value for testing equality of 4 subpopulation covariance matrices is less than 10^{-10}. GP.SIR indicates only one direction $\hat{\boldsymbol{\beta}}_1$ for the PCS, while partial

Table 11.8: Mussel Data

	Partial SIR			GP.SIR		
NH: d = m	$n\hat{F}_m$	D.F.	p-value	$n\hat{F}_m$	Trace	p-value
0	109.13	16	0	84.38	16	0
1	16.65	9	0.05	4.66	4.01	0.31
2	2.59	4	0.63	0.93	1.58	0.66

SIR indicates two important directions $\hat{\boldsymbol{\rho}}_1$ and $\hat{\boldsymbol{\rho}}_2$. The correlation between $\hat{\boldsymbol{\beta}}_1^T\mathbf{X}$ and $\hat{\boldsymbol{\rho}}_1^T\mathbf{X}$ is about 0.999 so partial SIR and GP.SIR find essentially the same first direction. In view of the illustration in Section 9.5 and the simulation results in Tables 11.3 and 11.5, we expect that the second partial SIR direction is spurious, arising because of differences in the site covariance matrices. The summary plot of M versus $\hat{\boldsymbol{\beta}}_1^T\mathbf{X}$ is shown in Figure 11.1.

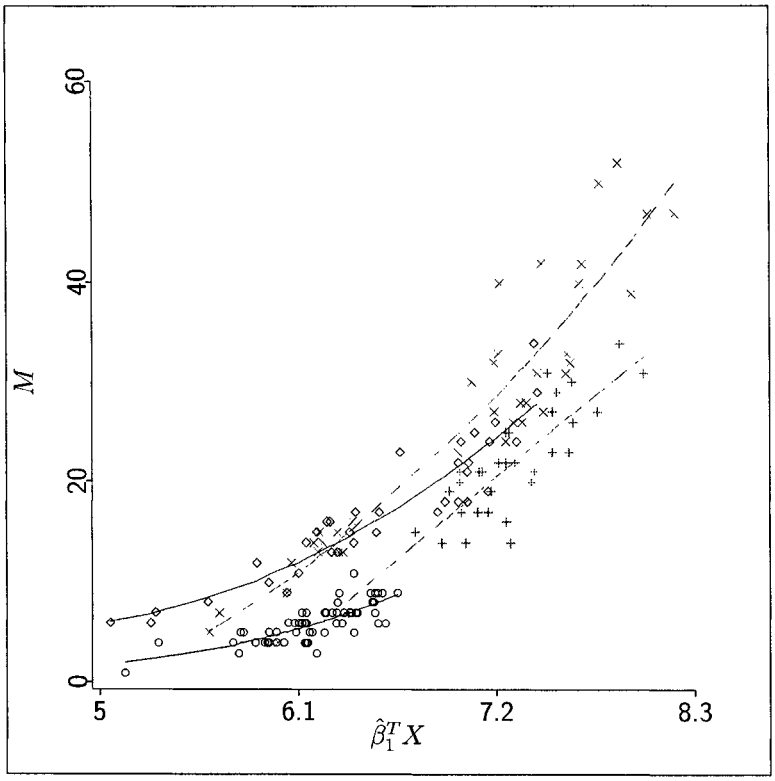


Figure 11.1: Summary plot for the mussel data based on general partial SIR. Locations are indicated by plotting symbol. The quadratic smooths are provided as visual enhancements.

Future Research

In this thesis, we consider sufficient dimension reduction via the central dimension reduction subspace for regression in a single population. Under the linearity condition, we have $\text{Span}\{\text{E}[\mathbf{Z}|Y]\} \subseteq \mathcal{S}_{Y|\mathbf{Z}}$ which connects dimension reduction with inverse regression. We developed an MDA family of dimension reduction methods by a minimum discrepancy approach, which uses quadratic inference functions constructed by linear combinations of sample versions of $\text{E}[\mathbf{Z}|Y]$. Within this MDA family, an efficient method—Optimal IRE—is proposed. Optimal IRE is optimal within the MDA family in two respects: the asymptotic efficiency for an estimated basis of the CS and the asymptotic chi-squared distribution for test statistic for dimension. Is Optimal IRE optimal globally among all estimating functions using sample versions of $\text{E}[\mathbf{Z}|Y]$? This needs further investigations.

The methods discussed in this thesis focus on extracting information about the CS from the vector $\text{E}[\mathbf{Z}|Y]$—the first moment of $\mathbf{Z}|Y$. Under assumptions of linearity and marginal covariance conditions we have $\text{Span}(\mathbf{I} - \mathbf{\Sigma}_{\mathbf{Z}|Y}) \subseteq \mathcal{S}_{Y|\mathbf{Z}}$. Based on this statement, Cook and Weisberg (1991) proposed another dimension reduction method—sliced average variance estima-

tion (SAVE)— which extracts information from the second moments of $\mathbf{Z}|Y$. Cook and Yin (2001) proposed a permutation method for estimating the dimension using SAVE. However, no asymptotic properties have been obtained for SAVE. The minimum discrepancy approach could be a key to a successful development of asymptotic theories about SAVE. This could set a stage for a unified approach to combine information from the first two moments.

For regression across multiple subpopulations, we developed a SIR-type method—general partial SIR, which extends partial SIR by removing the limiting condition of homogeneous subpopulation covariances. Therefore, GP.SIR can be used in far more circumstances than partial SIR. However, GP.SIR is not an optimal member of the MDA family as SIR is not in single population case. The idea of developing optimal methods for multiple subpopulations is straightforward based on the results in this thesis. More computational issues may arise along this line.

Application to classification problems is one of the recent developments in sufficient dimension reduction (Cook and Critchley 2000; Cook and Yin 2001). CCL illustrated the relation between the marginal central space (which is the CS without knowledge of subpopulations) and the partial central space, which opens a door for many applications. Dimension reduction methods like GP.SIR are very useful tools in this promising research field.

Appendix A

Notation

Definition 1. *For any matrix real* $\mathbf{A}$, $\mathbf{P_A}$ *is the orthogonal projection operator on the space spanned by the columns of* $\mathbf{A}$, $\mathrm{Span}\{\mathbf{A}\}$; $\mathbf{Q_A} = \mathbf{I} - \mathbf{P_A}$ *is the projection operator on the space orthogonal to* $Span\{\mathbf{A}\}$, *where* $\mathbf{I}$ *is the identity matrix.*

Definition 2. *The Kronecker product "*$\otimes$*" is defined as a matrix operator. Suppose* $\mathbf{A}$ *is an* $m \times n$ *matrix with* a_{ij} *being the* ij*-th element,* $\mathbf{B}$ *is a* $k \times l$ *matrix. Then,*

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{pmatrix},$$

an $mk \times nl$ *matrix.*

Properties of Projection and Kronecker Product

1. Let $[\mathbf{A}, \mathbf{B}]$ be the matrix that combines the columns of matrices $\mathbf{A}$ and $\mathbf{B}$.

   - For any matrices $\mathbf{A} \in \mathbb{R}^{l \times m}$, $\mathbf{B} \in \mathbb{R}^{l \times n}$, and $\mathbf{C} \in \mathbb{R}^{n \times m}$,

   $$\mathbf{Q}_{[\mathbf{A},\mathbf{B}]} = \mathbf{Q}_{[\mathbf{A}-\mathbf{BC},\mathbf{B}]}. \tag{A.1}$$

   - If $\text{Span}\{\mathbf{A}\}$ is orthogonal to $\text{Span}\{\mathbf{B}\}$, then

   $$\mathbf{Q}_{[\mathbf{A},\mathbf{B}]} = \mathbf{Q}_{\mathbf{A}}\mathbf{Q}_{\mathbf{B}} = \mathbf{Q}_{\mathbf{B}}\mathbf{Q}_{\mathbf{A}}. \tag{A.2}$$

2. $\text{rank}(\mathbf{A} \otimes \mathbf{B}) = \text{rank}(\mathbf{A}) \cdot \text{rank}(\mathbf{B})$.

   Proof: Suppose $\mathbf{A}$ and $\mathbf{B}$ have singular value decompositions as $\mathbf{A} = \mathbf{\Gamma}_{11}\mathbf{\Lambda}_1\mathbf{\Gamma}_{12}$ and $\mathbf{B} = \mathbf{\Gamma}_{21}\mathbf{\Lambda}_2\mathbf{\Gamma}_{22}$. Then, $\mathbf{A} \otimes \mathbf{B} = (\mathbf{\Gamma}_{11} \otimes \mathbf{\Gamma}_{21})(\mathbf{\Lambda}_1 \otimes \mathbf{\Lambda}_2)(\mathbf{\Gamma}_{12} \otimes \mathbf{\Gamma}_{22})$.

3. $\mathbf{P}_{\mathbf{A}\otimes\mathbf{B}} = \mathbf{P}_{\mathbf{A}} \otimes \mathbf{P}_{\mathbf{B}}$.

   Proof: $\mathbf{P}_{\mathbf{A}}\otimes\mathbf{P}_{\mathbf{B}} = (\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-}\mathbf{A}^T)\otimes(\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-}\mathbf{B}^T) = (\mathbf{A}\otimes\mathbf{B})((\mathbf{A}^T\mathbf{A})^{-}\mathbf{A}^T\otimes(\mathbf{B}^T\mathbf{B})^{-}\mathbf{B}^T)$. Thus $\text{Span}\{\mathbf{P}_{\mathbf{A}} \otimes \mathbf{P}_{\mathbf{B}}\} \subseteq \text{Span}\{\mathbf{A} \otimes \mathbf{B}\}$. On the other hand, $(\mathbf{P}_{\mathbf{A}} \otimes \mathbf{P}_{\mathbf{B}})(\mathbf{A} \otimes \mathbf{B})=(\mathbf{A} \otimes \mathbf{B})$, thus $\text{Span}\{\mathbf{A} \otimes \mathbf{B}\} \subseteq \text{Span}\{\mathbf{P}_{\mathbf{A}} \otimes \mathbf{P}_{\mathbf{B}}\}$.

4.

$$\mathbf{Q}_{\mathbf{A}\otimes\mathbf{B}} = \mathbf{P}_{\mathbf{A}} \otimes \mathbf{Q}_{\mathbf{B}} + \mathbf{Q}_{\mathbf{A}} \otimes \mathbf{I} = \mathbf{I} \otimes \mathbf{Q}_{\mathbf{B}} + \mathbf{Q}_{\mathbf{A}} \otimes \mathbf{P}_{\mathbf{B}} \tag{A.3}$$

Thus, if $\mathbf{A} \in \mathbb{R}^{m \times m}$ is full rank, $\mathbf{Q}_{\mathbf{A}\otimes\mathbf{B}} = \mathbf{I} \otimes \mathbf{Q}_{\mathbf{B}}$; if $\mathbf{B} \in \mathbb{R}^{n \times n}$ is full rank, $\mathbf{Q}_{\mathbf{A}\otimes\mathbf{B}} = \mathbf{Q}_{\mathbf{A}} \otimes \mathbf{I}$.

Proof:

$$\begin{aligned}\mathbf{Q}_{\mathbf{A}\otimes\mathbf{B}} &= \mathbf{I} - \mathbf{P}_{\mathbf{A}\otimes\mathbf{B}} \\ &= (\mathbf{P}_{\mathbf{A}} + \mathbf{Q}_{\mathbf{A}}) \otimes (\mathbf{P}_{\mathbf{B}} + \mathbf{Q}_{\mathbf{B}}) - \mathbf{P}_{\mathbf{A}\otimes\mathbf{B}} \\ &= \mathbf{P}_{\mathbf{A}} \otimes \mathbf{Q}_{\mathbf{B}} + \mathbf{Q}_{\mathbf{A}} \otimes \mathbf{P}_{\mathbf{B}} + \mathbf{Q}_{\mathbf{A}} \otimes \mathbf{Q}_{\mathbf{B}}\end{aligned}$$

5. For any matrices $\mathbf{A} \in \mathbb{R}^{l\times m}$, $\mathbf{B} \in \mathbb{R}^{n\times r}$, $\mathbf{C} \in \mathbb{R}^{l\times s}$, and $\mathbf{D} \in \mathbb{R}^{n\times t}$, if Span$\{\mathbf{A}\}$ and Span$\{\mathbf{C}\}$ are orthogonal or Span$\{\mathbf{B}\}$ and Span$\{\mathbf{D}\}$ are orthogonal, then

$$\mathbf{Q}_{\mathbf{A}\otimes\mathbf{B}}(\mathbf{C} \otimes \mathbf{D}) = \mathbf{C} \otimes \mathbf{D}.$$

Proof:

When Span$\{\mathbf{A}\}$ and Span$\{\mathbf{C}\}$ are orthogonal,

$$\begin{aligned}\mathbf{Q}_{\mathbf{A}\otimes\mathbf{B}}(\mathbf{C} \otimes \mathbf{D}) &= (\mathbf{I} \otimes \mathbf{Q}_{\mathbf{B}})(\mathbf{C} \otimes \mathbf{D}) + (\mathbf{Q}_{\mathbf{A}} \otimes \mathbf{P}_{\mathbf{B}})(\mathbf{C} \otimes \mathbf{D}) \\ &= \mathbf{C} \otimes \mathbf{Q}_{\mathbf{B}}\mathbf{D} + \mathbf{C} \otimes \mathbf{P}_{\mathbf{B}}\mathbf{D} \\ &= \mathbf{C} \otimes \mathbf{D}.\end{aligned}$$

Similarly, we can prove the statement when Span$\{\mathbf{B}\}$ and Span$\{\mathbf{D}\}$ are orthogonal.

Definition 3. *Let $\mathbf{a}_i$ be the i-th column of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$. A matrix operator vec$(\cdot)$ constructs a vector by stacking columns of a matrix such that*

$$\text{vec}(\mathbf{A}) = (\mathbf{a}_1^T, \mathbf{a}_2^T, \cdots, \mathbf{a}_n^T)^T.$$

One important property of vec$(\cdot)$ is that

$$\text{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})\,\text{vec}(\mathbf{B}). \tag{A.4}$$

Definition 4. *For any two $n \times n$ semi-positive definite matrices $\mathbf{A}$ and $\mathbf{B}$, if $\mathbf{A} - \mathbf{B}$ is semi-positive definite matrix, then we call $\mathbf{A} \geq \mathbf{B}$. The equality holds when $\mathbf{A} = \mathbf{B}$.*

Definition 5. *For any semi-positive matrix $\mathbf{A}$, $\mathbf{A}^{\frac{1}{2}}$ is a matrix such that $(\mathbf{A}^{\frac{1}{2}})^T = \mathbf{A}^{\frac{1}{2}}$ and $\mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}} = \mathbf{A}$.*

Definition 6. *Suppose $\mathbf{A}_i$ is an $n_i \times n_i$ dimensional matrix, $i = 1, 2, \ldots, k$. Let $n = \sum_{i=1}^{k} n_i$. Define*

$$\text{diag}\{\mathbf{A}_i\} = \begin{bmatrix} \mathbf{A}_1 & & & \\ & \mathbf{A}_2 & & \\ & & \ddots & \\ & & & \mathbf{A}_k \end{bmatrix},$$

an $n \times n$ dimensional block diagonal matrix.

Definition 7. *Suppose $\mathbf{A}_i$ is an $n_i \times m$ dimensional matrix, $i = 1, 2, \ldots, k$. Let $n = \sum_{i=1}^{k} n_i$. Define*

$$\text{stack}\{\mathbf{A}_i\} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_k \end{bmatrix},$$

an $n \times m$ dimensional matrix.

Appendix B

Lemmas for Optimization

Lemma 5. *Suppose* $\mathbf{A} = \sum_{i=1}^{n} \mathbf{a}_i \mathbf{a}_i^T$ *with a spectral decomposition* $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, *where* $\mathbf{a}_i \in \mathbb{R}^p$, $\mathbf{\Lambda} = \text{diag}\{\lambda_i\}$, *and* $\mathbf{U} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_p)$ *are its eigenvectors corresponding to eigenvalues* $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. *Let* $\kappa_m = \sum_{j=m+1}^{p} \lambda_j$, $m = 0, 1, \ldots, p-1$. *Then,*

$$\kappa_m = \min_{\boldsymbol{\beta} \in \mathbb{R}^{p \times m}, \boldsymbol{\gamma}_i \in \mathbb{R}^m} \sum_{i=1}^{n} (\mathbf{a}_i - \boldsymbol{\beta}\boldsymbol{\gamma}_i)^T (\mathbf{a}_i - \boldsymbol{\beta}\boldsymbol{\gamma}_i)$$

and $\text{Span}\{\hat{\boldsymbol{\beta}}\} = \text{Span}\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_m\}$. *Here* $\hat{\boldsymbol{\beta}}$ *is the value of* $\boldsymbol{\beta}$ *that minimizes* $\sum_{i=1}^{n} (\mathbf{a}_i - \boldsymbol{\beta}\boldsymbol{\gamma}_i)^T (\mathbf{a}_i - \boldsymbol{\beta}\boldsymbol{\gamma}_i)$.

Proof: Given $\boldsymbol{\beta}$, let $s(\boldsymbol{\beta}) = \min_{\boldsymbol{\gamma}_i \in \mathbb{R}^m} \sum_{i=1}^{n} (\mathbf{a}_i - \boldsymbol{\beta}\boldsymbol{\gamma}_i)^T (\mathbf{a}_i - \boldsymbol{\beta}\boldsymbol{\gamma}_i)$. It is easy to see that

$$\begin{aligned} s(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \mathbf{a}_i^T \mathbf{Q}_{\boldsymbol{\beta}} \mathbf{a}_i \\ &= \text{trace}(\mathbf{Q}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \mathbf{a}_i \mathbf{a}_i^T) \\ &= \text{trace}((\mathbf{I} - \mathbf{P}_{\boldsymbol{\beta}})\mathbf{A}) \\ &= \text{trace}(\mathbf{A}) - \text{trace}(\mathbf{P}_{\boldsymbol{\beta}} \mathbf{A}) \end{aligned}$$

Thus, minimizing $s(\boldsymbol{\beta})$ is equivalent to maximizing $\text{trace}(\mathbf{P}_{\boldsymbol{\beta}}\mathbf{A})$. Without loss of generality, assume $\boldsymbol{\beta}^T\boldsymbol{\beta} = \mathbf{I}_m$. Therefore,

$$\begin{aligned}\text{trace}(\mathbf{P}_{\boldsymbol{\beta}}\mathbf{A}) &= \text{trace}(\mathbf{P}_{\boldsymbol{\beta}}\mathbf{A}\mathbf{P}_{\boldsymbol{\beta}}) \\ &= \text{trace}(\boldsymbol{\beta}^T\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T\boldsymbol{\beta}) \\ &= \text{trace}(\boldsymbol{\Lambda}\mathbf{U}^T\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{U}).\end{aligned}$$

We have $\mathbf{U}^T\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{U} < \mathbf{U}^T\mathbf{U} = \mathbf{I}_p$ and $\text{trace}(\mathbf{U}^T\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{U}) = \text{trace}(\boldsymbol{\beta}^T\mathbf{U}\mathbf{U}^T\boldsymbol{\beta}) = \text{trace}(\mathbf{I}_m) = m$. Suppose $\{d_1, \ldots, d_p\}$ are the diagonal elements of $\mathbf{U}^T\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{U}$. Then, all $d_i \leq 1$, $i = 1, \ldots, p$, and $\sum_{i=1}^p d_i = m$. Therefore,

$$\text{trace}(\mathbf{P}_{\boldsymbol{\beta}}\mathbf{A}) = \sum_{i=1}^{p} d_i\lambda_i \leq \sum_{i=1}^{m} \lambda_i.$$

The equality holds only if $\mathbf{U}^T\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{U} = \text{diag}\{\mathbf{I}_m, 0\}$. Then,

$$\mathbf{P}_{\boldsymbol{\beta}} = \boldsymbol{\beta}\boldsymbol{\beta}^T = \mathbf{U}\text{diag}\{\mathbf{I}_m, 0\}\mathbf{U}^T = \sum_{i=1}^{m} \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T.$$

Lemma 6. *Consider the function*

$$f(\mathbf{b}) = \sum_{j=1}^{h}(\boldsymbol{\gamma}_j - c_j\mathbf{S}_j\mathbf{b})^T(\boldsymbol{\gamma}_j - c_j\mathbf{S}_j\mathbf{b})$$

where $c_j \in \mathbb{R}$, $\boldsymbol{\gamma}_j \in \mathbb{R}^p$, *and* $\mathbf{S}_j \in \mathbb{R}^{p\times p}$ *is a nonsingular matrix,* $j = 1, 2, \ldots, h$. *All* c_j, $\boldsymbol{\gamma}_j$, *and* $\mathbf{S}_j$ *are fixed. Here* $\mathbf{b} \in \mathbb{R}^p$ *and* $\|\mathbf{b}\| = 1$. *With the constraint that* $\mathbf{b}$ *is orthogonal to* $\text{Span}\{\mathbf{L}\}$, *where* $\mathbf{L} \in \mathbb{R}^{p\times m}$, *the argument that minimizes* $f(\mathbf{b})$ *is*

$$\hat{\mathbf{b}} = \mathbf{W}_2^{-1}[\mathbf{I} - \mathbf{L}(\mathbf{L}^T\mathbf{W}_2^{-1}\mathbf{L})^{-}\mathbf{L}^T\mathbf{W}_2^{-1}]\mathbf{W}_1.$$

where $\mathbf{W}_1 = \sum_{j=1}^{h} c_j\mathbf{S}_j^T\boldsymbol{\gamma}_j$ *and* $\mathbf{W}_2 = \sum_{j=1}^{h} c_j^2\mathbf{S}_j^T\mathbf{S}_j$.

Proof: The minimization with the constraint is equivalent to minimizing the function

$$k(\mathbf{g}) = \sum_{j=1}^{h} (\boldsymbol{\gamma}_j - c_j \mathbf{S}_j \mathbf{Q}_\mathbf{L} \mathbf{g})^T (\boldsymbol{\gamma}_j - c_j \mathbf{S}_j \mathbf{Q}_\mathbf{L} \mathbf{g})$$

where $\mathbf{g} \in \mathbb{R}^m$ and $\mathbf{Q}_\mathbf{L}$ is the projection on the orthogonal complement of Span$\{\mathbf{L}\}$. Immediately, we see that $k(\mathbf{g})$ is the sum of squared residuals from a multivariate ordinary least square fit,

$$\hat{\mathbf{g}} = \arg_\mathbf{g} \min k(\mathbf{g}) = (\mathbf{Q}_\mathbf{L} \mathbf{W}_2 \mathbf{Q}_\mathbf{L})^- \mathbf{Q}_\mathbf{L} \mathbf{W}_1$$

and

$$\begin{aligned}
\hat{\mathbf{b}} &= \mathbf{Q}_\mathbf{L} \hat{\mathbf{g}} \\
&= \mathbf{Q}_\mathbf{L} (\mathbf{Q}_\mathbf{L} \mathbf{W}_2 \mathbf{Q}_\mathbf{L})^- \mathbf{Q}_\mathbf{L} \mathbf{W}_1 \\
&= \mathbf{W}_2^{-\frac{1}{2}} \mathbf{W}_2^{\frac{1}{2}} \mathbf{Q}_\mathbf{L} (\mathbf{Q}_\mathbf{L} \mathbf{W}_2^{\frac{1}{2}} \mathbf{W}_2^{\frac{1}{2}} \mathbf{Q}_\mathbf{L})^- \mathbf{Q}_\mathbf{L} \mathbf{W}_2^{\frac{1}{2}} \mathbf{W}_2^{-\frac{1}{2}} \mathbf{W}_1 \\
&= \mathbf{W}_2^{-\frac{1}{2}} \mathbf{P}_{\mathbf{W}_2^{\frac{1}{2}} \mathbf{Q}_\mathbf{L}} \mathbf{W}_2^{-\frac{1}{2}} \mathbf{W}_1 \\
&= \mathbf{W}_2^{-\frac{1}{2}} \mathbf{Q}_{\mathbf{W}_2^{-\frac{1}{2}} \mathbf{L}} \mathbf{W}_2^{-\frac{1}{2}} \mathbf{W}_1 \qquad \text{(B.1)} \\
&= \mathbf{W}_2^{-\frac{1}{2}} [\mathbf{I} - \mathbf{W}_2^{-\frac{1}{2}} \mathbf{L} (\mathbf{L}^T \mathbf{W}_2^{-1} \mathbf{L})^- \mathbf{L}^T \mathbf{W}_2^{-\frac{1}{2}}] \mathbf{W}_2^{-\frac{1}{2}} \mathbf{W}_1 \\
&= \mathbf{W}_2^{-1} [\mathbf{I} - \mathbf{L} (\mathbf{L}^T \mathbf{W}_2^{-1} \mathbf{L})^- \mathbf{L}^T \mathbf{W}_2^{-1}] \mathbf{W}_1.
\end{aligned}$$

We know that $(\mathbf{W}_2^{\frac{1}{2}} \mathbf{Q}_\mathbf{L})^T \mathbf{W}_2^{-\frac{1}{2}} \mathbf{L} = 0$ and that $\text{rank}(\mathbf{W}_2^{\frac{1}{2}} \mathbf{Q}_\mathbf{L}) + \text{rank}(\mathbf{W}_2^{-\frac{1}{2}} \mathbf{L}) = p$. Therefore, $\mathbf{P}_{\mathbf{W}_2^{\frac{1}{2}} \mathbf{Q}_\mathbf{L}} + \mathbf{P}_{\mathbf{W}_2^{-\frac{1}{2}} \mathbf{L}} = \mathbf{I}_p$ and the equality in (B.1) holds.

Bibliography

Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis*, New York: Wiley.

Bura, E., and Cook, R. D. (2001a), "Estimating Structural Dimensions of Regressions via Parametric Inverse Regressions," *Journal of the Royal Statistical Society, Ser B*, 63, 393-410.

—- (2001b), "Extending Sliced Inverse Regression: the Weighted Chi-Squared Test," *Journal of the American Statistical Association*, 96, 996-1003.

Chiaromonte, F., Cook, R. D., and Li, B. (2002), "Sufficient Dimension Reduction in Regressions with Categorical Predictors," *The Annals of Statistics*, 30, 475-497.

Cook, R. D. (1994), "On the Interpretation of Regression Plots," *Journal of the American Statistical Association*, 89, 177-189.

—- (1996), "Graphics for Regressions with a Binary Response," *Journal of the American Statistical Association*, 91, 983-992.

—- (1998), *Regression Graphics: Ideas for Studying Regressions Through Graphics*, New York: Wiley.

Cook, R. D., and Critchley, F. (2000) "Identifying Regression Outliers and Mixtures Graphically," *Journal of the American Statistical Association*, 95:781-794.

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.

—- (1999), *Applied Regression including Computing and Graphics*, New York: Wiley.

Cook, R. D., and Yin, X. (2001) "Dimension Reduction and Visualization in Discriminant Analysis," *Australia and New Zealand Journal of Statistics*, 43(2),147-199.

Eaton, M. L., and Tyler, D. E. (1994), "The Asymptotic Distribution of Singular Values with Applications to Canonical Correlations and Correspondence Analysis," *Journal of Multivariate Analysis*, 34, 439-446.

Ferguson, T. (1958), "A Method of Generating Best Asymptotically Normal Estiamtes with Application to the Estimation of Bacterial Densities," *The Annals of Mathematical Statistics*, 29, 1046-162.

Field, C. (1993), "Tail Areas of Linear Combinations of Chi-Squares and Non-central Chi-squares," *Journal of Statistical Computation and Simulation*, 45, 243–248.

Friedman, J. and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association*, 76, 817-823.

Kiers, H. A. L. (2002), "Setting up Alternating Least Squares and Iterative Majorization Algorithms for Solving Various Matrix Optimization Problems," *Computational Statistics & Data Analysis*, 41, 157-170.

Li, K. C. (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86,316-342.

Rao, C. R. (1965), *Linear Statistical Inference and Its Applications* (1st ed.), New York: Wiley.

Ruhe, A., and Wedin, P. A. (1980), "Algorithms for Separable Nonlinear Least Squares Problems," *SIAM Review*, 22, 318-337.

Shapiro, A. (1986), "Asymptotic Theory of Overparameterized Structural Models," *Journal of the American Statistical Association*, 81,142-149.

Velilla, S. (1998), "Assessing the Number of Linear Components in a General Regression Problem," *Journal of the American Statistical Association*, 93,1088-1098.