Sufficient Dimension Reduction for Complex Data Structures

A THESIS

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Shanshan Ding

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

Doctor of Philosophy

R. Dennis Cook, Advisor

June, 2014

Acknowledgements

The years spent in graduate school at the University of Minnesota is one of the best time in my life. The journey of learning, studying, and doing research in Statistics has filled my life with enrichment, happiness, excitement and challenges. There are a number of people who have profound impact on my knowledge, ideas, and career goal, and to whom I am deeply indebted. Without them, my journey would not be so wonderful.

My deepest gratitude is extended to my advisor, Professor R. Dennis Cook. This thesis would not be accomplished without his tireless guidance, great inspiration, and tremendous support. He opened the door leading to academic research for me, and has continually inspired me with his scholarly wisdom and vision throughout the dissertation process. He has conveyed an innovative spirit in research and scholarship, a passion in teaching, and an earnest attitude in daily work. From him, I have seen a genius character, a scholar's erudition, and a mentor's noble personality and kind heart. These qualities have truly influenced me and made my life with vigor, dream, and positive energy. He is and will always be my role model.

My appreciation also goes to Professor Glen Meeden and Professor Christopher Nachtsheim for being my dissertation committee members and writing reference letters for my job application. I am thankful to Professor Birgit Grund who serves on my committee and provides insightful suggestions along the work. I am also grateful to Professor Tiefeng Jiang for his great advice and support in my job application and academic profession. I am indebted to Professor Galin Jones for enrolling me into the

graduate program of Statistics, to Professor Lan Wang for her kind help and encouragement during my graduate study, to Professor Adam Rothman for allowing me to audit his interesting lectures and sharing the lecture notes, and to Seth Mayottee for providing me with technical support on computer issues.

I would like to express my gratitude to all my friends and classmates who have helped me in one way or the other. Thank Dr. Xin Chen, Dr. Kofi Placid Adragni, and Dr. Zhihua Su for sharing their experiences and encouraging my grit in academic pursuit. Thank Dr. Liliana Forzani for her insightful discussion on dimension reduction problems. Thanks Dr. Qiqi Deng for her help with my internship application. Thank my officemates who brought me a lot of fun and made my PhD life more colorful. Thank my dear friends Joy Bai, Dr. Qian Li, Dr. Lu Xu, Shuling Li, Yalin Li, Junyan Shen, Dr. Shan Hu, Huimin Liu, Yingbo Hu, Sara Hosch, Nicholle Brokke, Dr. Chenxiao Da, Dr. Xingyuan Zeng, Jing Li, and many others for their great care and support.

The last but not the least, I would like to thank my parents for their fostering, and loving and caring me every time, every place, and in every way they can. My special thanks also go to my younger brother Deshi Ding, for sharing his ideas, values, and sense of humor with me, and to Wei Qian for all his love and support.

Dedication

To my mother Zhizhen Chen and my father Xiaochen Ding.

Abstract

Data with complex structures, such as array-valued predictors, or responses, are commonly encountered in modern statistical applications. Such data typically contain intrinsic relationship among the entries of each array-valued variable. Conventional sufficient dimension reduction (SDR) methods cannot efficiently utilize the data structures and are inappropriate for the complex data. In this thesis, we propose a class of sufficient dimension reduction methods, including model-based dimension reduction methods: dimension folding principal component analysis (PCA) and dimension folding principal fitted components (PFC), moment-based sufficient dimension reduction methods: tensor sliced inverse regression (SIR), and envelope methods to tackle data with array-valued predictors, or responses. The proposed methods can simultaneously reduce a predictor's, or a response's, multiple dimensions without losing any information in prediction or classification. We study the asymptotic properties of these methods and demonstrate their efficiency in both theoretical and numerical studies.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

With the rapid development of data storage and computing technology, high dimensional data are frequently collected in a large variety of areas, such as biomedical engineering, neuroimaging, genomics, social media analysis, and high frequency finance. Dimension reduction is among major techniques in studying high dimensional data.

The basic idea of dimension reduction is to reduce the number of random variables in a dataset from a high dimensional space to a low dimensional space. Considerable dimension reduction methods have been studied in literature. For instance, principal component analysis (PCA) can be considered as one of the earliest and the most commonly used dimension reduction methods in application. In addition, factor analysis, projection pursuit, independent component analysis, certain nonlinear dimension reduction techniques such as principal curves and multidimensional scaling, and sufficient dimension reduction (SDR) approaches are also popular in practice.

1.1 SDR

In this thesis, we study dimension reduction methods mainly under the framework of sufficient dimension reduction. Sufficient dimension reduction is introduced by Cook (1994, 1998a). It is a paradigm of exploring dependency information through dimension

reduction. Let $X \in \mathbb{R}^p$ be a p-dimensional random predictor vector and $Y \in \mathbb{R}^1$ be a response variable. Typical statistical problems study the relationship between Y and X in terms of the conditional distribution $Y|X$. When p is large, however, most statistical methods suffer the issue, so called the "curse of dimensionality". Therefore, it is desirable to reduce the dimension of the predictor while preserving the full relationship between Y and X. Sufficient dimension reduction serves to achieve this goal. The key idea of SDR is to reduce the dimension of the predictor vector X by replacing it with its projection $\mathrm{P}_{\mathcal{S}}X$ onto a subspace $\mathcal{S}$ of the predictor space without loss of information on the conditional distribution of $Y|X$. This requirement can be stated as $Y \perp\!\!\!\perp X|\mathrm{P}_{\mathcal{S}}X$, where ' $\perp\!\!\!\perp$ ' indicates independence. Under mild conditions (Cook 1998a), the intersection of all such dimension reduction subspaces $\mathcal{S} \subseteq \mathbb{R}^p$ is also a dimension reduction subspace and is called the central subspace, denoted as $\mathcal{S}_{Y|X}$.

In some cases, one might not concern sufficient reduction for the full conditional distribution $Y|X$, but only for certain aspects of the dependency of Y on X. For instance, one might be only interested in the conditional mean of $Y|X$, denoted as $\mathrm{E}(Y|X)$. In this case, a dimension reduction subspace is defined as the subspace $\mathcal{S}^{'} \subseteq \mathbb{R}^p$ such that $\mathrm{E}(Y|X) = \mathrm{E}(Y|\mathrm{P}_{\mathcal{S}'}X)$. Again, the smallest dimension reduction subspace, which is the intersection of all such dimension reduction subspaces, is of interest and it is called the central mean subspace (Cook and Li 2002), denotes as $\mathcal{S}_{\mathrm{E}(Y|X)}$. Depending on one's specific need, the goal of SDR is to estimate $\mathcal{S}_{Y|X}$, or $\mathcal{S}_{\mathrm{E}(Y|X)}$, or the smallest dimension reduction for the target of interest.

Research into sufficient dimension reduction has gained considerable momentum since early '90s. Numerous dimension reduction methods can be incorporated into the rationale of sufficient dimension reduction under certain conditions. Sliced inverse regression (SIR; Li 1991) and sliced average variance estimation (SAVE; Cook and Weisberg 1991) are two early techniques for sufficient dimension reduction. Since then, principal Hessian directions (PHD; Li 1992, Cook 1998b), iterative Hessian transformations (Cook and Li 2002), SDR for the conditional k-th moment (Yin and Cook

2002), SDR with categorical predictors (Chiaromonte et al. 2002), minimum average variance estimation (MAVE; Xia et al. 2002), bootstrap dimension reduction (Ye and Weiss 2003), inverse regression estimation (IRE; Cook and Ni 2005), directional regression (DR; Li and Wang 2007), SDR for non-elliptical distributed predictors (Li and Dong 2009), semiparametric SDR (Ma and Zhu 2012), and many other methods were developed to either improve the estimation for SDR, or to perform SDR under different settings. While a few of the proposed methods employ nonparametric and semiparametric techniques for estimation, most of them use the first two moments of $X|Y$ for estimation, so called moment-based methods. In contrast, Cook (2007), and Cook and Forzani (2008, 2009) presented model-based SDR techniques, including principal fitted components (PFC), that give the maximum likelihood estimators (MLE) of the central subspace based on normal inverse models of X on Y. Model-based SDR inherits the optimal properties from maximum likelihood estimation and thus is more efficient than moment-based methods when the normality assumption holds. More recently, Cook et al. (2010, 2013) proposed a nascent research area, "Envelope models", that combines the idea of SDR with multivariate analysis to achieve substantial gains in efficiency.

1.2 Outline

Although dimension reduction topics have been widely studied, the methods mainly focus on a simple data structure: $Y \in \mathbb{R}^1$ and $X \in \mathbb{R}^p$. In modern statistical applications, however, one often encounters more complex data structures, such as data with matrix- or array-valued predictors, or responses. In this thesis, we focus on such data and develop model-based SDR, moment-based SDR, and envelope models for array-valued data.

In Chapter 2, we propose model-based dimension folding methods mainly for data with matrix-valued predictors. The methods can be treated as extensions of conventional principal components analysis (PCA) and principal fitted components (PFC). We refer

to them as dimension folding PCA and dimension folding PFC. The proposed methods can simultaneously reduce a predictor's multiple dimensions and inherit asymptotic properties from maximum likelihood estimation. They provide robust estimation and are computationally efficient. Dimension folding PFC gains further efficiency by effective use of the response information. Both theoretical and numerical results are provided to demonstrate the advantages.

In Chapter 3, we develop an efficient moment-based SDR method by extending SIR to general array (tensor)-valued predictors and refer to it as tensor SIR. Tensor SIR is constructed based on tensor decomposition to reduce an array-valued predictor's multiple dimensions simultaneously. The proposed method provides fast and efficient estimation. It circumvents high-dimensional covariance matrix inversion that researchers often suffer when dealing with such data. We further investigate its asymptotic properties and show its advantages by simulation studies and a real data application.

Inspired by the idea of envelopes proposed by Cook et al. (2010), we establish matrix-variate regressions and their envelope models for data with matrix-valued predictors and responses in Chapter 4. The proposed methods can be naturally extended to array-valued regressions for array-valued predictors and responses. We study the estimation procedures and their asymptotic properties for the cases - with and without envelope structures. Under the envelope framework, immaterial information can be eliminated in estimation and the number of parameters can be notably reduced when the matrix-variate dimension is large. Therefore, the estimation can be much more accurate and efficient. We investigate these properties by both theoretical and numerical studies.

In chapter 5, we discuss some future works regarding SDR for complex data structures.

1.3 Notations

To facilitate our discussion, the following notations are used in the thesis. The symbol $U \perp\!\!\!\perp V|Z$ indicates the conditional independence of U and V given Z, and $\sim$ means identically distributed. For positive integers p and q, $\mathbb{R}^{p\times q}$ denotes the class of all $p \times q$ matrices. For $A \in \mathbb{R}^{p\times q}$, span$(A)$ denotes the subspace spanned by the columns of A, $\mathrm{P}_A = A(A^TA)^\dagger A^T$ denotes a projection operator onto span(A) relative to the usual inner product, and $\mathrm{Q}_A = I_p - \mathrm{P}_A$, where $\dagger$ is the Moore-Penrose inverse, and I_p is the $p \times p$ identity matrix. For a symmetric and positive definite matrix $\Sigma \in \mathbb{R}^{p\times p}$, $P_{A(\Sigma)} = A(A^T\Sigma A)^\dagger A^T\Sigma$ denotes a projection operator onto span(A) relative to the Σ-inner product. The Σ-inner product is defined as $< X, X >= X^T\Sigma X$ for $X \in \mathbb{R}^p$.

For a subspace $\mathcal{S} \subseteq \mathbb{R}^p$ and a square matrix $B \in \mathbb{R}^{p\times p}$, $B\mathcal{S} = \{B\nu : \nu \in \mathcal{S}\}$. The symbol $\mathcal{S}^\perp$ stands for the orthogonal complement of $\mathcal{S}$ relative to the usual inner product. A basis matrix for $\mathcal{S}$ is any semi-orthogonal matrix whose columns form a basis for $\mathcal{S}$. A matrix $A \in \mathbb{R}^{p\times q}(q < p)$ is a semi-orthogonal matrix if $A^TA = I_q$. When A is a basis matrix of $\mathcal{S}$, we use A_0 to denote a semi-orthogonal basis of $\mathcal{S}^\perp$, where (A, A_0) forms an orthogonal matrix. The notation $\mathcal{G}(u, r)$ stands for the Grassman manifold of dimension u in $\mathbb{R}^r$, which is a set of all u dimensional subspaces in $\mathbb{R}^r$.

For two square matrices $B, C \in \mathbb{R}^{p\times p}$, $\mathcal{S}_d(B)$ denotes the span of the d eigenvectors of B corresponding to its d largest eigenvalues, and $\mathcal{S}_d(B, C) = B^{-\frac{1}{2}}\mathcal{S}_d(B^{-\frac{1}{2}}CB^{-\frac{1}{2}})$. When $B \geq 0$, $|B|_0$ indicates the product of non-zero eigenvalues of B. The notation $\otimes$ means the Kronecker product, and $\oplus$ denotes the direct sum of subspaces. For instance, the direct sum of m subspaces $\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_m$ is defined as $\bigoplus_{i=1}^{m} \mathcal{V}_i = \{v_1 + v_2 + \cdots + v_m : v_1 \in \mathcal{V}_1, v_2 \in \mathcal{V}_2, \ldots, v_m \in \mathcal{V}_m\}$. The symbol $||\cdot||_{\mathrm{F}}$ stands for the Frobenius norm of a matrix or an array.

We use "vec" to indicate the vectorization operator that stacks the columns of a matrix into a vector, use "vech" to denote the half vectorization operator that stacks elements from the upper triangular or lower triangular part of a symmetric matrix into

a vector column-wise, and use "avar" to denote the asymptotic covariance matrix.

Chapter 2

Dimension folding PCA and PFC

In modern statistical applications, data with matrix- or array-valued predictors, such as longitudinal data with p predictors observed over q times, EEG (electroencephalography) data, FMRI (functional Magnetic Resonance Imaging) data and general image data, are often encountered. The EEG data studied by Li et al. (2010) contains 122 subjects that are divided into alcoholic and control groups. For each subject, the predictor contains measurements from 64 channels of electrodes placed on the subject's scalp and sampled at 256 times. Thus the predictor is formed as a matrix of dimension 256×64, and the response is a binary variable indicating groups. The data structure can be represented as $Y \in \mathbb{R}^1$ and $X \in \mathbb{R}^{p_L \times p_R}$. Traditional dimension reduction methods are inadequate to analyze such complex data structures since they can only reduce the predictor's dimension by vectorizing it, thus losing important information on its matrix structure.

In face recognition and image analysis, certain unsupervised dimension reduction techniques were developed to deal with such data, based only on the marginal distribution of X. These methods include 2DPCA (Yang et al. 2004), $(2D)^2$PCA (Zhang

and Zhou 2005), GLRAM (Ye 2005), Unified PCA (Shan et al. 2008), probabilistic higher-order PCA (Yu, Bi and Ye 2011), etc. Li et al. (2010) proposed supervised and moment-based dimension folding approaches that extend SIR, SAVE, and DR to data with matrix-valued predictors, in order to reduce the predictor's row and column dimensions simultaneously without loss of information on $Y|X$. The idea of dimension folding can be expressed as the condition: $Y \perp\!\!\!\perp X|\Gamma_2^T X\Gamma_1$ or, equivalently, $Y \perp\!\!\!\perp \text{vec}(X)|(\Gamma_1 \otimes \Gamma_2)^T\text{vec}(X)$, where $\Gamma_1 \in \mathbb{R}^{p_R\times d_R}$ and $\Gamma_2 \in \mathbb{R}^{p_L\times d_L}$ have the smallest column dimensions d_R and d_L ($d_R \leq p_R$, $d_L \leq p_L$). The subspace $\text{span}(\Gamma_1 \otimes \Gamma_2)$ or, equivalently, $\text{span}(\Gamma_1) \otimes \text{span}(\Gamma_2)$ is called the central dimension folding (CDF) subspace for $Y|X$, and denoted as $\mathcal{S}_{Y|\circ X\circ}$.

Like conventional moment-based methods, moment-based dimension folding approaches are generally more efficient for discrete than for continuous responses, since their performance depends on how to slice the response variable in order to estimate the conditional mean or variance of $X|Y$. The estimation can be inadequate if the number of slices is not selected properly. Moreover, the moment-based dimension folding methods may not possess good asymptotic properties since they require inverting the high dimensional covariance matrix $\hat{\Sigma} = \widehat{\text{cov}}[\text{vec}(X)]$. When the predictor X contains a large number of rows and columns, computational complexity and singularity issues intrude. As a result, pre-screening is often necessary. To resolve these issues and improve efficiency, we propose model-based dimension folding methods, to be called dimension folding PCA and dimension folding PFC, that retain the key idea of dimension folding and obtain the MLE of the central dimension folding subspace. Dimension folding PFC gains further efficiency by effective use of the response information. The proposed methods circumvent directly inverting $\hat{\Sigma}$ and thus are more applicable to high dimensional data. In addition, dimension folding PCA and PFC provide robust estimators. They can be treated as generalized versions of conventional PCA and PFC since they include them as special cases.

The remainder of this chapter is organized as follows. In Section 2.1 we introduce

dimension folding PCA and its estimation. Section 2.2 is devoted to the development of dimension folding PFC. Section 2.3 provides robustness results. Prediction methods are discussed in Section 2.4. Section 2.5 and 2.6 contain illustrations of the performance of our methods with simulation studies and data analysis. Discussion is given in Section 2.7. Technical details are given in Section 2.8.

2.1 Dimension folding PCA

Dimension folding PCA is a preliminary step to developing dimension folding PFC. It performs dimension reduction for data with matrix-valued predictors by reducing the predictor's row and column dimensions simultaneously, so the predictor's matrix information can be preserved. It is built on a normal inverse model of the predictor $X \in \mathbb{R}^{p_L \times p_R}$ on a latent matrix $\nu \in \mathbb{R}^{d_L \times d_R}$ and provides the MLE of the central dimension folding subspace.

Here is a brief review of conventional PCA methods. PCA was originally considered as a well-established data-analytic method not associated with any probabilistic model. Model-based PCA can be traced back to Tipping and Bishop (1999), where the PCA model was formulated as

$$X = \mu + \Gamma\nu + \sigma\varepsilon. \tag{2.1}$$

In their case, $X \in \mathbb{R}^p$ is the predictor vector, $\mu \in \mathbb{R}^p$ is the overall mean of X, $\Gamma \in \mathbb{R}^{p\times d}(d \leq p)$ is a coefficient matrix with rank d, $\nu \in \mathbb{R}^d$ is a latent random vector, and $\varepsilon \in \mathbb{R}^p$ is the random error. Additionally, ν and ε are assumed to be independent and both have standard multivariate normal distributions with zero means and identity covariance matrices. A random error with this structure is called an isotropic error. The identity covariance assumption for ν is not a restriction, since one can always combine a non-identity covariance matrix with Γ. Thus, the parameter Γ itself is not identified but span(Γ) is identified.

Under (2.1), it can be shown that the maximum likelihood estimator of span(Γ)

corresponds to the subspace spanned by the first d eigenvectors of the sample covariance matrix $\hat{\Sigma}$ of X, which is the principal subspace obtained from data-analytic PCA. Cook (2007) proposed that when the latent variable ν is replaced by some fixed, centered but unobserved values $\nu_1, ..., \nu_n$, (2.1) can be considered as the regression of X on ν. Then $R(X) = \Gamma^T X$ is a sufficient reduction satisfying $X|\Gamma^T X, \nu \sim X|\Gamma^T X$, where '$\sim$' stands for equivalence. The MLE of span(Γ) is the same as the estimator obtained from (2.1) with the normal assumption for ν.

2.1.1 Formulation of dimension folding PCA

Dimension folding PCA incorporates the idea of dimension folding into the conventional PCA model (2.1). To achieve this, we assume that the matrix-valued predictor X is matrix normally distributed and has some intrinsic structure among its rows and columns to convey its matrix structure. The model is built on the inverse regression of the predictor as

$$X = \mu + \Gamma_2 \nu \Gamma_1^T + \sigma\varepsilon, \tag{2.2}$$

where $X \in \mathbb{R}^{p_L \times p_R}$, $\Gamma_1 \in \mathbb{R}^{p_R \times d_R}$ $(d_R \leq p_R)$ and $\Gamma_2 \in \mathbb{R}^{p_L \times d_L}$ $(d_L \leq p_L)$ are semi-orthogonal matrices that reduce the column and row dimensions of X, $\mu \in \mathbb{R}^{p_L \times p_R}$ is the overall mean of X, and $\nu \in \mathbb{R}^{d_L \times d_R}$ is a latent matrix with mean zero. The random error ε is assumed to be independent of ν and have a matrix normal distribution. The matrix normal distribution is briefly reviewed in the appendix. As dimension folding PCA is a starting model, we simplify the error to be isotropic, so ε is $N_{p_L \times p_R}(0_{p_L \times p_R}, I_{p_R}, I_{p_L})$. More general error structures will be discussed in the dimension folding PFC section. In (2.2), neither Γ_1 nor Γ_2 is identified: if Γ_1, Γ_2 and ν are replaced by $\Gamma_2 A_2$, $\Gamma_1 A_1$ and $A_2^{-1}\nu(A_1^T)^{-1}$, equation (2.2) remains the same, where A_1 and A_2 are any nonsingular matrices. Thus, the dimension folding PCA model depends on Γ_1 and Γ_2 only through their column spaces. Under (2.2), ν contains the coordinates of the centered conditional mean $\mathrm{E}(X|\nu) - \mu$ relative to Γ_1 and Γ_2, and the relationship $\mathrm{E}(X|\nu) - \mu = P_{\Gamma_2}[\mathrm{E}(X|\nu) -$

$\mu]P_{\Gamma_1}$ holds. Therefore, the predictor's important row and column signals are preserved by span(Γ_1) and span(Γ_2).

Model (2.2) reflects the homogeneous characteristic among the rows and columns of the centered conditional mean $\mathrm{E}(X|\nu) - \mu$, because its column information is retained by the same Γ_1 over all rows and its row information is preserved by Γ_2 over all of its columns. This feature can be found in many data sets with matrix-valued predictors. For example, in the EEG data, the rows and columns of the predictors indicate the time and location measurements for each subject. It is reasonable to believe that the signals provided by the scalp locations are consistent over time, and vice versa. This is one major distinction between dimension folding PCA and conventional PCA, which omits the predictor's intrinsic matrix information and simply converts it to a vector. In addition to preserving the predictors' matrix structure, another benefit of (2.2) is to greatly reduce number of parameters in estimation and improve accuracy. Meanwhile, when the column dimension of X is one, (2.2) is equivalent to the conventional PCA model (2.1) under the setting of Cook (2007). Thus, it is a generalization of the conventional model.

Model (2.2) can also be written in a vectorization version as

$$\mathrm{vec}(X) = \mathrm{vec}(\mu) + (\Gamma_1 \otimes \Gamma_2)\mathrm{vec}(\nu) + \sigma\mathrm{vec}(\varepsilon). \tag{2.3}$$

Here vec(ε) has a multivariate normal distribution $N(0_{p_L p_R}, I_{p_L p_R})$. In this way, dimension folding PCA implies that under the isotropic error assumption, the centered conditional means $\mathrm{E}[\mathrm{vec}(X|\nu)] - \mathrm{vec}(\mu)$ fall in the subspace spanned by the columns of $\Gamma_1 \otimes \Gamma_2$.

A proposition connects the inverse regression models (2.2) and (2.3) to the dimension folding conditions.

Proposition 2.1. *(a) Under (2.2), the distribution of $\nu|X$ is the same as the distribution of $\nu|\Gamma_2^T X \Gamma_1$ over all values of X; (b) under (2.3), the distribution of $\nu|vec(X)$ is the same as the distribution of $\nu|(\Gamma_1 \otimes \Gamma_2)^T vec(X)$ for all values of X.*

Based on Proposition 2.1, $R(X) = \Gamma_2^T X \Gamma_1$ is a sufficient reduction (folding) satisfying $X \perp\!\!\!\perp \nu \mid \Gamma_2^T X \Gamma_1$. Since both Γ_1 and Γ_2 have the minimum column dimensions, $\text{span}(\Gamma_1 \otimes \Gamma_2)$ forms the central dimension folding subspace $\mathcal{S}_{\nu|\circ X\circ}$.

2.1.2 Estimation of dimension folding PCA

The parameters in (2.2) are estimated based on maximum likelihood. We assume that for each observation X_i of X, $i = 1, ..., n$, there is a corresponding coordinate matrix ν_i, such that $X_i = \mu + \Gamma_2 \nu_i \Gamma_1^T + \sigma\varepsilon$, where ν_i is fixed and $\sum_{i=1}^{n} \nu_i = 0$ without loss of generality. In general, we are not able to find a closed-form solution for the MLE of the central dimension folding subspace. Yet we can apply a fast and stable algorithm that uses three eigen-based iterations and provides connections to the conventional PCA model.

For an independent sample $\{X_i\}$, according to (2.23), the full log likelihood of (2.2) can be written as

$$\begin{aligned} l(\mu, \mathcal{S}_{\Gamma_1}, \mathcal{S}_{\Gamma_2}, \sigma^2, \nu_1, \cdots, \nu_n) = &-\frac{np_L p_R}{2}\log(2\pi) - \frac{np_L p_R}{2}\log\sigma^2 \\ &- \frac{1}{2\sigma^2}\sum_{i=1}^{n} \text{tr}[(X_i - \mu - \Gamma_2\nu_i\Gamma_1^T)^T(X_i - \mu - \Gamma_2\nu_i\Gamma_1^T)], \end{aligned} \tag{2.4}$$

where $\mathcal{S}_{\Gamma_1}$ and $\mathcal{S}_{\Gamma_2}$ denote the column spaces $\text{span}(\Gamma_1)$ and $\text{span}(\Gamma_2)$. It is easy to see that the MLE $\hat{\mu} = \bar{X}$ since $\sum_{i=1}^{n} \nu_i = 0$. Then for any arbitrary σ^2, maximizing (2.4) is equivalent to minimizing $\sum_{i=1}^{n} \text{tr}[(X_i - \bar{X} - \Gamma_2\nu_i\Gamma_1^T)^T(X_i - \bar{X} - \Gamma_2\nu_i\Gamma_1^T)]$, which can be solved based on the following.

Proposition 2.2. *Suppose that $X_i \in \mathbb{R}^{p_L \times p_R}$, $i = 1, ..., n$, are observed matrices. Let $(\hat{\Gamma}_1, \hat{\Gamma}_2, \hat{\nu}_1, ..., \hat{\nu}_n)$ be minimizers of*

$$\sum_{i=1}^{n} \text{tr}[(X_i - G_2\omega_i G_1^T)^T(X_i - G_2\omega_i G_1^T)] \tag{2.5}$$

over all $G_1 \in \mathbb{R}^{p_R \times d_R}$, $G_2 \in \mathbb{R}^{p_L \times d_L}$, and $\omega_i \in \mathbb{R}^{d_L \times d_R}$, $i = 1, ..., n$. Then

(i) For fixed G_1, the columns of the minimizer $\hat{\Gamma}_2$ are given by the d_L eigenvectors of the matrix $\hat{\Sigma}_L = \sum_{i=1}^{n} X_i \mathrm{P}_1 X_i^T/n$ corresponding to its d_L largest nonzero eigenvalues, where $\mathrm{P}_1 = G_1 G_1^T$.

(ii) For fixed G_2, the columns of the minimizer $\hat{\Gamma}_1$ consist of the d_R eigenvectors of the matrix $\hat{\Sigma}_R = \sum_{i=1}^{n} X_i^T \mathrm{P}_2 X_i/n$ corresponding to its d_R largest nonzero eigenvalues, where $\mathrm{P}_2 = G_2 G_2^T$.

(iii) For fixed G_1 and G_2, the minimizer $\hat{\nu}_i = G_2^T X_i G_1$, $i = 1, ..., n$.

Based on Proposition 2.2, for fixed G_1 and G_2, if ω_i is replaced by $\hat{\nu}_i = G_2^T X_i G_1$, the objective function (2.5) is $L_1 = \mathrm{tr}(\sum_{i=1}^{n} X_i^T X_i) - \mathrm{tr}[\sum_{i=1}^{n} (X_i^T \mathrm{P}_2 X_i)\mathrm{P}_1]$. Then for fixed P_2, L_1 is minimized by choosing the columns of G_1 to be the first d_R eigenvectors of $\sum_{i=1}^{n} X_i^T \mathrm{P}_2 X_i$. So we need to choose P_2 to minimize $L_{12} = \sum_{k=1}^{d_R} \lambda_k(\sum_{i=1}^{n} X_i^T \mathrm{P}_2 X_i)$, where $\lambda_k(A)$ indicates the k-th eigenvalue of A. This can be treated as an optimization problem over a Grassmann manifold but it is hard to solve because eigenvalues are involved in the objective function. Instead, we apply an iterative algorithm that can solve the problem efficiently. We assume that the predictors are centered.

1. Generate an initial value of $\Gamma_{10} \in \mathbb{R}^{p_L \times d_L}$ and let $\hat{\Gamma}_1 = \Gamma_{10}$.

2. For given $\hat{\Gamma}_1$, compute the matrix $\hat{\Sigma}_L = \sum_{i=1}^{n} X_i \hat{\Gamma}_1 \hat{\Gamma}_1^T X_i^T/n$ and find its first d_L eigenvectors, denoted as $\hat{v}_1, \hat{v}_2, ..., \hat{v}_{d_L}$. Estimate Γ_2 as $\hat{\Gamma}_2 = [\hat{v}_1, \hat{v}_2, ..., \hat{v}_{d_L}]$.

3. For given $\hat{\Gamma}_2$, compute $\hat{\Sigma}_R = \sum_{i=1}^{n} X_i^T \hat{\Gamma}_2 \hat{\Gamma}_2^T X_i/n$; find the first d_R eigenvectors of $\hat{\Sigma}_R$, denoted as $\hat{l}_1, \hat{l}_2, ..., \hat{l}_{d_R}$, which form the columns of $\hat{\Gamma}_1$ as $\hat{\Gamma}_1 = [\hat{l}_1, \hat{l}_2, ..., \hat{l}_{d_R}]$.

4. For given $\hat{\Gamma}_1$ and $\hat{\Gamma}_2$, compute $\hat{\nu}_i = \hat{\Gamma}_2^T X_i \hat{\Gamma}_1$, $i = 1, ..., n$.

5. Repeat Step 2 to 4 and iterate each time using the updated $\hat{\Gamma}_1$ and $\hat{\Gamma}_2$ until $\sum_{i=1}^{n} \mathrm{tr}[(X_i - \Gamma_2 \nu_i \Gamma_1^T)^T (X_i - \Gamma_2 \nu_i \Gamma_1^T)]$ converges.

The MLE of the central dimension folding subspace $\mathcal{S}_{\nu|\circ X\circ}$ is then equal to $\mathrm{span}(\hat{\Gamma}_1) \otimes \mathrm{span}(\hat{\Gamma}_2)$. Consequently, $\hat{\sigma}^2$ is equal to $\frac{1}{np_L p_R} \sum_{i=1}^{n} \mathrm{tr}[(X_i - \hat{\Gamma}_2 \hat{\nu}_i \hat{\Gamma}_1^T)^T (X_i - \hat{\Gamma}_2 \hat{\nu}_i \hat{\Gamma}_1^T)]$. The

estimators obtained from the dimension folding model inherit the asymptotic properties of likelihood estimation under normality.

As with most optimization procedures, the proposed algorithm can convergence to a local minimum. It has a linear convergence rate. Our experience shows that the convergence behavior depends on the gaps between the eigenvalues of $\hat{\Sigma}_L$ and the gaps between the eigenvalues of $\hat{\Sigma}_R$. The larger the gaps, the more likely the algorithm obtains a global solution. Meanwhile, according to our empirical study, the algorithm is quite stable with use of random initial values of Γ_{10}. When a better initial value is required, one can choose the first d_R eigenvectors of $\sum_{i=1}^n X_i^T X_i/n$ as an initial Γ_{10}, where $\sum_{i=1}^n X_i^T X_i/n$ is the sample row covariance matrix of X.

The proposed estimation procedure has connections with conventional PCA and is easily interpreted. It can be seen that when the column reduction matrix Γ_1 is known, the estimator of the row reduction Γ_2 is the same as that of Γ in the conventional PCA model (2.1) with the original predictor X_i replaced by $X_i\Gamma_1$. Although here $X_i\Gamma_1$ is a matrix instead of a vector, the estimation logic remains the same. Similarly, if Γ_2 is known, the column reduction Γ_1 can be obtained from the conventional PC model with X_i replaced by $\Gamma_2^T X_i$.

Compared to conventional PCA, dimension folding PCA is computationally efficient for dealing with matrix-valued predictors. The algorithm has three major steps at each iteration. An efficient way to compute $\hat{\Sigma}_L$ is to perform multiplication for X_i and $\hat{\Gamma}_1$ first and then multiply it by its transpose. Thus, the total computation cost of $\hat{\Sigma}_L$ is $O(np_Ld_R(p_L + p_R))$. The eigen-decomposition of $\hat{\Sigma}_L$ requires $O(p_L^2 d_L)$ operations. Similarly, it takes $O(np_Rd_L(p_L + p_R))$ and $O(p_R^2 d_R)$ operations to compute $\hat{\Sigma}_R$ and its eigenspace. The computation of $\hat{\nu}_i$ is of order $O(p_Ld_R(p_R + d_L))$. Therefore, dimension folding PCA totally requires at most $O(\max(p_L, p_R)^2\max(d_L, d_R)nm)$ operations, where m is the number of iterations. Conventional PCA targeting vectorized X costs $O(p_L^2 p_R^2 n)$ operations, which is more expensive under the mild condition that $\max(d_L, d_R)m < \min(p_L, p_R)^2$.

2.1.3 Relationship with tensor PCA

Higher-order tensor decompositions have been widely studied in applied mathematics and engineering. Among them, the Tucker decomposition is considered as a higher order form of PCA, or tensor PCA (Kolda and Bader 2009). Here we discuss the connections of dimension folding PCA with tensor PCA. The key idea of tensor PCA is to decompose a tensor into a core tensor multiplied by a component matrix along each mode. Thus, in a two-mode tensor case where $X \in \mathbb{R}^{p_L \times p_R}$, we have $X \approx GCH^T$, where $C \in \mathbb{R}^{d_L \times d_R}$ is the core matricized two-way tensor, and $G \in \mathbb{R}^{p_L \times d_L}$ and $H \in \mathbb{R}^{p_R \times d_R}$ are the component matrices. If d_L and d_R are less than p_L and p_R, the core tensor C is considered as a compressed version of X. Thus, dimension reduction of the original tensor can be achieved. There are several ways to compute the Tucker decomposition. Major algorithms are developed to minimize the mean-squared loss function

$$f(G, H, C) = ||X - \hat{X}||_F^2 = ||X - GCH^T||_F^2. \tag{2.6}$$

This loss function has the equivalent form of the last term in our objective function (2.4). Kroonenberg and De Leeuw (1980) proposed an iterative least squares algorithm (ALS), called TUCKALS3 for computing a Tucker decomposition of three-way arrays. This method was further refined by De Lathauwer et al. (2000), where they enhanced the approximation by directly calculating the dominant subspaces rather than their individual singular vectors. From this aspect, the algorithm we presented for dimension folding PCA is equivalent to a sample version of the method in Lathauwer, Moor and Vandewalle (2000) for two-mode tensors.

Tensor PCA is a well-established data-analytic method but is not associated with any probabilistic model. Dimension folding PCA can be treated as a model-based tensor PCA. It gains properties from maximum likelihood estimation when the predictors are approximately normally distributed. The normality assumption, however, is not essential in our model and can be relaxed to a general distribution. In this case, dimension folding PCA is equivalent to tensor PCA. The robustness of the dimension folding

model regarding its normality assumption will be further discussed in Section 2.3.2.

2.2 Dimension folding PFC

Although dimension folding PCA can reduce the predictor's row and column dimensions simultaneously, it performs dimension folding marginally and the relationship between the predictor and the response is omitted. Instead of regressing X on a latent matrix ν, dimension folding PFC models the inverse regression of $X|Y$ and provides more informative estimation of the central dimension folding subspace $\mathcal{S}_{Y|\circ X\circ}$.

2.2.1 Formulation of dimension folding PFC

The dimension folding PFC model can be formed in several ways depending on the relations between the predictors and response. One way is to fit the inverse regression by taking the true model to be

$$X = \mu + \Gamma_2\beta_2 f(Y)\beta_1^T\Gamma_1^T + \varepsilon \tag{2.7}$$

or, equivalently,

$$\text{vec}(X) = \text{vec}(\mu) + (\Gamma_1 \otimes \Gamma_2)(\beta_1 \otimes \beta_2)\text{vec}(f(Y)) + \text{vec}(\varepsilon), \tag{2.8}$$

where $f(Y) \in \mathbb{R}^{r_L \times r_R}$ contains elements formalized as functions of Y, $\beta_1 \in \mathbb{R}^{d_R \times r_R}(d_R \leq r_R)$ and $\beta_2 \in \mathbb{R}^{d_L \times r_L}(d_L \leq r_L)$ are the coefficient matrices of rank d_R and d_L, and ε is the random error independent of Y. It can be isotropic following the matrix normal distribution $\sigma N_{p_L \times p_R}(0_{p_L \times p_R}, I_{p_R}, I_{p_L})$ or more general with $N_{p_L \times p_R}(0_{p_L \times p_R}, \Omega, M)$ error. In Section 2.3.2, we show that the normality assumption is not necessary in order to obtain consistent estimation. The other terms in (2.7) are defined as in Section 2.1.1. Based on (2.7), each coordinate X_{ij} of X is a linear function of the elements in $f(Y)$ plus a random error. In addition,

$$(\Gamma_2^T X\Gamma_1)_{ij} = (\Gamma_2^T\mu\Gamma_1)_{ij} + \sum_{k=1}^{r_L}\sum_{l=1}^{r_R}\beta_{ik}^{(2)}\beta_{lj}^{(1)}f(Y)_{kl} + (\Gamma_2^T\varepsilon\Gamma_1)_{ij},$$

where $\beta_{ik}^{(2)}$ denotes the ik-th element of β_2, $\beta_{lj}^{(1)}$ denotes the lj-th element of β_1^T, and $f(Y)_{kl}$ is the kl-th element of $f(Y)$, $i = 1, ..., d_L$, $j = 1, ..., d_R$. This shows a multiplicative coefficient structure.

The function $f(Y)$ is determinable in some cases, for instance when inverse response plots (Cook 1998 (Chapter 10)) of X_{ij} versus Y are informative about $f(Y)$, or when the response Y is categorical. In other cases, one can approximate $f(Y)$ by a series of basis functions or piecewise basis functions. Usually $f(Y)$ can be chosen as a diagonal matrix with dimension $r_L = r_R = r$. We use this matrix form in the rest of this chapter. When using polynomial approximations, $f(Y)$ is then a diagonal matrix with diagonal elements of $Y, Y^2, ..., Y^r$. Correspondingly, the conditional expectation $[\Gamma_2^T \mathrm{E}(X|Y)\Gamma_1]_{ij}$ is $(\Gamma_2^T \mu \Gamma_1)_{ij} + \sum_{k=1}^{r} \beta_{ik}^{(2)} \beta_{kj}^{(1)} Y^k$, which often captures the main regression shape of X on Y when r is relative large. In fact, in Section 2.3.1 we show that in order to receive a consistent estimator for the central dimension folding subspace, the selected fitting function does not need to be very close to the true function, it is only required to be correlated to it. This indicates that an approximation with a finite dimension for $f(Y)$ is generally adequate.

When the response Y is categorical, the fitting function $f(Y)$ can be naturally determined. For instance, suppose that Y has h categories, then $f(Y)$ can be simply chosen as a diagonal matrix of dimension $r = h-1$ and its k-th diagonal element can be specified as $\mathrm{diag}(f(Y))_k = I(Y \in J_k) - n_k/n$, $k = 1, ..., h-1$, where J_k indicates the k-th category, n_k is the number of observation in J_k, and $I(\cdot)$ is the indicator function. The sample solution of dimension folding PFC with a categorical response is not equivalent to that obtained by dimension folding SIR (Li et al. 2010). Dimension folding PFC is more efficient in estimation, does not involve computations relative to $\mathrm{vec}(X)$.

Compared with slicing-based methods, dimension folding PFC provides the flexibility to formulate the relationship between X and Y. It can more effectively use the response information by choosing an appropriate fitting function to perform dimension folding. Slicing function can be considered as one special choice for fitting $f(Y)$ but

it is generally less accurate when Y is continuous. A proposition identifies the central dimension folding subspace for the dimension folding model (2.7).

Proposition 2.3. *Under (2.7), when the random error ε is isotropic the central dimension folding subspace $\mathcal{S}_{Y|\circ X\circ} = span(\Gamma_1) \otimes span(\Gamma_2)$; when ε has a general matrix normal distribution $N_{p_L\times p_R}(0_{p_L\times p_R}, \Omega, M)$, the central dimension folding subspace $\mathcal{S}_{Y|\circ X\circ} = span(\Omega^{-1}\Gamma_1) \otimes span(M^{-1}\Gamma_2)$.*

Other ways to formulate the dimension folding PFC model are discussed in Section 2.7. We focus on estimating model (2.7) with both isotropic error and general error in the next section. Without loss of generality, the predictor X and the fitting function $f(Y)$ are assumed to be centered.

2.2.2 Estimation of dimension folding PFC

Isotropic error

When ε is isotropic with distribution $\sigma N_{p_L\times p_R}(0_{p_L\times p_R}, I_{p_R}, I_{p_L})$, the central dimension folding subspace $\mathcal{S}_{Y|\circ X\circ}$ is equal to $\text{span}(\Gamma_1) \otimes \text{span}(\Gamma_2)$. For a random sample of size n from (Y, X), the MLE of $\mathcal{S}_{Y|\circ X\circ}$ is obtained based on the log likelihood function of (2.7):

$$l(\mu, \mathcal{S}_{\Gamma_1}, \mathcal{S}_{\Gamma_2}, \sigma^2, \beta_1, \beta_2) = -\frac{np_Lp_R}{2}\log(2\pi) - \frac{np_Lp_R}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\times \sum_{i=1}^{n} \text{tr}((X_i - \mu - \Gamma_2\beta_2 f(Y_i)\beta_1^T\Gamma_1^T)^T(X_i - \mu - \Gamma_2\beta_2 f(Y_i)\beta_1^T\Gamma_1^T). \tag{2.9}$$

It is easy to see that the MLE $\hat{\mu} = \bar{X}$. Thus for any arbitrary σ^2, maximizing (2.9) is equivalent to minimizing the empirical expectation

$$\text{E}_n\{\text{tr}[(X - \Gamma_2\beta_2 f(Y)\beta_1^T\Gamma_1^T)^T(X - \Gamma_2\beta_2 f(Y)\beta_1^T\Gamma_1^T)]\} \tag{2.10}$$

over X and Y.

Proposition 2.4. *Suppose that $X \in \mathbb{R}^{p_L \times p_R}$ is a random matrix and $Y \in \mathbb{R}^1$ is a random variable. Let $(\hat{\Gamma}_1, \hat{\Gamma}_2, \hat{\beta}_1, \hat{\beta}_2)$ be minimizers of*

$$\mathrm{E}_n\{\mathrm{tr}[(X - G_2 b_2 f(Y) b_1^T G_1^T)^T (X - G_2 b_2 f(Y) b_1^T G_1^T)]\}. \tag{2.11}$$

over all $G_1 \in \mathbb{R}^{p_R \times d_R}$, $G_2 \in \mathbb{R}^{p_L \times d_L}$, $b_1 \in \mathbb{R}^{d_R \times r_R}$, and $b_2 \in \mathbb{R}^{d_L \times r_L}$. Then
(i) For fixed G_1 and b_1, the columns of the minimizer $\hat{\Gamma}_2$ over G_2 are given by the d_L eigenvectors of the matrix

$$\Sigma_{\mathrm{fit}_L} = \mathrm{E}_n(XG_1 f^{*^T})[\mathrm{E}_n(f^* f^{*^T})]^{-1}\mathrm{E}_n(f^* G_1^T X^T)$$

corresponding to its d_L largest nonzero eigenvalues, where $f^* = f(Y)b_1^T$. The minimizer $\hat{\beta}_2 = \hat{\Gamma}_2^T \mathrm{E}_n(XG_1 f^{*^T})[\mathrm{E}_n(f^* f^{*^T})]^{-1}$.
(ii) For fixed G_2 and b_2, the columns of the minimizer $\hat{\Gamma}_1$ over G_1 consist of the d_R eigenvectors of the matrix

$$\Sigma_{\mathrm{fit}_R} = \mathrm{E}_n(X^T G_2 f^*)[\mathrm{E}_n(f^{*^T} f^*)]^{-1}\mathrm{E}_n(f^{*^T} G_2^T X)$$

corresponding to its d_R largest nonzero eigenvalues, where $f^* = b_2 f(Y)$. The minimizer $\hat{\beta}_1 = \hat{\Gamma}_1^T \mathrm{E}_n(X^T G_2 f^*)[\mathrm{E}_n(f^{*^T} f^*)]^{-1}$.

Similar to Proposition 2.2, after replacing G_2 and b_2 with their optimum solutions $\hat{\Gamma}_2$ and $\hat{\beta}_2$ obtained from Proposition 2.4(i), the problem becomes an optimization over a Grassmann manifold, but it is complicated to solve. Instead, we choose a simple iterative algorithm to estimate the likelihood function (2.9) as follows.

1. Generate initial values of Γ_{10} and β_{10} and let $\hat{\Gamma}_1 = \Gamma_{10}$ and $\hat{\beta}_1 = \beta_{10}$.

2. For given $\hat{\Gamma}_1$ and $\hat{\beta}_1$, compute the matrix $\hat{\Sigma}_{\mathrm{fit}_L} = \mathbb{X}_L^T \mathrm{P}_{\mathbb{F}_L} \mathbb{X}_L / n$, where $\mathbb{X}_L = (X_1\hat{\Gamma}_1, ..., X_n\hat{\Gamma}_1)^T$, $\mathbb{F}_L = (f_1^*, ..., f_n^*)^T$ with $f_i^* = f(Y_i)\hat{\beta}_1^T$. Then the term $\mathrm{P}_{\mathbb{F}_L}\mathbb{X}_L$ represents the fitted values from the multivariate regression of $X\hat{\Gamma}_1$ on $f(Y)\hat{\beta}_1^T$. Therefore, $\hat{\Sigma}_{\mathrm{fit}_L}$ is the sample column covariance matrix of the fitted values of $X\hat{\Gamma}_1$. Then the columns of $\hat{\Gamma}_2$ are estimated by the first d_L eigenvectors of $\hat{\Sigma}_{\mathrm{fit}_L}$ and $\hat{\beta}_2 = \hat{\Gamma}_2^T \mathbb{X}_L^T \mathbb{F}_L (\mathbb{F}_L^T \mathbb{F}_L)^{-1}$.

3. For given $\hat{\Gamma}_2$ and $\hat{\beta}_2$, compute the matrix $\hat{\Sigma}_{\text{fit}_R} = \mathbb{X}_R^T \mathrm{P}_{\mathbb{F}_R} \mathbb{X}_R / n$, where $\mathbb{X}_R = (X_1^T \hat{\Gamma}_2, ..., X_n^T \hat{\Gamma}_2)^T$, $\mathbb{F}_R = (f_1^{*^T}, ..., f_n^{*^T})^T$ with $f_i^* = \hat{\beta}_2 f(Y_i)$. The term $\mathbb{X}_R^T \mathrm{P}_{\mathbb{F}_R}$ represents the fitted values from the multivariate regression of $\hat{\Gamma}_2^T X$ on $\hat{\beta}_2 f(Y)$. Then $\hat{\Sigma}_{\text{fit}_R}$ represents the sample row covariance matrix of the fitted values of $\hat{\Gamma}_2^T X$. The columns of $\hat{\Gamma}_1$ are given by the first d_R eigenvectors of $\hat{\Sigma}_{\text{fit}_R}$ and $\hat{\beta}_1 = \hat{\Gamma}_1^T \mathbb{X}_R^T \mathbb{F}_R (\mathbb{F}_R^T \mathbb{F}_R)^{-1}$.

4. Repeat Steps 2-3 and iterate each time with the updated estimators until the objective function (2.10) converges.

The MLE of the central dimension folding subspace is then given by $\text{span}(\hat{\Gamma}_1) \otimes \text{span}(\hat{\Gamma}_2)$. Correspondingly, σ^2 is estimated by

$$\frac{1}{np_L p_R} \sum_{i=1}^{n} \text{tr}((X_i - \hat{\Gamma}_2 \hat{\beta}_2 f(Y_i) \hat{\beta}_1^T \hat{\Gamma}_1^T)^T (X_i - \hat{\Gamma}_2 \hat{\beta}_2 f(Y_i) \hat{\beta}_1^T \hat{\Gamma}_1^T)).$$

It can be seen that the estimators $\hat{\Gamma}_1$ and $\hat{\Gamma}_2$ obtained from dimension folding PFC have similar expressions as those achieved by dimension folding PCA. The only difference is that we perform eigen-decomposition for the sample row (column) covariance matrix of the fitted values of the linear regressions $\hat{\Gamma}_2^T X$ $(X\hat{\Gamma}_1)$ on $\hat{\beta}_2 f(Y)$ $(f(Y)\hat{\beta}_1^T)$. In this way, the redundant information of X that is not related to Y is eliminated. Thus, dimension folding PFC is more precise in estimation and prediction. The estimators obtained from this algorithm can be treated as a generalized version of the results attained in conventional PFC.

From a computational perspective, the proposed algorithm is more economical than conventional PFC and dimension folding SIR. Its major costs come from the computation of $\hat{\Sigma}_{\text{fit}_L}$ and $\hat{\Sigma}_{\text{fit}_R}$. For $\hat{\Sigma}_{\text{fit}_L}$, computing $\mathbb{X}_L$ and $\mathbb{F}_L$ requires $np_L p_R d_R$ and $nr_L r_R d_R$ operations, and computing $\mathbb{X}_L^T \mathbb{F}_L$ and $\mathbb{F}_L^T \mathbb{F}_L$ requires $nd_R p_L r_L$ and $nd_R r_L^2$ operations. The inverse of $\mathbb{F}_L^T \mathbb{F}_L$ costs $O(r_L^3)$. Therefore, the total cost of $\hat{\Sigma}_{\text{fit}_L}$ is at most $O(\max(nd_R, r_L)\max(p_L, p_R, r_L, r_R)^2)$. Similarly, the cost of $\hat{\Sigma}_{\text{fit}_R}$ is of order $O(\max(nd_L, r_R)\max(p_L, p_R, r_L, r_R)^2)$. Thus, dimension folding PFC with an isotropic

error requires at most $O(\max(nd_L, nd_R, r_L, r_R)\max(p_L, p_R, r_L, r_R)^2 m)$ operations with m iterations. Analogously, it can be shown that the computations of conventional PFC and dimension folding SIR targeting on $\text{vec}(X)$ take at least $O(\max(n, p_L p_R)\max(p_L p_R, r)r)$ and $O(p_L^2 p_R^2 \max(p_L p_R, n)k)$ operations, which are in general more than dimension folding PFC when p_L and p_R are relative large. Here r is the dimension of the fitting function in conventional PFC and k is the iteration number in dimension folding SIR.

General error

In this section, we consider a general error structure for ε with the matrix normal distribution $N_{p_L \times p_R}(0_{p_L \times p_R}, \Omega, M)$. Based on this covariance structure, the dimension folding models reveal another homogeneous characteristic among the predictor's rows and columns. Let $e_i = (0, ..., 0, 1, 0, ...0)^T$ denote the p_L-dimensional vector with ith component equal to one, $i = 1, ..., p_L$. Then $e_i^T X = (\text{vec}(e_i^T X))^T$ and $\text{var}(e_i^T X|Y) = \text{var}[(I \otimes e_i^T)\text{vec}(X)|Y] = (I \otimes e_i^T)(\Omega \otimes M)(I \otimes e_i) = m_{ii}\Omega$, where m_{ii} is the ith diagonal component of M. This implies that the conditional covariance matrices of the predictor's row vectors are all proportional to Ω. Similarly, the predictor's column conditional covariance matrices are all proportional to M. Thus, the second-order moments also reflect the predictor's intrinsic row and column structure, which the conventional PC and PFC models are not able to catch.

Another notable advantage is that the high-dimensional covariance matrix $\Sigma = \text{var}[\text{vec}(X)] \in \mathbb{R}^{p_L p_R \times p_L p_R}$ can be decomposed into two smaller matrices $\Omega \in \mathbb{R}^{p_R \times p_R}$ and $M \in \mathbb{R}^{p_L \times p_L}$. Therefore, one can circumvent inverting the sample covariance matrix $\hat{\Sigma}$ in estimation. This is beneficial when the sample size is relative small.

For estimation, note that if Ω and M are known, the problem reduces to the isotropic dimension folding PFC since one can standardize X_i to $Z_i = M^{-\frac{1}{2}} X_i \Omega^{-\frac{1}{2}}$. When Ω and M are unknown, the log likelihood function becomes:

$$l(\mu, \mathcal{S}_{\Gamma_1}, \mathcal{S}_{\Gamma_2}, \beta_1, \beta_2, \Omega, M)$$

$$= -\frac{np_Lp_R}{2}\log(2\pi) - \frac{np_L}{2}\log|\Omega| - \frac{np_R}{2}\log|M| \\ - \frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}\{\Omega^{-1}(X_i - \mu - \Gamma_2\beta_2 f(Y_i)\beta_1^T\Gamma_1^T)^T M^{-1}(X_i - \mu - \Gamma_2\beta_2 f(Y_i)\beta_1^T\Gamma_1^T)\}. \tag{2.12}$$

It is easy to see that the MLE of μ is $\bar{X}$. The other parameters can be estimated by alternating iterations with one group of parameters fixed. Let $\mathbb{X}_L = (X_1\Omega^{-\frac{1}{2}}, ..., X_n\Omega^{-\frac{1}{2}})^T$, $\mathbb{F}_L = (f(Y_1)\beta_1^T\Gamma_1^T\Omega^{-\frac{1}{2}}, ..., f(Y_n)\beta_1^T\Gamma_1^T\Omega^{-\frac{1}{2}})^T$, and $\mathbb{X}_R = (X_1^T M^{-\frac{1}{2}}, ..., X_n^T M^{-\frac{1}{2}})^T$, $\mathbb{F}_R = (f(Y_1)^T\beta_2^T\Gamma_2^T M^{-\frac{1}{2}}, ..., f(Y_n)^T\beta_2^T\Gamma_2^T M^{-\frac{1}{2}})^T$. Define $\hat{\Sigma}_{\mathrm{fit}_L} = \mathbb{X}_L^T \mathrm{P}_{\mathbb{F}_L} \mathbb{X}_L/np_R$, $\hat{M}_{\mathrm{res}} = \tilde{M} - \hat{\Sigma}_{\mathrm{fit}_L} = \mathbb{X}_L^T\mathbb{X}_L/np_R - \hat{\Sigma}_{\mathrm{fit}_L}$, and $\hat{\Sigma}_{\mathrm{fit}_R} = \mathbb{X}_R^T \mathrm{P}_{\mathbb{F}_R}\mathbb{X}_R/np_L$, $\hat{\Omega}_{\mathrm{res}} = \tilde{\Omega} - \hat{\Sigma}_{\mathrm{fit}_R} = \mathbb{X}_R^T\mathbb{X}_R/np_L - \hat{\Sigma}_{\mathrm{fit}_R}$, where $\tilde{\Omega}$ and $\tilde{M}$ are sample row and column covariance matrices. Then the MLEs can be obtained based on the following.

Proposition 2.5. *Suppose that $X_i \in \mathbb{R}^{p_L \times p_R}$, $i = 1, ..., n$ are observed and centered matrices, and let $(\hat{\Gamma}_1, \hat{\Gamma}_2, \hat{\beta}_1, \hat{\beta}_2, \hat{\Omega}, \hat{M})$ be the minimizers of (2.12).*

(i) For fixed Ω, Γ_1, and β_1, if $\hat{U}_L\hat{\Lambda}_L\hat{U}_L^T$ be the eigen-decomposition of $\hat{M}_{\mathrm{res}}^{-\frac{1}{2}}\hat{\Sigma}_{\mathrm{fit}_L}\hat{M}_{\mathrm{res}}^{-\frac{1}{2}}$ and $\hat{D}_L$ is the diagonal matrix with the first d_L eigenvalues of $\hat{\Lambda}_L$ replaced by zeros, then $\hat{M} = \hat{M}_{\mathrm{res}} + \hat{M}_{\mathrm{res}}^{\frac{1}{2}}\hat{U}_L\hat{D}_L\hat{U}_L^T\hat{M}_{\mathrm{res}}^{\frac{1}{2}}$, $\hat{\Gamma}_2 = \hat{M}^{\frac{1}{2}}$ times the first d_L eigenvectors of $\hat{M}^{-\frac{1}{2}}\hat{\Sigma}_{\mathrm{fit}_L}\hat{M}^{-\frac{1}{2}}$, and $\hat{\beta}_2 = \hat{\Gamma}_2^T\hat{M}^{-1}\mathbb{X}_L^T\mathbb{F}_L(\mathbb{F}_L^T\mathbb{F}_L)^{-1}$.

(ii) For fixed M, Γ_2, and β_2, if $\hat{U}_R\hat{\Lambda}_R\hat{U}_R^T$ is the eigen-decomposition of $\hat{\Omega}_{\mathrm{res}}^{-\frac{1}{2}}\hat{\Sigma}_{\mathrm{fit}_R}\hat{\Omega}_{\mathrm{res}}^{-\frac{1}{2}}$ and $\hat{D}_R$ is the diagonal matrix with the first d_R eigenvalues of $\hat{\Lambda}_R$ replaced by zeros, then $\hat{\Omega} = \hat{\Omega}_{\mathrm{res}} + \hat{\Omega}_{\mathrm{res}}^{\frac{1}{2}}\hat{U}_R\hat{D}_R\hat{U}_R^T\hat{\Omega}_{\mathrm{res}}^{\frac{1}{2}}$, $\hat{\Gamma}_1 = \hat{\Omega}^{\frac{1}{2}}$ times the first d_R eigenvectors of $\hat{\Omega}^{-\frac{1}{2}}\hat{\Sigma}_{\mathrm{fit}_R}\hat{\Omega}^{-\frac{1}{2}}$, and $\hat{\beta}_1 = \hat{\Gamma}_1^T\hat{\Omega}^{-1}\mathbb{X}_R^T\mathbb{F}_R(\mathbb{F}_R^T\mathbb{F}_R)^{-1}$.

To estimate the parameters in (2.12), one can begin with initial estimates of Ω, Γ_1, and β_1, then iterate the two steps in Proposition 2.5 until the log likelihood function (2.12) converges. The computational cost of dimension folding PFC under a general

error is in general less expensive than that of conventional PFC and dimension folding SIR. We summarize the results for all models in Table 2.1.

Table 2.1: Comparison of computation complexity

Method		Computation complexity
PCA	DF-PCA	$O(\max(p_L, p_R)^2 \max(d_L, d_R) nm)$
	PCA	$O(p_L^2 p_R^2 n)$
isotropic PFC	DF-PFC	$O(\max(nd_L, nd_R, r_L, r_R)\max(p_L, p_R, r_L, r_R)^2 m)$
	PFC	$O(\max(n, p_L p_R)\max(p_L p_R, r) r)$
general PFC	DF-PFC	$O(\max(np_L, np_R, r_L, r_R)\max(p_L, p_R, r_L, r_R)^2 m)$
	PFC	$O(\max(n, p_L p_R)\max(p_L p_R, r)^2)$
SIR	DF-SIR	$O(p_L^2 p_R^2 \max(p_L p_R, n) k)$

Remark 1. According to Proposition 2.5, $\hat{M}$ is invertible when $\hat{M}_{\text{res}}$ is invertible. The existence of $\hat{M}_{\text{res}}^{-1}$ only requires that $p_L \leq np_R - 1$ and $\text{Rank}(I - \text{P}_{\mathbb{F}_L}) = p_L$. The latter condition is generally satisfied since the nonzero eigenvalues of $\text{P}_{\mathbb{F}_L}$ are unlikely to be exactly equal to one and they are unlikely to be all identical. Hence it is usually guaranteed that $\hat{M}^{-1}$ and $\hat{\Omega}^{-1}$ exist if $p_L \leq np_R - 1$ and $p_R \leq np_L - 1$ or, equivalently, $n > \max(\frac{p_L}{p_R}, \frac{p_R}{p_L}) - 1$.

Remark 2. The maximum matrix dimension required in Proposition 2.5 is $np_L \times np_L$ or $np_R \times np_R$, from $\text{P}_{\mathbb{F}_L}$ or $\text{P}_{\mathbb{F}_R}$. This dimension could be very large ($> 30000 \times 30000$) in some cases (e.g. the EEG data) and exceed the storage limit in R software. In this case, one can apply an equivalent iteration algorithm that i) chooses moment estimators of Ω and M as initial values of $\hat{\Omega}$ and $\hat{M}$; ii) standardizes the predictors as $Z_i = \hat{M}^{-\frac{1}{2}} X_i \hat{\Omega}^{-\frac{1}{2}}$; iii) applies isotropic dimension folding PFC to the standardized data; iv) updates $\hat{\Omega}$ and $\hat{M}$ according to (2.26) and (2.27), the MLEs of matrix normal distribution (Dutilleul 1999) described in the supplement file; v) repeats ii)-iv) using the updated parameter values until the likelihood function converges.

Remark 3. Although the proposed algorithms are quite efficient for estimating the

central dimension folding subspace based on random initial values, using the conventional PFC model to obtain initial values can guarantee consistency of the estimators when the fitted function $f(Y)$ is misspecified. This is discussed in Sections 2.3 and 2.8.

Corollary 2.1 provides five equivalent forms of the MLE of the central dimension folding subspace. We applied the original form $\mathcal{S}_{d_R}(\hat{\Omega}, \hat{\Sigma}_{\text{fit}_R}) \otimes \mathcal{S}_{d_L}(\hat{M}, \hat{\Sigma}_{\text{fit}_L})$ in our simulation and data analysis.

Corollary 2.1. *The MLE of $\mathcal{S}_{Y|\circ X\circ}$ under (2.7) with an general error is* $\mathcal{S}_{d_R}(\hat{\Omega}, \hat{\Sigma}_{\text{fit}_R}) \otimes \mathcal{S}_{d_L}(\hat{M}, \hat{\Sigma}_{\text{fit}_L})$. *It is equivalent to* $\mathcal{S}_{d_R}(\hat{\Omega}_{\text{res}}, \hat{\Sigma}_{\text{fit}_R}) \otimes \mathcal{S}_{d_L}(\hat{M}_{\text{res}}, \hat{\Sigma}_{\text{fit}_L}) = \mathcal{S}_{d_R}(\tilde{\Omega}, \hat{\Sigma}_{\text{fit}_R}) \otimes \mathcal{S}_{d_L}(\tilde{M}, \hat{\Sigma}_{\text{fit}_L}) = \mathcal{S}_{d_R}(\hat{\Omega}, \tilde{\Omega}) \otimes \mathcal{S}_{d_L}(\hat{M}, \tilde{M}) = \mathcal{S}_{d_R}(\hat{\Omega}_{\text{res}}, \tilde{\Omega}) \otimes \mathcal{S}_{d_L}(\hat{M}_{\text{res}}, \tilde{M})$.

2.3 Robustness

In this section, we study the robustness of the estimator $\mathcal{S}_{d_R}(\hat{\Omega}, \hat{\Sigma}_{\text{fit}_R}) \otimes \mathcal{S}_{d_L}(\hat{M}, \hat{\Sigma}_{\text{fit}_L})$ when $f(Y)$ in model (2.7) is misspecified and the normality assumption is violated.

2.3.1 Misspecification of $f(Y)$

Under (2.7), we now assume that the true fitting function $f(Y)$ is misspecified by using the user-selected function $h(Y)$ in place of $f(Y)$. It can be shown that the estimator of the central dimension folding subspace is still consistent under certain conditions. To simplify the notation, let $g = \beta_2 f(Y)\beta_1^T$ and $l = \kappa_2 h(Y)\kappa_1^T$ be the misspecified fitting components. Note that g and l are both centered. We take $\rho_L = \text{var}_c^{-\frac{1}{2}}(g)\text{cov}_c(g,l)\text{var}_c^{-\frac{1}{2}}(l)$ to be the $d_L \times d_L$ column correlation matrix between the elements of g and l, where $\text{var}_c(g) = \text{E}(gg^T)$ is the column variance of g, $\text{var}_c(l) = \text{E}(ll^T)$ is the column variance of l, and $\text{cov}_c(g,l) = \text{E}(gl^T)$ is the column covariance matrix between g and l; let $\rho_R = \text{var}_r^{-\frac{1}{2}}(g)\text{cov}_r(g,l)\text{var}_r^{-\frac{1}{2}}(l)$ be the $d_R \times d_R$ row correlation matrix between the elements of g and l, where $\text{var}_r(g) = \text{E}(g^Tg)$ and $\text{var}_r(l) = \text{E}(l^Tl)$ are row variance matrices of g and l, and $\text{cov}_r(g,l) = \text{E}(g^Tl)$ is the row covariance matrix between g and l.

Proposition 2.6. *$\mathcal{S}_{d_R}(\hat{\Omega}, \hat{\Sigma}_{\text{fit}_R}) \otimes \mathcal{S}_{d_L}(\hat{M}, \hat{\Sigma}_{\text{fit}_L})$ is a $\sqrt{n}$ consistent estimator of* $\text{span}(\Omega^{-1}\Gamma_1) \otimes \text{span}(M^{-1}\Gamma_2)$ *if and only if ρ_L has rank d_L and ρ_R has rank d_R.*

Thus $\mathcal{S}_{d_R}(\hat{\Omega}, \hat{\Sigma}_{\text{fit}_R}) \otimes \mathcal{S}_{d_L}(\hat{M}, \hat{\Sigma}_{\text{fit}_L})$ can still be a reasonable estimator when $f(Y)$ is misspecified and the normality assumption is violated, as long as the row and columns correlations between the true fitting function and the selected fitting function have full ranks. This result is a generalization of Theorem 3.5 in Cook and Forzani (2008), and it is a mild condition. Nevertheless, in applications care should be taken when selecting $f(Y)$ in order to obtain better estimates. Polynomial approximations can be simple and good choices.

2.3.2 Normality assumption

In applications, when the matrix-valued predictors do not satisfy the normality assumption, transformations such as log power are commonly used in literature (Gasser, Bächer, and Möcks 1982) to achieve relative normality.

In addition, we show that the normality assumption is not essential for our model-based dimension folding methods. Suppose the random error ε in model (2.2) follows a general distribution with mean zero and covariance matrices I_{P_R} and I_{P_L}. The unknown parameters in this model can be estimated by minimizing $\sum_{i=1}^{n} ||(X_i - \mu - \Gamma_2 \nu_i \Gamma_1^T)||_F^2 = \sum_{i=1}^{n} \text{tr}[(X_i - \mu - \Gamma_2 \nu_i \Gamma_1^T)^T (X_i - \mu - \Gamma_2 \nu_i \Gamma_1^T)]$. Here the estimates $\text{span}(\hat{\Gamma}_1)$ and $\text{span}(\hat{\Gamma}_2)$ have the same expression as what we obtained under normality. Moreover, this objective function is equivalent to the loss function (2.6) of the two-mode tensor PCA. The asymptotic normality and asymptotic efficiency of the projection matrix $P_{\hat{\Gamma}_1 \otimes \hat{\Gamma}_2}$ onto the estimated principal subspace $\text{span}(\hat{\Gamma}_1 \otimes \hat{\Gamma}_2)$ were developed by Hung et al. (2012). Hence without normality, one can still obtain $\sqrt{n}$ consistent estimators for the principal subspaces.

In terms of sufficient dimension reduction, the normality assumption can be relaxed to the elliptically symmetric condition required by dimension folding SIR. Suppose $\text{vec}(\varepsilon) \sim \text{EC}_{p_L p_R}(0, \Omega \otimes M, Q)$, where $\text{EC}_{p_L p_R}(0, \Omega \otimes M, Q)$ is an elliptical contoured distribution with mean zero, row and column covariance matrices Ω and M, and a density generator $Q(.)$. Let $\tilde{Y} = sI(Y \in J_s), s = 1, ..., h$, be the slice indicator function, where $J_1, ..., J_h$ are h non-overlapping slices. Let $\tilde{\zeta} = (\Omega \otimes M)^{-1}\text{E}[\text{vec}(X)|\tilde{Y}]$, and let $\mathcal{E}^{\otimes}(\tilde{\zeta})$ be the Kronecker envelope of $\tilde{\zeta}$. According Li et al. (2010), $\mathcal{E}^{\otimes}(\tilde{\zeta})$ is the dimension folding SIR subspace. It is defined as $\mathcal{S}_{\circ\tilde{\zeta}} \otimes \mathcal{S}_{\tilde{\zeta}\circ}$, the Kronecker product of the two smallest subspaces $\mathcal{S}_{\circ\tilde{\zeta}}$ and $\mathcal{S}_{\tilde{\zeta}\circ}$, such that $\text{span}(\tilde{\zeta}) \subseteq \mathcal{S}_{\circ\tilde{\zeta}} \otimes \mathcal{S}_{\tilde{\zeta}\circ}$. The relationships between the dimension folding SIR subspace ($\mathcal{S}_{fSIR}$), dimension folding PFC subspace ($\mathcal{S}_{fPFC}$), and central dimension folding subspace ($\mathcal{S}_{Y|\circ X\circ}$) are shown below.

Proposition 2.7. *Under (2.7), when the random error is elliptically contoured distributed as* $\text{EC}_{p_L p_R}(0, \Omega \otimes M, Q)$, $\mathcal{S}_{fSIR} \subseteq \mathcal{S}_{fPFC} \subseteq \mathcal{S}_{Y|\circ X\circ}$, *where* $\mathcal{S}_{fPFC} = \text{span}(\Omega^{-1}\Gamma_1) \otimes \text{span}(M^{-1}\Gamma_2)$.

Thus, under the elliptically symmetric condition, the subspace $\text{span}(\Omega^{-1}\Gamma_1) \otimes \text{span}(M^{-1}\Gamma_2)$ given by dimension folding PFC is not guaranteed to be the true central dimension folding subspace but a subspace of it. It contains the dimension folding SIR subspace at the population level and its sample estimate can be more accurate since the fitting function $f(Y)$ is generally more efficient than a slicing function. Therefore, under this minimum condition, dimension folding PFC is still useful. Both algorithms in Section 2.2.2 provide $\sqrt{n}$ consistent estimators for $\mathcal{S}_{fPFC}$ without normality, because the algorithm for the isotropic error case in Section 2.2.2 is equivalent to a least square estimation and the consistent estimation of the algorithm for the general error case in Section 2.2.2 is given by Proposition 2.6, which does not rely on normality.

Similarly, Proposition 2.7 holds for dimension folding PCA in terms of $\mathcal{S}_{\nu|\circ X\circ}$ and $\zeta = \text{E}[\text{vec}(X)|\nu]$. Hence dimension folding PCA and PFC are beneficial under the minimum elliptically symmetric condition.

2.4 Prediction

The ultimate purpose of dimension folding is to serve regression and classification. Dimension folding SIR, SAVE, and DR proposed by Li et al. (2010) provide good prediction results in the classification case. Dimension folding PFC can further improve prediction accuracy for classification problems. In the regression case, where the response variable is continuous, the function of moment-based dimension folding methods is limited. Slicing could miss useful information on the response variable and the choice of slice number is a big issue. Dimension folding PFC can overcome this shortcoming and provide better prediction results.

We propose two prediction approaches. Based on our knowledge, there is no well-established method for predicting a univariate responses from a matrix-valued predictor directly. Thus, we consider the prediction of Y from $\text{vec}(X)$ instead. The first approach is to regress Y on $\text{vec}(X)$ in two steps. By applying dimension folding PCA or PFC, one can obtain the MLE of the central dimension folding subspace $\hat{\mathcal{S}}_{Y|\circ X\circ} = \text{span}(\hat{\Gamma}_1) \otimes \text{span}(\hat{\Gamma}_2)$ under an isotropic error, or $\hat{\mathcal{S}}_{Y|\circ X\circ} = \text{span}(\hat{\Omega}^{-1}\hat{\Gamma}_1) \otimes \text{span}(\hat{M}^{-1}\hat{\Gamma}_2) = \mathcal{S}_{d_R}(\hat{\Omega}, \hat{\Sigma}_{\text{fit}_R}) \otimes \mathcal{S}_{d_L}(\hat{M}, \hat{\Sigma}_{\text{fit}_L})$ under a general error. After dimension folding, one has a new predictor $\hat{\Gamma}_2^T X \hat{\Gamma}_1$, or $\hat{\Gamma}_2^T \hat{M}^{-1} X \hat{\Omega}^{-1} \hat{\Gamma}_1$, with smaller row and column dimensions compared to the original predictor X. The second step is to fit a model, such as a general additive model (GAM), to estimate the mean function $E[Y|\text{vec}(\hat{\Gamma}_2^T X \hat{\Gamma}_1)]$ or $E[Y|\text{vec}(\hat{\Gamma}_2^T \hat{M}^{-1} X \hat{\Omega}^{-1} \hat{\Gamma}_1)]$, and then perform prediction based on it.

The second method was motivated by a nonparametric prediction technique of Adragni and Cook (2009). Let $f(X)$ and $f(X|Y)$ be the density functions of X and $X|Y$. Let $R(X)$ denote a sufficient folding assumed to have a density. Then $\text{E}[Y|X = x] = \text{E}\{Y f[R(x)|Y]\}/\text{E}\{f[R(x)|Y]\}$. This provides the key idea of this nonparametric prediction approach because the estimated prediction function $\hat{\text{E}}[Y|X = x]$ can be written as $\hat{\text{E}}[Y|X = x] = \sum_{i=1}^n \omega_i(x) Y_i$, where $\omega_i(x) = \hat{f}[\hat{R}(x)|Y_i] / \sum_{i=1}^n \hat{f}[\hat{R}(x)|Y_i]$.

Once the density function $f(X|Y)$ is estimated, the predicted value $\hat{Y}$ can be easily obtained since it is the weighted average of the observed responses. This method is applicable to our proposed dimension folding models since the conditional distribution of $X|Y$ is known through the model assumptions. According to (2.23) in Section 2.8, when the random error ε is isotropic we have

$$\begin{aligned}\hat{f}[\hat{R}(x)|Y_i] &= \hat{f}[\hat{R}(\mathrm{vec}(x))|Y_i] \\ &\propto \exp\{-(2\hat{\sigma}^2)^{-1}||(\hat{\Gamma}_1 \otimes \hat{\Gamma}_2)^T[\mathrm{vec}(x) - \mathrm{vec}(\hat{X}_i)]||^2\} \\ &= \exp\{-(2\hat{\sigma}^2)^{-1}||\hat{R}(\mathrm{vec}(x)) - \hat{R}(\mathrm{vec}(\hat{X}_i))||^2\},\end{aligned} \tag{2.13}$$

where $\mathrm{vec}(\hat{X}_i) = \mathrm{vec}(\bar{X}) + (\hat{\Gamma}_1 \otimes \hat{\Gamma}_2)(\hat{\beta}_1 \otimes \hat{\beta}_2)\mathrm{vec}(f(Y_i))$ is the predicted value of $\mathrm{vec}(x)|Y_i$ and the reduction $\hat{R}(\mathrm{vec}(x)) = (\hat{\Gamma}_1 \otimes \hat{\Gamma}_2)^T\mathrm{vec}(x)$. When ε has a general covariance structure, the estimated conditional density is

$$\begin{aligned}\hat{f}[\hat{R}(x)|Y_i] = \hat{f}[\hat{R}(\mathrm{vec}(x))|Y_i] \propto \exp\{-\frac{1}{2}||[(\hat{\Gamma}_1 \otimes \hat{\Gamma}_2)^T(\hat{\Omega} \otimes \hat{M})^{-1}(\hat{\Gamma}_1 \otimes \hat{\Gamma}_2)]^{-\frac{1}{2}} \\ [\hat{R}(\mathrm{vec}(x)) - \hat{R}(\mathrm{vec}(\hat{X}_i))]||^2\},\end{aligned} \tag{2.14}$$

where $\hat{R}(\mathrm{vec}(x)) = (\hat{\Gamma}_1 \otimes \hat{\Gamma}_2)^T(\hat{\Omega} \otimes \hat{M})^{-1}\mathrm{vec}(x)$.

Each method outperforms the other under certain conditions. The inverse regression prediction relies on the density function $f[R(X)|Y]$ but does not make any parametric assumption on modeling $Y|X$, while forward regression prediction usually assumes a parametric model on $Y|X$ or it depends on the estimation of $Y|X$. Thus, the inverse prediction method shows its advantages when the distribution of the random error ε in model (2.7) is known or can be well estimated. The forward prediction is beneficial when the assumption made on $Y|X$ is reasonable.

In addition, the choice of $f(Y)$ can affect the prediction accuracy. Consider the mean squared error $\mathrm{MSE} = \mathrm{E}[Y - \hat{Y}(X)]^2$ for which the minimum prediction error is achieved when $\hat{Y}(X)$ is the conditional mean $\mathrm{E}(Y|X)$. According to Proposition 2.6, when the row and column correlations of the selected fitting function $\kappa_2 h(Y)\kappa_1$ and the true function both have full ranks, which indicates that the two are correlated, the estimator

of the central dimension folding subspace is $\sqrt{n}$ consistent. For the forward prediction method, we have $\hat{Y}(X) = \hat{\mathrm{E}}(Y|\hat{R}(X)) = \hat{\mathrm{E}}(Y|\hat{\Gamma}_2^T\hat{M}^{-1}X\hat{\Omega}^{-1}\hat{\Gamma}_1)$. If one chooses $\hat{\mathrm{E}}(Y|R(X))$ to be a consistent estimator for $\mathrm{E}(Y|R(X))$, such as the Nadaraya-Watson estimator, then under mild regularity conditions, $\hat{Y}(X) \to \mathrm{E}(Y|R(X)) = \mathrm{E}(Y|X)$ when the selected fitting function is correlated to the true function. Thus the prediction error can reach its minimum asymptotically if the condition in Proposition 2.6 is satisfied. For the inverse prediction method, we have

$$\hat{\mathrm{E}}[Y|X=x] = \frac{1}{n}\sum_{i=1}^{n}\hat{f}[\hat{R}(x)|Y_i]Y_i \Big/ \frac{1}{n}\sum_{i=1}^{n}\hat{f}[\hat{R}(x)|Y_i].$$

Assuming that $f(Y)$ is known, then it can be shown that $\frac{1}{n}\sum_{i=1}^{n}\hat{f}[\hat{R}(x)|Y_i] \to \mathrm{E}\{f[R(X)|Y]\}$ and $\frac{1}{n}\sum_{i=1}^{n}\hat{f}[\hat{R}(x)|Y_i]Y_i \to \mathrm{E}\{Yf[R(X)|Y]\}$ at $\sqrt{n}$ rate. Then $\hat{\mathrm{E}}[Y|X=x]$ converges to $\mathrm{E}[Y|X=x]$ and the prediction error is asymptotically minimized. This result does not hold for misspecified $f(Y)$ because the density function $f[R(X)|Y]$ is misspecified in this case. Yet we can expect that the closer the approximation of the fitting function, the more likely we obtain good prediction.

2.5 Simulation studies

2.5.1 Evaluation of estimation accuracy

We assess the accuracy of our proposed dimension folding methods and compare it to that of conventional methods. We measure the difference between the estimated projection matrices and true projection matrices for the central dimension folding subspace and denote it as "PCDF_Error"; for conventional PCA and PFC, we evaluate the estimation error of the projection matrices of the central subspace and denote it as "PCS_Error". Specifically,

$$\text{PCDF\_Error} = \|P_{\hat{\mathcal{S}}_{Y|\circ X\circ}} - P_{\mathcal{S}_{Y|\circ X\circ}}\|_F^2 \tag{2.15}$$

$$\text{PCS\_Error} = \|P_{\hat{\mathcal{S}}_{Y|\text{vec}(X)}} - P_{\mathcal{S}_{Y|\circ X\circ}}\|_F^2. \tag{2.16}$$

To evaluate the performance of the dimension folding PCA model (2.2), the data were generated as follows: Let $d_L = d_R = 2$ and $p_L = p_R = p$, with sample size $n = 100$. The components of Γ_1 and Γ_2 were generated from $N(0, 1)$ and the components ν_i before centering were generated from $N(1, 2)$, $i = 1, ..., n$. The vectorized isotropic error ε was obtained from the multivariate normal with mean zero and covariance matrix $0.8I_{p_L p_R}$. We chose $p = 5, 10, 15, 20$ and 30, and ran each simulation 1000 times. The notations "DF-PCA", "DF-PFC" and "DF-SIR" were used to denote dimension folding PCA, dimension folding PFC, and dimension folding SIR in figures and tables.

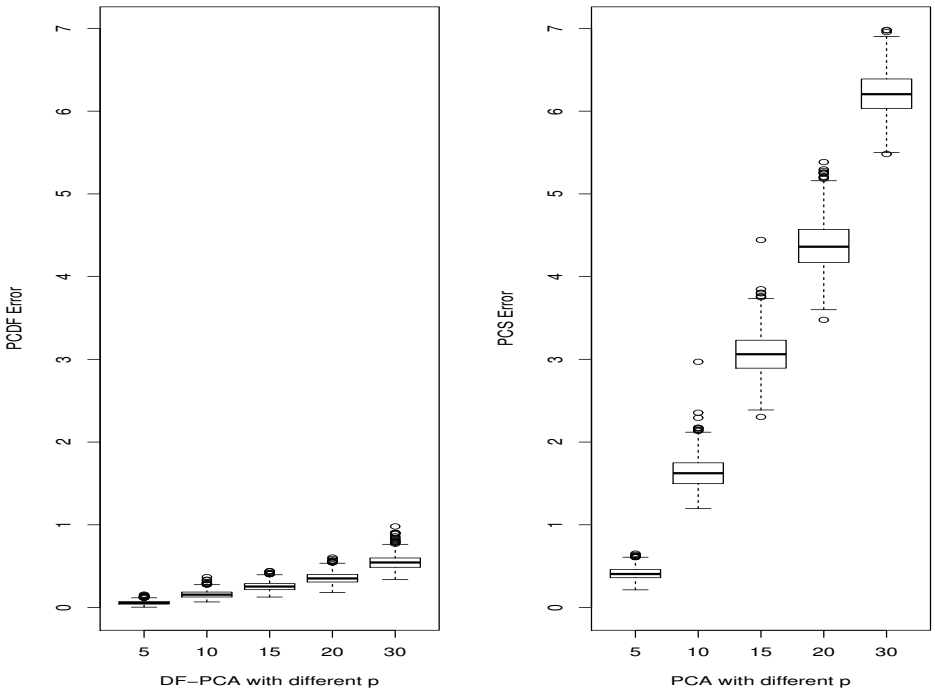


Figure 2.1: The comparison results of DF-PCA and PCA

Figure 2.1 shows that for all selected dimensions of p, dimension folding PCA was

noticeably more accurate than PCA. As the predictor's dimension increases both methods showed ascending estimation distance from the true projection space, but dimension folding PCA had the slower error increase in both the mean and standard deviation.

For dimension folding PFC, we did simulations for both isotropic and general error cases. When the general error structure was considered, we chose $p_L = p_R = 3$, $d_L = d_R = 2$ and $r_L = r_R = 4$. Conventional PFC and dimension folding SIR both required $n > p_L \times p_R$ with a general error and we used small matrices $p_L \times p_R = 9$ in this case. The sample size was selected as $n = 30, 50, 80, 100$ and 150. The components of Γ_1 and Γ_2 were generated from $N(0, 1)$. The elements of β_1 and β_2 were generated from $N(1, 2)$ and absolute normal $|N(2, 2)|$. The responses Y_i, $i = 1, ..., n$ were obtained from $N(0, 1)$, and $f(Y_i) = \text{diag}(Y_i, Y_i^2, Y_i^3, Y_i^4)$. The covariance matrices were

$$\Omega = \begin{pmatrix} 0.50 & -0.25 & 0.00 \\ -0.25 & 0.50 & -0.25 \\ 0.00 & -0.25 & 0.50 \end{pmatrix} \qquad M = \begin{pmatrix} 0.886 & 0.266 & 0.062 \\ 0.266 & 0.248 & 0.048 \\ 0.062 & 0.048 & 0.015. \end{pmatrix}$$

For the isotropic error case, we chose $p_L = p_R = 10$ and $\sigma = 0.8$, with sample size $n = 120, 150, 200, 300, 500$. The other parameters were kept the same as those in the general error case. We ran the simulation 1000 times for each sample size.

Figure 2.2 summarizes the results under the general error setting. It can be seen that the central dimension folding subspaces were estimated precisely based on the estimation procedures proposed in Section 2.2.2 except for some extreme outliers. Although the plots appear with dense outliers, the actual percentages of these outliers were less than 5% under 1000 repetitions. Some outliers like the one with estimation error close to 3 at $n = 150$ could be due to the algorithm getting caught in a local minimum. Conventional PFC had much higher estimation errors for all sample sizes.

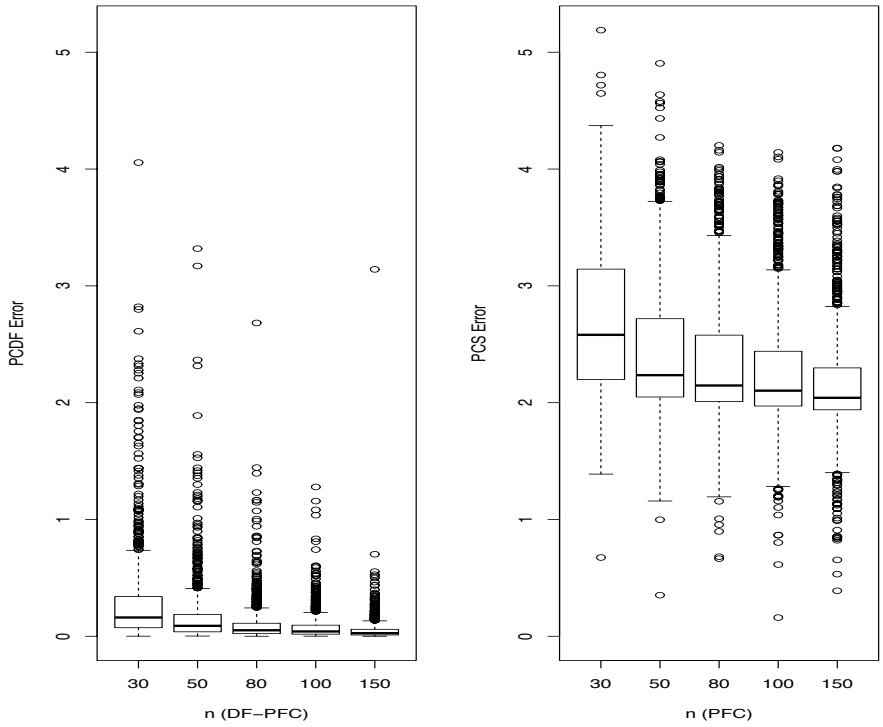
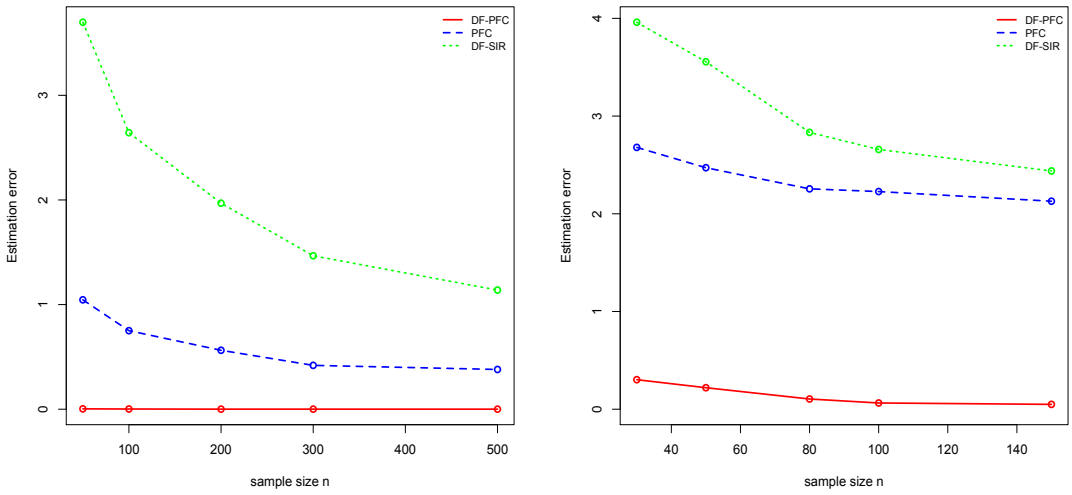


Figure 2.2: The comparison results of DF-PFC and PFC under general errors

We further compared the model-based methods to dimension folding SIR. For the latter, 8 slices were selected for the response variable. Based on our simulation results, it was the best choice among 6, 8, 10 and 15 slices.

The mean estimation errors are shown in Figure 2.3, based on 1000 repetitions. It can be seen that dimension folding PFC provided the most accurate estimations for the central dimension folding subspace over all sample sizes. Although conventional PFC was less accurate than dimension folding PFC, it still beat dimension folding SIR to a large extent. Dimension folding SIR failed to obtain precise estimation because the conditional mean $\mathrm{E}(X|Y)$ was not adequately estimated by slicing the responses. The PFC methods benefitted from careful fitting of the inverse regression of X on Y.



(a) isotropic error

(b) general error

Figure 2.3: The comparison results of DF-PFC, DF-SIR and PFC

2.5.2 Choice of d_L and d_R

In the previous sections, the reduced row and column dimensions d_L and d_R were assumed known. In applications, one can apply an information criterion, say AIC or BIC, to select optimal dimensions by minimizing the objective function $-2L(d_L, d_R) + h(n)g(d_L, d_R)$. Here $L(d_L, d_R)$ is the log likelihood function of the estimated model, $h(n)$ is $\log(n)$ for BIC and 2 for AIC, and $g(d_L, d_R)$ is the number of parameters to be estimated. One can also use the likelihood ratio test statistic $\Lambda(d_{L_0}, d_{R_0}) = 2(L(\min(r_L, p_L), \min(r_R, p_R)) - L(d_{L_0}, d_{R_0}))$ to perform sequential tests for increasing values of d_L and d_R.

We illustrate these procedures using the simulated samples obtained from the isotropic error setting. Here $d_L = d_R = 2$, $p_L = p_R = 10$, and $n = 200$ were chosen for both dimension folding PCA and dimension folding PFC. The simulations were repeated 1000 times. All three methods were able to correctly identify the true dimensions over 95%

of the time. When we took the true dimensions to be $(d_L, d_R) = (1,1), (1,2)$ and $(2,1)$, the percentages of the precise identifications were over 90% for all methods.

Table 2.2: Percentages of correct identifications

	DF-PCA			DF-PFC		
	AIC	BIC	LRT(p-val.)	AIC	BIC	LRT(p-val.)
$d_L = d_R = 1$	100	100	100	94.8	100	90.6
$d_L = 1, d_R = 2$	100	100	100	98.5	99.6	92.0
$d_L = 2, d_R = 1$	100	100	100	98.1	99.6	93.2
$d_L = d_R = 2$	100	100	100	99.9	99.8	95.8

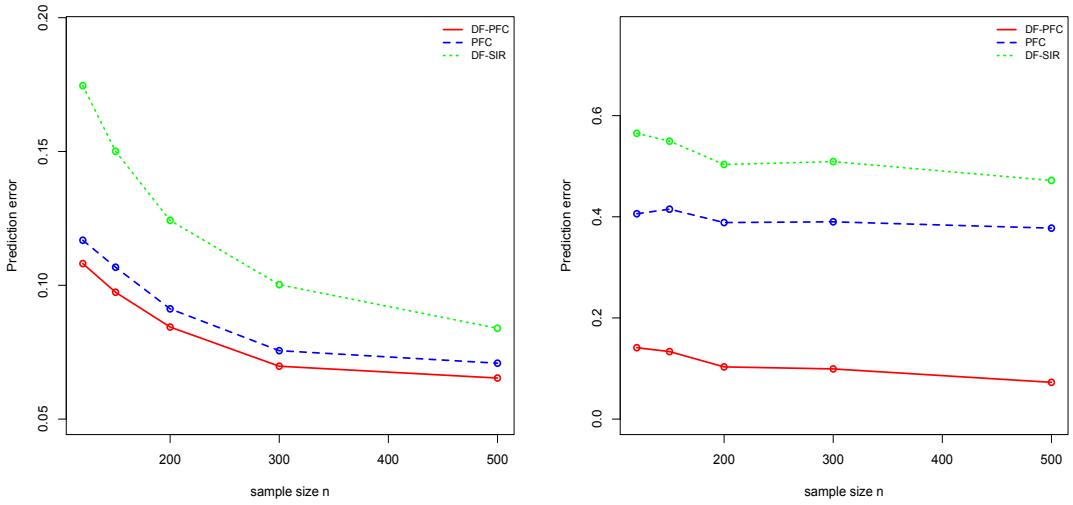
2.5.3 Prediction

We evaluated the prediction performance of dimension folding PFC, conventional PFC, and dimension folding SIR using the simulated data under the isotropic error from Section 2.5.1. The two prediction methods in Section 2.4 were applied. For the first method, we fitted a generalized additive model of Y on the reduced predictor $(\hat{\Gamma}_1 \otimes \hat{\Gamma}_2)^T \text{vec}(X)$ to the original data and then generated new data for prediction. The new data are denoted by (X_i^*, Y_i^*), $i = 1, ..., n_{\text{new}}$, where $n_{\text{new}} = n/4$. The average prediction error was calculated as:

$$\text{PE} = \sum_{i=1}^{n_{\text{new}}} (Y_i^* - \hat{\text{E}}(Y|X = X_i^*))^2 / n_{\text{new}}. \tag{2.17}$$

This procedure was repeated for 1000 data sets and the averaged prediction error $\sum_{i=1}^{1000} \text{PE}_i/1000$ was used to assess the prediction accuracy of the three methods.

For the nonparametric prediction approach, we used the same data and evaluation scenario except for using different prediction functions for $\hat{\text{E}}(Y|X = X_i^*)$. For dimension folding PFC and conventional PFC, the density function $f(X|Y)$ was obtained based on their model assumptions. For dimension folding SIR, $f(X|Y)$ was estimated based on the matrix normal distribution.



(a) prediction with model fitting

(b) nonparametric prediction

Figure 2.4: Prediction results under isotropic errors

Figure 2.4(a) shows the prediction results with generalized additive model fitting. It illustrates the potential advantages of using an inverse regression model to estimate the conditional expectation $\mathrm{E}(X|Y)$, or $\mathrm{E}[\mathrm{vec}(X)|Y]$, instead of using a slicing method. Dimension folding PFC predicted best over all sample sizes. Though conventional PFC omits the predictor's matrix structure, it still gave more accurate results than did dimension folding SIR. Figure 2.4(b) shows the prediction performance according to the second prediction approach. It provided smaller prediction errors for dimension folding PFC and relatively large errors for conventional PFC and dimension folding SIR.

2.6 Data analysis

We applied dimension folding PFC to two data sets, one with a discrete response, the other with a continuous response. For the discrete response case, the EEG data used in Li et al. (2010) was studied, while Dow Jones industrial stock data was used for the

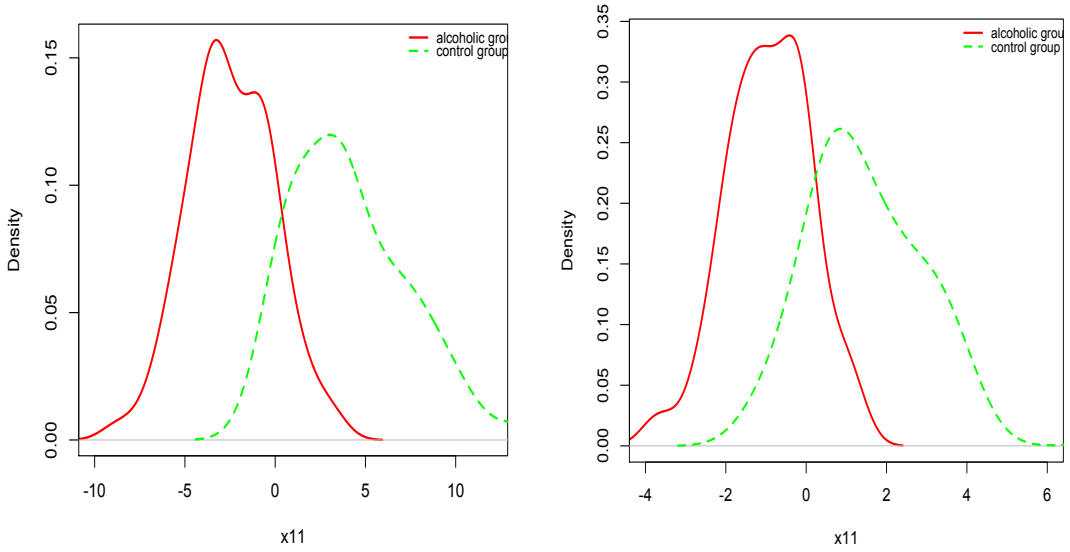
second case.

2.6.1 EEG data

The primary goal of this study was to explore the relationship between alcoholism and the pattern of voltage values over times and channels. Let $(X_1, Y_1), ..., (X_{122}, Y_{122})$ denote the observed data, where X_i is a 256×64 matrix and Y_i is a binary univariate variable, $i = 1, ..., 122$. It is easy to see that error structure is not isotropic. In this case, conventional PFC is not applicable since $n \ll p_L \times p_R$. We applied dimension folding PFC with a general error to these data. Since our proposed estimation procedures circumvent vectorization of the predictors, we were able to handle the original EEG data without pre-screening work, as in Li et al. (2010). In our case, the maximum dimension of a matrix inversion is 256 by 256 ($\hat{M}^{-1}$), instead of the 256×64 by 256×64 ($\hat{\Sigma}^{-1}$) required for the moment-based dimension folding methods. According to Remark 1 in Section 2.2.2, both inverse matrices $\hat{M}^{-1}$ and $\hat{\Omega}^{-1}$ exist for the original EEG data, because $n > \max(\frac{p_L}{p_R}, \frac{p_R}{p_L}) - 1$.

For a categorial response Y of h categories, $f(Y)$ can be naturally chosen as a diagonal matrix with its kth diagonal element $\text{diag}(f(Y))_k = I(Y \in H_k) - n_k/n$, $k = 1, ..., h-1$. Thus, for the EEG data, we have $d_L = d_R = r_L = r_R = r = 1$. Then the sufficient reduction $\hat{\Gamma}_2^T \hat{M}^{-1} X \hat{\Omega}^{-1} \hat{\Gamma}_1$ obtained by the dimension folding PFC model is a univariate variable, labeled as X_{11}. Figure 2.5(a) shows good separation of the two groups by X_{11} without pre-screening the original predictors. Figure 2.5(b) shows the corresponding result after pre-screening the predictors to smaller dimensions $(p_L^*, p_R^*) = (15, 15)$ with the screening method in Li et al. (2010). Pre-screening the predictors loses information about the original data as the two groups cannot be separated quite as well as in (a).

To obtain classification results, we applied quadratic discriminant analysis and leave-one-out cross validation. Without pre-screening the original predictors, dimension folding PFC with a general error correctly classified 107 subjects out of the total 122 subjects



(a) DF-PFC without screen

(b) DF-PFC with screen

Figure 2.5: Density plot with the new reduced predictor X_{11}

based on X_{11}; after pre-screening the predictors, it classified 102 out of the 122 subjects. In comparison, dimension folding DR and dimension folding SIR provided 97 and 94 out of 122 correct decisions, using $(p_L^*, p_R^*) = (15, 15)$ and $(d_L, d_R) = (1, 2)$.

2.6.2 Dow Jones stock data

We used Dow Jones industrial stock data from January 2001 to December 2010. The response is the monthly Dow Jones industrial average index change rate. If m_i denotes the Dow Jones industrial average monthly index for the i-th month, the responses $Y_i = (m_i - m_{i-1})/m_{i-1}$, $i = 1, ..., n$, are the index change rates, assumed to be independent. For each response (month), the predictor was formed by 19 daily stock price change rates over the 30 Dow Jones companies. We chose 19 daily stock price change rates

because there are usually 19-23 trading days each month. Hence the predictor for each observation is a 19×30 matrix and the response is a univariate continuous variable. We deleted the observations in September 2001 and September 2008 due to the incidents of terrorism and the financial crisis, leaving $n = 118$ observation months. The final data set consisted of $(X_1, Y_1), ..., (X_{118}, Y_{118})$ observations. Primary interest was in association between monthly stock index change rates and the daily stock price change rates from the individual companies.

Dimension folding PFC with both isotropic and general errors, dimension folding SIR, and the Lasso were applied to our study. We evaluated the prediction performance for the first three methods using the prediction approach with OLS fitting of Y on the reduced predictor $\text{vec}(\hat{\Gamma}_2^T X \hat{\Gamma}_1)$, as proposed in Section 2.4. Four sets of dimensions, $(d_L, d_R) = (1, 1)$, $(d_L, d_R) = (1, 2)$, $(d_L, d_R) = (2, 1)$, and $(d_L, d_R) = (2, 2)$, were selected. The function $f(Y)$ was chosen as a diagonal matrix with its diagonal elements formed by (Y, Y^2, Y^3, Y^4) for dimension folding PFC. Dimension folding SIR was studied with slicing numbers 6 and 8. We also applied the Lasso to select important signals in $\text{vec}(X)$ and performed prediction. The 10-fold cross validation method was used to evaluate the prediction performance using (2.17) for all methods. The results are summarized in Table 2.3.

Table 2.3: Prediction results (×1000) with 10 folded cross validations

	DF-PFC		DF-SIR		Lasso
	(isotropic)	(general)	(6 slices)	(8 slices)	
$d_L = d_R = 1$	9.1	12.3	15.6	13.6	15.4
$d_L = 1, d_R = 2$	8.7	12.4	10.7	9.8	15.4
$d_L = 2, d_R = 1$	9.6	11.0	12.3	11.0	15.4
$d_L = d_R = 2$	10.0	10.1	12.8	11.0	15.4

It can be seen that isotropic dimension folding PFC provided smaller prediction errors than all other methods. Since the dependence of the stock price change rates is

not strong from day to day and from company to company, dimension folding PFC under a general error structure could be overparametrized and thus the prediction errors were likely to be increased. Dimension folding SIR presented less accurate results than the isotropic dimension folding PFC model over all selected dimensions and slicing numbers. Lasso showed relatively large prediction errors.

2.7 Discussion

Our dimension folding PCA and PFC methods provide likelihood-based dimension folding solutions for matrix-valued predictors that can be applied to a broad range of applications with categorical or continuous responses. The fitting components $f(Y)$ in the dimension folding models possess the flexibility to capture the useful information on response and provide more accurate estimation for the conditional mean $\mathrm{E}(X|Y)$ than the moment-based dimension folding approaches. The assumption on the covariance structure of the random error provides another benefit for the model-based methods since one can circumvent inverting the high dimensional covariance matrix of $\mathrm{vec}(X)$. In addition, the MLEs obtained from our algorithms have good interpretations and connections to the conventional PCA and PFC methods, and are robust to model assumptions.

There are different formulations for the dimension folding PFC model. Model (2.7) provides a multiplicative coefficient structure $\beta_2 f(Y)\beta_1^T$ for the fitted function. Instead, one can model dimension folding PFC with an additive coefficient structure, an interactive coefficient structure, or a general coefficient structure, respectively, as

$$X = \mu + \Gamma_2[\beta_2 f(Y)\mathbf{e}_{r_R,d_R} + \mathbf{e}_{d_L,r_L} f(Y)\beta_1^T]\Gamma_1^T + \varepsilon, \tag{2.18}$$

$$X = \mu + \Gamma_2[\beta_2 f(Y)\mathbf{e}_{r_R,d_R} + \mathbf{e}_{d_L,r_L} f(Y)\beta_1^T + \beta_2 f(Y)\beta_1^T]\Gamma_1^T + \varepsilon, \tag{2.19}$$

$$X = \mu + \Gamma_2 \mathrm{vec}^{-1}\{\beta g(Y)\}\Gamma_1^T + \varepsilon, \tag{2.20}$$

where $\mathbf{e}_{r_R,d_R}$ is a $r_R \times d_R$ matrix with all elements equal to one, and $\mathbf{e}_{d_L,r_L}$ is similarly defined. If $f(Y)$ is diagonal and its diagonal elements are formed by polynomial basis functions, then under (2.18) the folded conditional mean $[\Gamma_2^T \mathrm{E}(X|Y)\Gamma_1]_{ij} = \sum_{k=1}^{r}(\beta_{ik}^{(2)} + \beta_{kj}^{(1)})Y^k$, where the coefficients are additive. When the multiplicative or additive coefficient model itself is not sufficient to formulate the relationship between X and Y, (2.19) might be needed. In this case, $[\Gamma_2^T \mathrm{E}(X|Y)\Gamma_1]_{ij} = \sum_{k=1}^{r}(\beta_{ik}^{(2)} + \beta_{kj}^{(1)} + \beta_{ik}^{(2)}\beta_{kj}^{(1)})Y^k$. This is called the dimension folding PFC model with the interactive coefficient structure. More generally, one might not impose any constraints on the coefficients and adopt (2.20), where "vec^{-1}" stands for the matrixing operation. Then with polynomial basis functions as the components of $g(Y)$, the folded conditional mean $[\Gamma_2^T \mathrm{E}(X|Y)\Gamma_1]_{ij} = \sum_{k=1}^{r_L r_R} \beta_{(j-1)d_L+i,k}Y^k$, where $\beta_{(j-1)d_L+i,k}$ is the element in $[(j-1)d_L + i]$-th row and k-th column of β. The choice of a particular dimension folding PFC model depends on the intrinsic row and column structure of $X|Y$. To estimate model (2.20), one can apply the estimation procedure in Section 2.2.2, though the algorithm cannot be directly used for the dimension folding PFC model with the additive or the interactive coefficient structure. Instead, one can use numerical algorithms with least square iterations.

The proposed dimension folding models can also be generalized to array-valued predictors. Let $\mathcal{X} = \{\mathcal{X}_{i_1 \ldots i_m} : i_1 = 1, ..., p_1, ..., i_m = 1, ..., p_m\}$ be an m-way random array of dimension $p_1 \times \cdots \times p_m$ and Y be a univariate random response. The goal of dimension folding is to find a smaller m-array of dimension $d_1 \times \cdots \times d_m$ such that the conditional distribution of $Y|\mathcal{X}$ is retained the same if $\mathcal{X}$ is replaced by the reduced array. In this case, the dimension folding PCA and PFC models are formulated as

$$\mathrm{vec}\ (\mathcal{X}) = \mathrm{vec}(\mu) + (\Gamma_1 \otimes \Gamma_2 \otimes \cdots \otimes \Gamma_m)\mathrm{vec}\ (\nu_i) + \mathrm{vec}\ (\varepsilon), \tag{2.21}$$

$$\mathrm{vec}(\mathcal{X}) = \mathrm{vec}(\mu) + (\Gamma_1 \otimes \Gamma_2 \otimes \cdots \otimes \Gamma_m) \cdot (\beta_1 \otimes \beta_2 \otimes \cdots \otimes \beta_m) \cdot \mathrm{vec}(f(Y)) + \mathrm{vec}(\varepsilon), \tag{2.22}$$

respectively. Here $\Gamma_i \in \mathbb{R}^{p_{m-i} \times d_{m-i}}$, $i = 1, ..., m$, ν_i is a m-way array of dimension $d_1 \times \cdots \times d_m$, $\beta_i \in \mathbb{R}^{d_{m-i} \times r_{m-i}}$, $i = 1, ..., m$, $f(Y)$ is a m-way array of dimension

$r_1 \times \cdots \times r_m$, and vec (ε) has a multivariate normal distribution with mean $0_{p_1 \cdots p_m \times p_1 \cdots p_m}$ and covariance matrices $\Omega_1 \otimes \Omega_1 \otimes \cdots \otimes \Omega_m$. It can be shown that the dimension folding subspace with m-way array-valued predictors is span$\{(\Omega_1 \otimes \Omega_1 \otimes \cdots \otimes \Omega_m)^{-1}(\Gamma_1 \otimes \Gamma_2 \otimes \cdots \otimes \Gamma_m)\}$, which can be estimated by adapting the numerical algorithms in Section 2.1.2 and Section 2.2.2.

2.8 Appendix

2.8.1 Matrix normal distribution

A matrix-valued distribution (De Waal 1985) is a probability distribution of a random matrix. The matrix normal distribution is a generalization of the multivariate normal distribution to matrix-valued random variables. Let $X = \{X_{ij} : i = 1, ..., p_L, j = 1, , p_R\}$ be a matrix-valued variable. Its expected value and covariance matrix are defined as $\mathrm{E}[X] = (\mathrm{E}[X_{ij}]) = \mu$ and $\mathrm{var}(X) = \mathrm{E}[\mathrm{vec}(X - \mathrm{E}[X])\mathrm{vec}^T(X - \mathrm{E}[X])] = \Sigma$. Then X has a matrix normal distribution if its covariance can be decomposed as the Kronecker product of two positive definite matrices Ω and M, and vec(X) follows a multivariate normal distribution with mean vec(μ) and covariance matrix $\Sigma = \Omega \otimes M$. The matrix normal distribution is denoted as $\mathrm{N}_{p_L \times p_R}(\mu, \Omega, M)$. Its density function is defined through the distribution of vec(X) and is given by

$$\begin{aligned} f_X(x) &= f_{\mathrm{vec}(X)}(\mathrm{vec}(x)) \\ &= (2\pi)^{-\frac{p_L p_R}{2}} |\Omega|^{-\frac{p_L}{2}} |M|^{-\frac{p_R}{2}} \exp\{-\frac{1}{2}\mathrm{tr}(\Omega^{-1}(x-\mu)^T M^{-1}(x-\mu))\}. \end{aligned} \tag{2.23}$$

The second moments of X are $\mathrm{E}[(X-\mu)(X-\mu)^T] = M\mathrm{tr}(\Omega)$ and $\mathrm{E}[(X-\mu)^T(X-\mu)] = \Omega\mathrm{tr}(M)$. Thus, $\Omega = \mathrm{E}[(X-\mu)^T(X-\mu)]/\mathrm{tr}(M)$ is called the row covariance matrix and $M = \mathrm{E}[(X-\mu)(X-\mu)^T]/\mathrm{tr}(\Omega)$ is called the column covariance matrix. The rows or columns of X are independent if and only if Ω or M is diagonal. In addition, if both Ω and M are scalar matrices, X is called isotropic, which means that X has an isotropic variance.

The MLE algorithm for the matrix normal distribution was proposed by Dutilleul (1999). The MLE of μ is $\bar{x}$. For fixed M, the MLE $\hat{\Omega}$ is given by

$$\hat{\Omega} = \frac{1}{np_L}\sum_{i=1}^{n}(X_i - \bar{X})^T M^{-1}(X_i - \bar{X}); \tag{2.24}$$

and for fixed Ω, the MLE $\hat{M}$ is

$$\hat{M} = \frac{1}{np_R}\sum_{i=1}^{n}(X_i - \bar{X})\Omega^{-1}(X_i - \bar{X})^T. \tag{2.25}$$

Dutilleul (1999) showed that the MLEs of Ω and M estimated from (2.24) and (2.25) are positive definite if and only if $n \geq \max(p_L/p_R, p_R/p_L) + 1$, so a large sample size is not required in order to invert the estimated covariance matrices, as long as the relative ratios of the two dimensions are not too large.

Based on this result, for the general dimension folding PFC model with a log likelihood function (2.12), the MLE $\hat{\Omega}$ is given by

$$\hat{\Omega} = \frac{1}{np_L}\sum_{i=1}^{n}(X_i - \bar{X} - \Gamma_2\beta_2 f_i\beta_1^T\Gamma_1^T)^T M^{-1}(X_i - \bar{X} - \Gamma_2\beta_2 f_i\beta_1^T\Gamma_1^T), \tag{2.26}$$

for fixed Γ_1, Γ_2, β_1, β_2 and M; and the MLE $\hat{M}$ is

$$\hat{M} = \frac{1}{np_R}\sum_{i=1}^{n}(X_i - \bar{X} - \Gamma_2\beta_2 f_i\beta_1^T\Gamma_1^T)\Omega^{-1}(X_i - \bar{X} - \Gamma_2\beta_2 f_i\beta_1^T\Gamma_1^T)^T, \tag{2.27}$$

for fixed Γ_1, Γ_2, β_1, β_2 and Ω.

2.8.2 Proofs

Proof of Propositions 2.1 and 2.3. We demonstrate the proof of Proposition 2.1 first. The condition $\nu|X \sim \nu \mid \Gamma_2^T X\Gamma_1$ is equivalent to $(X|\Gamma_2^T X\Gamma_1, \nu) \sim X|\Gamma_2^T X\Gamma_1$. Treating ν as a parameter matrix and X as data, we can show that $\Gamma_2^T X\Gamma_1$ is a sufficient statistic for $X|\nu$. Since Γ_1 and Γ_2 have the smallest column dimensions, it is equivalent to prove that $\Gamma_2^T X\Gamma_1$ is a minimum sufficient statistic for $X|\nu$. To show this, let $f(X|\nu)$ be the

conditional density function of $X|\nu$, we consider the the log likelihood ratio based on model (2.2):

$$\log\frac{f(X|\nu)}{f(Z|\nu)} = -\frac{1}{2}\text{tr}[(X-\mu)^T(X-\mu) - (Z-\mu)^T(Z-\mu)] + \text{tr}[(\nu(\Gamma_1^T(X-Z)^T\Gamma_2)].$$

It can be seen that $\log f(X|\nu)/f(Z|\nu)$ is a constant in ν if and only if $(\Gamma_1^T(X-Z)^T\Gamma_2) = 0$. Thus, $\Gamma_2^T X\Gamma_1$ is a minimum sufficient statistic and the condition $\nu|X \sim \nu \mid \Gamma_2^T X\Gamma_1$ holds. Similarly, it can be shown that $(\Gamma_1\otimes\Gamma_2)^T\text{vec}(X)$ is a minimum sufficient statistic for $\text{vec}(X)|\nu$ based the log likelihood ratio in (2.3).

For proposition 2.3, when the random error is isotropic, the result can be directly obtained from proposition 2.1. When the random error has a general matrix normal distribution, let $Z = M^{-\frac{1}{2}}X\Omega^{-\frac{1}{2}}$, then $\text{vec}(Z) = (\Omega\otimes M)^{-\frac{1}{2}}\text{vec}(X)$ has covariance $I_{p_Lp_R}$. Transforming model (2.8) into Z scale, we have $\mathcal{S}_{Y|\circ Z\circ} = (\Omega\otimes M)^{-\frac{1}{2}}\text{span}(\Gamma_1\otimes\Gamma_2)$. Based on Proposition 1 in Li, et al. (2010), $\mathcal{S}_{Y|\circ X\circ} = (\Omega^{-\frac{1}{2}}\otimes M^{-\frac{1}{2}})\mathcal{S}_{Y|\circ Z\circ} = \text{span}(\Omega^{-1}\Gamma_1)\otimes\text{span}(M^{-1}\Gamma_2)$.

Proof of Propositions 2.2 and 2.4. We first prove Proposition 2.2. It is easy to see that

$$\begin{aligned}\sum_{i=1}^n \text{tr}[(X_i - G_2\omega_i G_1^T)^T(X_i - G_2\omega_i G_1^T)] = \text{tr}(\sum_{i=1}^n X_i^TX_i) - 2\text{tr}(\sum_{i=1}^n X_i^TG_2\omega_iG_1^T)\\ + \text{tr}(\sum_{i=1}^n \omega_i^T\omega_i)\end{aligned} \tag{2.28}$$

Minimizing (2.28) over G_1, G_2 and ω_i is the same as minimizing $L = \text{tr}(\sum_{i=1}^n \omega_i^T\omega_i) - 2\text{tr}(\sum_{i=1}^n X_i^TG_2\omega_iG_1^T)$. For fixed G_1 and G_2, to obtain the minimizer ν_i over ω_i, we take the first derivative of L corresponding to ω_i and have $\partial L/\partial\omega_i = 2\omega_i - 2(G_2^TX_iG_1)$. Since the second derivate of L on ω_i is positive, the minimum L is obtained when $\hat{\nu}_i = G_2^TX_iG_1$, $i = 1, ..., n$. Thus, the objective function L becomes

$$L = -\text{tr}[G_1^T(\sum_{i=1}^n X_i^T\text{P}_2X_i)G_1], \tag{2.29}$$

where $\mathrm{P}_2 = G_2 G_2^T$. For fixed G_2, L is minimized by choosing the columns of the minimizer $\hat{\Gamma}_1$ over G_1 to be the d_R eigenvectors of $\sum_{i=1}^{n} X_i^T \mathrm{P}_2 X_i$ (or $\sum_{i=1}^{n} X_i^T \mathrm{P}_2 X_i/n$) corresponding to its d_R largest nonzero eigenvalues. Similarly, (2.29) can be written as $L = -\mathrm{tr}[G_2^T(\sum_{i=1}^{n} X_i \mathrm{P}_1 X_i^T) G_2]$, where $\mathrm{P}_1 = G_1 G_1^T$. Then for fixed G_1, the minimizer $\hat{\Gamma}_2$ over G_2 is obtained when its columns are composed by the d_L eigenvectors of $\sum_{i=1}^{n} X_i \mathrm{P}_1 X_i^T$ (or $\sum_{i=1}^{n} X_i \mathrm{P}_1 X_i^T/n$) corresponding to its d_L largest nonzero eigenvalues.

To prove Proposition 2.4, for fixed G_1 and b_1, let $f^* = f(Y) b_1^T$ and $G_{20} \in \mathbb{R}^{p_L \times (p_L - d_L)}$ be the orthogonal compliment of G_2, then we have

$$\begin{aligned}
&\mathrm{E}_n\{\mathrm{tr}[(X - G_2 b_2 f(Y) b_1^T G_1^T)^T (X - G_2 b_2 f(Y) b_1^T G_1^T)]\} \\
=&\mathrm{E}_n\{\mathrm{tr}[(X - G_2 b_2 f^* G_1^T)^T (G_2 G_2^T + G_{20} G_{20}^T)(X - G_2 b_2 f^* G_1^T)]\} \\
=&\mathrm{E}_n\{\mathrm{tr}[(G_2^T X - b_2 f^* G_1^T)(G_2^T X - b_2 f^* G_1^T)^T]\} + \mathrm{E}_n\{\mathrm{tr}[(G_{20}^T X)^T (G_{20}^T X)]\}
\end{aligned} \tag{2.30}$$

We first find the minimizer $\hat{\beta}_2$ over b_2 assuming other terms are fixed. By taking the first derivative of the last equation in (2.30) corresponding to b_2, we have $\partial L_1/\partial b_2 = -2 G_2^T \mathrm{E}_n(X G_1 f^{*^T}) + 2 b_2 \mathrm{E}_n(f^* f^{*^T})$, then $\hat{\beta}_2 = G_2^T \mathrm{E}_n(X G_1 f^{*^T})[\mathrm{E}_n(f^* f^{*^T})]^{-1}$. Replacing b_2 with $\hat{\beta}_2$, the objective function (2.11) becomes

$$\begin{aligned}
&\mathrm{E}_n\{\mathrm{tr}[(X - G_2 \hat{\beta}_2 f^* G_1^T)^T (X - G_2 \hat{\beta}_2 f^* G_1^T)]\} \\
=&\mathrm{E}_n[\mathrm{tr}(X X^T)] - \mathrm{tr}\{P_{G_2} \mathrm{E}_n(X G_1 f^{*^T})[\mathrm{E}_n(f^* f^{*^T})]^{-1} \mathrm{E}_n(X G_1 f^{*^T})^T\}.
\end{aligned}$$

Therefore, the minimizer $\hat{\Gamma}_2$ over G_2 has its columns formed by the first d_L eigenvectors of

$$\mathrm{E}_n(X G_1 f^{*^T})[\mathrm{E}_n(f^* f^{*^T})]^{-1} \mathrm{E}_n(f^* G_1^T X^T),$$

and correspondingly $\hat{\beta}_2 = \hat{\Gamma}_2^T \mathrm{E}_n(X G_1 f^{*^T})[\mathrm{E}_n(f^* f^{*^T})]^{-1}$.

Similarly, given G_2 and b_2, let $f^* = b_2 f(Y)$ and we have

$$\begin{aligned}
&\mathrm{E}_n\{\mathrm{tr}[(X - G_2 b_2 f(Y) b_1^T G_1^T)(X - G_2 b_2 f(Y) b_1^T G_1^T)^T]\} \\
=&\mathrm{E}_n\{\mathrm{tr}[(X^T - G_1 b_1 f^{*^T} G_2^T)^T (X^T - G_1 b_1 f^{*^T} G_2^T)]\}.
\end{aligned}$$

The same procedure for estimating $\hat{\Gamma}_2$ and $\hat{\beta}_2$ can be applied to obtain $\hat{\Gamma}_1$ and $\hat{\beta}_1$. Hence the columns of $\hat{\Gamma}_1$ consist of the first d_R eigenvectors of the matrix

$$\mathrm{E}_n(X^T G_2 f^*)[\mathrm{E}_n(f^{*^T} f^*)]^{-1}\mathrm{E}_n(f^{*^T} G_2^T X),$$

and $\hat{\beta}_1 = \hat{\Gamma}_1^T \mathrm{E}_n(X^T G_2 f^*)[\mathrm{E}_n(f^{*^T} f^*)]^{-1}$.

Proof of Proposition 2.5 and Corollary 2.1. To prove Proposition 2.5(i), for fixed Ω, Γ_1 and β_1, let $X^* = X\Omega^{-\frac{1}{2}}$, and $f^* = f(Y)\beta_1^T\Gamma_1^T\Omega^{-\frac{1}{2}}$. The log likelihood function (2.12) under centered predictors becomes

$$l(\mathcal{S}_{\Gamma_2}, \beta_2, M) = C - \frac{np_R}{2}\log|M| - \frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}\{(X_i^* - \Gamma_2\beta_2 f_i^*)^T M^{-1}(X_i^* - \Gamma_2\beta_2 f_i^*)\},$$

where $C = -\frac{np_L p_R}{2}\log(2\pi) - \frac{np_L}{2}\log|\Omega|$. Treating Γ_2 and M fixed, by taking derivatives of the log likelihood corresponding to β_2, it is easy to obtain that

$$\hat{\beta}_2 = (\Gamma_2^T M^{-1}\Gamma_2)^{-1}\Gamma_2^T M^{-1}\mathbb{X}_L^T\mathbb{F}_L(\mathbb{F}_L^T\mathbb{F}_L)^{-1}.$$

Substituting $\hat{\beta}_2$ back, after some algebra we have

$$l(\mathcal{S}_{\Gamma_2}, M) = C - \frac{np_R}{2}\log|M| - \frac{np_R}{2}\{\mathrm{tr}(M^{-\frac{1}{2}}\tilde{M}M^{-\frac{1}{2}}) - \mathrm{tr}(P_{M^{-\frac{1}{2}}\Gamma_2}M^{-\frac{1}{2}}\hat{\Sigma}_{\mathrm{fit}_L}M^{-\frac{1}{2}}\}.$$

Now treating M fixed, the log likelihood is maximized when the columns of $M^{-\frac{1}{2}}\Gamma_2$ contain the first d_L eigenvectors of $M^{-\frac{1}{2}}\hat{\Sigma}_{\mathrm{fit}_L}M^{-\frac{1}{2}}$. Since $\hat{M}_{\mathrm{res}} = \tilde{M} - \hat{\Sigma}_{\mathrm{fit}_L}$, then the log likelihood reduces to

$$\begin{aligned}
&l(M)\\
&= C - \frac{np_R}{2}\log|M| - \frac{np_R}{2}\{\mathrm{tr}(M^{-\frac{1}{2}}\hat{M}_{\mathrm{res}}M^{-\frac{1}{2}}) - \mathrm{tr}((I - P_{M^{-\frac{1}{2}}\Gamma_2})M^{-\frac{1}{2}}\hat{\Sigma}_{\mathrm{fit}_L}M^{-\frac{1}{2}}\}\\
&= C - \frac{np_R}{2}\log|M| - \frac{np_R}{2}\mathrm{tr}(M^{-1}\hat{M}_{\mathrm{res}}) - \frac{np_R}{2}\sum_{i=d_L+1}^{p_L}\lambda_i(M^{-1}\hat{\Sigma}_{\mathrm{fit}_L}).
\end{aligned}$$

The MLE of M is $\hat{M} = \hat{M}_{\mathrm{res}} + \hat{M}_{\mathrm{res}}^{\frac{1}{2}}\hat{U}_L\hat{D}_L\hat{U}_L^T\hat{M}_{\mathrm{res}}^{\frac{1}{2}}$. This proof can be done in the same way as for Theorem 3.1 in Cook and Forzani (2008). Thus it is omitted. Substitute $\hat{M}$

back to the estimate of $M^{-\frac{1}{2}}\Gamma_2$, we have $\hat{\Gamma}_2 = \hat{M}^{\frac{1}{2}}$ times the first d_L eigenvectors of $\hat{M}^{-\frac{1}{2}}\hat{\Sigma}_{\text{fit}_L}\hat{M}^{-\frac{1}{2}}$ and further $\hat{\beta}_2 = \hat{\Gamma}_2^T\hat{M}^{-1}\mathbb{X}_L^T\mathbb{F}_L(\mathbb{F}_L^T\mathbb{F}_L)^{-1}$.

The results in Proposition 2.5(ii) can be simply obtained by taking transpose of (2.7) and then following the above procedure.

To prove Corollary 2.1, let $A = (I_{p_L} + \hat{D}_L)^{-1}$ and $\tilde{U}_L = \hat{M}_{\text{res}}^{-\frac{1}{2}}\hat{U}_L A^{\frac{1}{2}}$. Applying Lemma A.1 in Cook and Forzani (2008), we have $\text{span}(\tilde{U}_L) = \mathcal{S}_{d_L}(\hat{M}, \hat{\Sigma}_{\text{fit}_L})$. Since A is a full rank diagonal matrix, $\text{span}(\tilde{U}_L)$ is equal to $\text{span}(\hat{M}_{\text{res}}^{-\frac{1}{2}}\hat{U}_L)$, where $\hat{U}_L$ are the first d_L eigenvectors of $\hat{M}_{\text{res}}^{-\frac{1}{2}}\hat{\Sigma}_{\text{fit}_L}\hat{M}_{\text{res}}^{-\frac{1}{2}}$. This implies that $\mathcal{S}_{d_L}(\hat{M}, \hat{\Sigma}_{\text{fit}_L}) = \mathcal{S}_{d_L}(\hat{M}_{\text{res}}, \hat{\Sigma}_{\text{fit}_L})$. Similarly, one can show that $\mathcal{S}_{d_R}(\hat{\Omega}, \hat{\Sigma}_{\text{fit}_R}) = \mathcal{S}_{d_R}(\hat{\Omega}_{\text{res}}, \hat{\Sigma}_{\text{fit}_R})$. Thus the second form holds. Since $\hat{\Sigma}_{\text{fit}_L} = \tilde{M} - \hat{M}_{\text{res}}$, it is easy to see that $\tilde{M}^{-1}\hat{\Sigma}_{\text{fit}_L}$ and $\hat{M}_{\text{res}}^{-1}\hat{\Sigma}_{\text{fit}_L}$ have the same eigenvectors. This provides the result: $\mathcal{S}_{d_L}(\tilde{M}, \hat{\Sigma}_{\text{fit}_L}) = \tilde{M}^{-\frac{1}{2}}\mathcal{S}_{d_L}(\tilde{M}^{-\frac{1}{2}}\hat{\Sigma}_{\text{fit}_L}\tilde{M}^{-\frac{1}{2}}) = \mathcal{S}_{d_L}(\tilde{M}^{-1}\hat{\Sigma}_{\text{fit}_L}) = \mathcal{S}_{d_L}(\hat{M}_{\text{res}}, \hat{\Sigma}_{\text{fit}_L})$. Similarly, $\mathcal{S}_{d_R}(\hat{\Omega}_{\text{res}}, \hat{\Sigma}_{\text{fit}_R}) = \mathcal{S}_{d_R}(\tilde{\Omega}, \hat{\Sigma}_{\text{fit}_R})$. The third form is proved. The last two forms hold since $\tilde{\Omega} = \hat{\Omega}_{\text{res}} + \hat{\Sigma}_{\text{fit}_R}$ and $\tilde{M} = \hat{M}_{\text{res}} + \hat{\Sigma}_{\text{fit}_L}$.

Proof of Proposition 2.6. Recall that $g = \beta_2 f(Y)\beta_1^T$ is the true fitting function and $l = \kappa_2 h(Y)\kappa_1^T$ is the selected fitting function. Let h_i denote $h(Y_i)$ and h_Y denote $h(Y)$. By applying conventional PFC model, we can obtain proper initial values for our algorithm to prove the consistency of our estimates. To do so, we choose a nonzero vector $v \in \mathbb{R}^{p_L}$. Recall that $f(Y)$ is a diagonal fitted matrix with dimensions $r \times r$. Based on model (2.7), we have $X^T v = \Gamma_1\beta_1 f(Y)\beta_2^T\Gamma_2^T v + \varepsilon^T v = \Gamma_1\beta_1 f(Y)\omega + \varepsilon^T v$, where $\omega = \beta_2^T\Gamma_2^T v$ is a r dimensional vector and $\text{var}(\varepsilon^T v) = a\Omega$ with a constant $a = v^T M v$. Let $\tilde{f} \in \mathbb{R}^r$ denote a vector containing the diagonal elements in $f(Y)$, then $f(Y)$ can be written as $\text{diag}(\tilde{f})$. Since $\text{diag}(\tilde{f})\omega = \text{diag}(\omega)\tilde{f}$, it follows that $X^T v = \Gamma_1\beta_1\text{diag}(\omega)\tilde{f} + \varepsilon^T v = \Gamma\beta\tilde{f} + \varepsilon^T v$, where $\Gamma = \Gamma_1$ and $\beta = \beta_1\text{diag}(\omega)$. This forms a conventional PFC model and the unknown parameters Γ, β and Ω can be estimated based on it. Conventional PFC provides $\sqrt{n}$ consistent estimator for the true subspace $\text{span}(\Omega^{-1}\Gamma)$, even when the function $\tilde{f}$ is misspecified by $\tilde{h}$ but they are sufficiently

correlated (Cook and Forzani 2008), where $\text{diag}(\tilde{h}) = h(Y)$. Thus, we can apply conventional PFC to get proper initial values of Γ_1, κ_1 and Ω as $\hat{\Gamma}$, $\hat{\kappa}$ and $\hat{\Omega}$. Let $X^* = X_i\hat{\Omega}^{-\frac{1}{2}}$, $h^* = h_Y\hat{\kappa}^T\hat{\Gamma}^T\hat{\Omega}^{-\frac{1}{2}}$. Then $\hat{\Sigma}_{\text{fit}_L} = (\sum_{i=1}^{n} X_i^* h_i^{*^T}/n)(\sum_{i=1}^{n} h_i^* h_i^{*^T}/n)^{-1}(\sum_{i=1}^{n} h_i^* X_i^{*^T}/n)/p_R$ converges to $\Sigma_{\text{fit}_L} = \text{E}(X\Omega^{-1}\Gamma_1\kappa h_Y^T)Q^{-1}\text{E}(X\Omega^{-1}\Gamma_1\kappa h_Y^T)^T/p_R$, where $Q = \text{var}_c(h_Y\kappa^T) = \text{E}(h_Y\kappa^T\kappa h_Y)$, $\kappa = \kappa_1\text{diag}(\omega)$ and $\omega = \kappa_2^T\Gamma_2^T v$. Using (2.7), we have $\text{E}(X\Omega^{-1}\Gamma_1\kappa h_Y^T) = \text{E}(\Gamma_2 g\kappa h_Y^T) = \Gamma_2\text{cov}_c(g, h_Y\kappa^T) = \Gamma_2\text{cov}_c(g, h_Y\text{diag}(\omega)\kappa_1^T) = \Gamma_2 V\text{diag}(\omega)$, where $V = \text{cov}_c(g, h_Y\kappa_1^T)$. Thus, $\Sigma_{\text{fit}_L} = \Gamma_2 V\text{diag}(\omega)Q^{-1}\text{diag}(\omega)V^T\Gamma_2^T/p_R$. As early defined, $\tilde{M} = \sum_{i=1}^{n} X_i^* X_i^{*^T}/np_R = \sum_{i=1}^{n} X_i\Omega^{-1}X_i^T/np_R$. It follows that $\tilde{M}$ converges at $\sqrt{n}$ rate to $M^* = \text{E}[X\Omega_1^{-1}X^T]/p_R = (\Gamma_2\text{var}_c(g)\Gamma_2^T + M)/p_R$. The last equation is obtained based on (2.7).

From Corollary 2.1, we know $\mathcal{S}_{d_L}(\hat{M}, \hat{\Sigma}_{\text{fit}_L}) = \mathcal{S}_{d_L}(\tilde{M}, \hat{\Sigma}_{\text{fit}_L})$, that is equivalent to $\mathcal{S}_{d_L}(\tilde{M}^{-1}\hat{\Sigma}_{\text{fit}_L})$. Hence $\mathcal{S}_{d_L}(\hat{M}, \hat{\Sigma}_{\text{fit}_L})$ converges to $\mathcal{S}_{d_L}(M^{*^{-1}}\Sigma_{\text{fit}_L})$ at $\sqrt{n}$ rate. Using the fact that $(\Gamma_2 C\Gamma_2^T + M)^{-1} = M^{-1} - M^{-1}\Gamma_2(C^{-1} + \Gamma_2^T M^{-1}\Gamma_2)^{-1}\Gamma_2^T M^{-1}$, we have $\text{span}(M^{*^{-1}}\Sigma_{\text{fit}_L}) = \text{span}\{(\Gamma_2\text{var}_c(g)\Gamma_2^T + M)^{-1}\Gamma_2 V\text{diag}(\omega)Q^{-1}\text{diag}(\omega)V^T\Gamma_2^T\} \subseteq \text{span}\{(\Gamma_2\text{var}_c(g)\Gamma_2^T + M)^{-1}\Gamma_2\} = \text{span}(M^{-1}\Gamma_2)$. Since Γ_2 has full rank d_L and $\text{diag}(\omega)$ has full rank r (Its diagonal elements are all nonzeros with probability one.), we have $\text{span}(M^{*^{-1}}\Sigma_{\text{fit}_L}) = \text{span}(M^{-1}\Gamma_2)$ if and only if the rank of $V = \text{cov}_c(g, h_Y\kappa_1^T)$ is equal to d_L. Since $\rho_L = \text{var}_c^{-\frac{1}{2}}(g)\text{cov}_c(g, l)\text{var}_c^{-\frac{1}{2}}(l) = \text{var}_c^{-\frac{1}{2}}(g)\text{cov}_c(g, h_Y\kappa_1^T)\kappa_2^T\text{var}_c^{-\frac{1}{2}}(l)$ and κ_2 has rank d_L, the rank of ρ_L is equal to the rank of $\text{cov}_c(g, h_Y\kappa_1^T)$.

Similarly, by following the above steps one can show that $\mathcal{S}_{d_R}(\hat{\Omega}^{-1/2}, \hat{\Sigma}_{\text{fit}_R})$ converges to $\text{span}(\Omega^{-1}\Gamma_1)$ at $\sqrt{n}$ rate if and only if $\text{cov}_r(g, h_Y^T\kappa_2^T)$ or, equivalently, ρ_R has rank d_R, based on the fact that $\text{span}(\hat{M}^{-1}\hat{\Gamma}_2) = \mathcal{S}_{d_L}(\hat{M}, \hat{\Sigma}_{\text{fit}_L})$ is $\sqrt{n}$ consistent to $\text{span}(M^{-1}\Gamma_2)$.

Proof of Proposition 2.7. Assume that $E(X) = 0$. Let $Z = M^{-\frac{1}{2}}X\Omega^{-\frac{1}{2}}$ and let

$\mathcal{S}_{Y|\circ Z\circ} = \text{span}(\alpha_1 \otimes \alpha_2)$. Under the elliptically symmetric condition, we have

$$\begin{aligned}(\Omega \otimes M)^{-\frac{1}{2}}\text{E}[\text{vec}(X)|Y] = \text{E}[\text{vec}(Z)|Y] &= \text{E}\{\text{E}[\text{vec}(Z)|(\alpha_1 \otimes \alpha_2)^T\text{vec}(Z), Y]|Y\} \\ &= \text{E}\{\text{E}[\text{vec}(Z)|(\alpha_1 \otimes \alpha_2)^T\text{vec}(Z)]|Y\} \\ &= P_{\alpha_1\otimes\alpha_2}\text{E}[\text{vec}(Z)|Y].\end{aligned} \tag{2.31}$$

Thus, $(\Omega \otimes M)^{-\frac{1}{2}}\text{E}[\text{vec}(X)|Y] \in \mathcal{S}_{Y|\circ Z\circ}$. From model (2.7), we can observe that $\text{E}[\text{vec}(Z)|Y] = (\Omega \otimes M)^{-\frac{1}{2}}(\Gamma_1 \otimes \Gamma_2)(\beta_1 \otimes \beta_2)\text{vec}(f(Y))$. Hence $(\Omega \otimes M)^{-\frac{1}{2}}\text{span}(\Gamma_1 \otimes \Gamma_2) = \text{span}\{\text{E}[\text{vec}(Z)|Y] : \text{over all } Y\} \subseteq \mathcal{S}_{Y|\circ Z\circ}$. By the invariance property $\mathcal{S}_{Y|\circ Z\circ} = (\Omega \otimes M)^{\frac{1}{2}}\mathcal{S}_{Y|\circ X\circ}$, we have $\mathcal{S}_{fPFC} = (\Omega \otimes M)^{-1}\text{span}(\Gamma_1 \otimes \Gamma_2) = \text{span}\{\zeta = (\Omega \otimes M)^{-1}\text{E}[\text{vec}(X)|Y] : \text{over all } Y\} \subseteq \mathcal{S}_{Y|\circ X\circ}$.

Dimension folding SIR can be formulated with $f(Y)$ specified by $h(Y) = \text{diag}\{I(Y \in J_1) - \frac{n_1}{n}, ..., I(Y \in J_{h-1}) - \frac{n_{h-1}}{n}\}^T = \text{diag}(I(\tilde{Y} = 1) - \frac{n_1}{n}, ..., I(\tilde{Y} = h-1)/(h-1) - \frac{n_{h-1}}{n})^T$. Then $\tilde{\zeta} = (\Omega \otimes M)^{-1}\text{E}[\text{vec}(X)|\tilde{Y}] = (\Omega \otimes M)^{-1}(\Gamma_1 \otimes \Gamma_2)(\beta_1 \otimes \beta_2)\text{vec}(h(Y))$. It follows that $\text{span}(\tilde{\zeta}) = \text{span}\{(\Omega \otimes M)^{-1}\text{E}[\text{vec}(X)|\tilde{Y}] : \text{over all } \tilde{Y}\} \subseteq \text{span}\{(\Omega \otimes M)^{-1}(\Gamma_1 \otimes \Gamma_2)\} = \mathcal{S}_{fPFC}$. According to Theorem 1 in Li et al. (2010), the dimension folding SIR subspace $\mathcal{S}_{fSIR}$ is equal to the Kronecker envelope $\mathcal{E}^{\otimes}(\zeta)$, which is the Kronecker product of the two smallest subspaces $\mathcal{S}_{\circ\zeta}\otimes\mathcal{S}_{\zeta\circ}$ such that $\text{span}(\zeta) \subseteq \mathcal{S}_{\circ\zeta}\otimes\mathcal{S}_{\zeta\circ}$. Therefore, $\mathcal{S}_{fSIR} \subseteq \mathcal{S}_{fPFC}$.

Chapter 3

Tensor sliced inverse regression

In this chapter, we propose an efficient SDR method by extending sliced inverse regression (SIR) to data with general m-way array-valued (m-mode tensor-valued) predictors, and refer to it as tensor SIR. We further study its asymptotic properties and demonstrate its advantages by both theoretical and numerical results. Since the method is developed based on the tensor data structure and tensor decompositions, we use "tensor" instead of "array" to facilitate the description of the data throughout this chapter.

3.1 Motivation

SIR was proposed by Li (1991). It is a major supervised dimension reduction technique in non-parametric regression problems. It assumes that the response variable $Y \in \mathbb{R}^1$ depends on the predictor $X \in \mathbb{R}^p$ only through $K(K < p)$ unknown linear combinations of the predictor. Let $B = (\beta_1, \beta_2, ..., \beta_K) \in \mathbb{R}^{p\times K}$. This relationship can be described as $Y|X \sim Y|B^T X$. To build SIR into the sufficient dimension reduction framework, $B^T X$ is called a sufficient reduction of X (Cook 1994, 1998). The matrix B itself is not identified since it can be replaced by any non-singular transformation of its columns. However, the linear space $\mathcal{S}_B = \text{span}(B)$ is identified. As a consequence of this structure one can reduce the dimension of the predictor X by replacing it with its projection $\text{P}_{\mathcal{S}_B}$

onto the subspace $\mathcal{S}_B$, without loss information on the conditional distribution of $Y|X$. That is,

$$Y|X \sim Y|\mathrm{P}_{\mathcal{S}_B}X \text{ or } Y \perp\!\!\!\perp X|\mathrm{P}_{\mathcal{S}_B}X. \tag{3.1}$$

When K is the smallest column rank of B such that (3.1) holds, the subspace $\mathcal{S}_B$ is the central dimension reduction subspace (CS), denoted as $\mathcal{S}_{Y|X}$. The goal of SIR is to estimate $\mathcal{S}_{Y|X}$. We will provide a brief review for the SIR procedure in Section 3.2.1.

Conventional SIR is simple and useful for dimension reduction of vector-valued predictor $X \in \mathbb{R}^p$. However, it is inefficient to tackle problems with more general tensor-valued predictors, such as an $m-$mode tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_m}$. This type of data is commonly encountered in applications. For instance, EEG (electroencephalography) signals in biomedical engineering, gene expression in bioinformatics and images in pattern recognition are usually formed as two-mode tensors (matrices). Video sequences, spatial data and data in social networks often contain three- or multiple mode tensor predictors. Such data are often referred to as multivariate relational data because the tensor-valued predictors represents intrinsic spatial, repeated measured, or other correlated structure among variables. In the EEG data, for example, the brain signals of each subject forms a 256×64 matrix-valued (two-mode tensor-valued) predictor with its rows and columns representing time and location information respectively. Vectorizing such higher order predictors could and typically does lose important information about the data structure and decrease estimation accuracy.

Sufficient dimension reduction (SDR) for tensor-valued predictors has received increasing interest in recent literature. Pioneering work was done by Li et al. (2010), where the authors proposed the idea of dimension folding and developed a class of moment-based dimension folding methods, including dimension folding SIR, to reduce a tensor predictor's multiple dimensions simultaneously. Their methods apply to many moment-based dimension reduction approaches but, as will be shown in later sections, are not very efficient for dealing with higher-order tensor predictors. Other works include longitudinal SIR studied by Pfeiffer et al. (2012) and dimension folding PCA and

PFC developed by Ding and Cook (2013). These two studies focused only on two-mode tensor predictors.

In this chapter, we propose higher-order SDR by extending SIR to general $m-$mode tensor-valued predictors and refer to it as tensor SIR. We provide asymptotic properties for tensor SIR estimates and compare tensor SIR with aforementioned methods in the two-mode tensor case. The proposed method outperforms dimension folding SIR by: (i) reducing the number of parameters and alleviating computation cost; (ii) circumventing high-dimensional covariance matrix inversion; and iii) having easy interpretation and good theoretical properties. In comparison to longitudinal SIR, tensor SIR places fewer restrictions on the covariance structure of $\text{vec}(\mathcal{X})$. It provides the maximum likelihood estimation of the sufficient reduction when $X|Y$ is matrix-normally distributed and $\text{cov}[\text{vec}(X)]$ has a Kronecker structure.

The rest of this chapter is organized as follows. Section 3.2 introduces tensor SIR for two-mode tensor predictors, called two-tensor SIR. Section 3.3 is devoted to the development of tensor SIR for more general $m-$mode tensor predictors. We develop the asymptotic properties for the proposed methods in Section 3.4. Section 3.5 establishes connections between tensor SIR and other high-order SDR methods. Sections 3.6 and 3.7 contain simulation results and data analyses. Discussion is given in Section 3.8. Technical details are given in Section 3.9.

3.2 Two-tensor SIR

Without loss of generality, we assume that the predictors discussed in this chapter have mean zero.

3.2.1 A review of SIR

In the classical setting, $X \in \mathbb{R}^p$ is a predictor vector and $Y \in \mathbb{R}^1$ is a response variable. Let Σ and $\hat{\Sigma}$ be the covariance and sample covariance matrices of X respectively. Assume

that $\mathcal{S}_{Y|X} = \text{span}(\eta)$ is the central subspace for $Y|X$, where $\eta \in \mathbb{R}^{p\times d}$ $(d \leq p)$. Then under the linearity condition (Condition 3.1 in Li (1991), or Proposition 4.2 in Cook (1998)), we have

$$\text{E}(X|Y) = \text{E}[\text{E}(X|\eta^T X, Y)|Y] = \text{E}[\text{E}(X|\eta^T X)|Y] = P^T_{\eta(\Sigma)}\text{E}(X|Y). \tag{3.2}$$

This indicates that $\Sigma^{-1}\text{E}(X|Y) \in \text{span}(\eta)$, and correspondingly $\Sigma^{-1}\text{span}\{\text{cov}[\text{E}(X|Y)]\} \subseteq \mathcal{S}_{Y|X}$. Conventional SIR estimates $\mathcal{S}_{Y|X}$ by the sample estimate $\hat{\Sigma}^{-\frac{1}{2}}$ times the leading d eigen-vectors of $\widehat{\text{cov}}[\hat{\Sigma}^{-\frac{1}{2}}\text{E}(X|Y)]$.

3.2.2 Two-tensor SIR

To introduce the idea of tensor SIR, we first consider a two-mode tensor-valued (matrix-valued) predictor $X \in \mathbb{R}^{p_1\times p_2}$. The response Y is still univariate. In this case, we call tensor SIR as two-tensor SIR to distinguish it from higher order cases.

The sufficient dimension reduction for $X \in \mathbb{R}^{p_1\times p_2}$ is defined as follows.

Definition 3.1 (Li et al. 2010)**.** *Let $B_1 \in \mathbb{R}^{p_1\times d_1}$ $(d_1 \leq p_1)$ and $B_2 \in \mathbb{R}^{p_2\times d_2}$ $(d_2 \leq p_2)$ be two semi-orthogonal matrices that satisfy*

$$Y \perp\!\!\!\perp X|B_1^T X B_2, \text{ or equivalently, } Y \perp\!\!\!\perp \text{vec}(X)|(B_2 \otimes B_1)^T\text{vec}(X). \tag{3.3}$$

i) Then span$(B_2) \otimes$ span(B_1) is called a dimension folding subspace. ii) If d_1 and d_2 both are the smallest column dimensions such that (3.3) holds, then span$(B_2) \otimes$ span(B_1) is called the central tensor (dimension folding) subspace (CTS) for $Y|X$, denoted as $\mathcal{S}_{Y|\circ X\circ}$.

The key idea of SDR for a matrix-valued predictor is to reduce the predictor's row and column dimensions simultaneously without loss of information on $Y|X$. Li et al. (2010) proposed dimension folding SIR for estimating the CTS for $X \in \mathbb{R}^{p_1\times p_2}$. Their method relies on the linearity condition on vec(X) and does not employ the predictor's matrix structure in estimation. Two-tensor SIR considers a matrix-formed

linearity condition and leads to more efficient estimation for the CTS. We propose the methodology below and provide a connection between the two methods in Section 3.5.

For a matrix-valued predictor X, it is reasonable to consider the linearity conditions built in a matrix form. Let $\alpha \in \mathbb{R}^{p_1 \times d_1}(d_1 \leq p_1)$ and $\beta \in \mathbb{R}^{p_2 \times d_2}(d_2 \leq p_2)$ be two full rank matrices. Assume that the conditional means $\mathrm{E}(X|\alpha^T X)$ and $\mathrm{E}(X|X\beta)$ are linear functions of $\alpha^T X$ and $X\beta$ respectively. In other words, there exist two uniquely defined matrices $A \in \mathbb{R}^{p_1 \times d_1}$ and $B \in \mathbb{R}^{p_2 \times d_2}$ such that

$$\mathrm{E}(X|\alpha^T X) = A\alpha^T X, \;\; \mathrm{E}(X|X\beta) = X\beta B^T. \tag{3.4}$$

The next lemma gives the explicit forms for A and B.

Lemma 3.1. *Let $\Omega_1 = \mathrm{E}(XX^T)$ and $\Omega_2 = \mathrm{E}(X^T X)$ be the column and row covariance matrices of X. If condition (3.4) holds for full rank matrices α and β, then $A = \Omega_1\alpha(\alpha^T\Omega_1\alpha)^{-1}$ and $B = \Omega_2\beta(\beta^T\Omega_2\beta)^{-1}$.*

Suppose that $\mathcal{S}_{Y|\circ X\circ} = \mathrm{span}(B_2 \otimes B_1)$. According to Lemma 3.1, we see that

$$\mathrm{E}(X|Y) = \mathrm{E}[\mathrm{E}(X|B_1^T X, Y)|Y] = \mathrm{E}[\mathrm{E}(X|B_1^T X)|Y] = P^T_{B_1(\Omega_1)}\mathrm{E}(X|Y), \tag{3.5}$$

Similarly, we observe $\mathrm{E}(X|Y) = \mathrm{E}(X|Y)P_{B_2(\Omega_2)}$. Therefore, the following equations hold:

$$\mathrm{E}(X|Y) = P^T_{B_1(\Omega_1)}\mathrm{E}(X|Y)P_{B_2(\Omega_2)}, \tag{3.6}$$

or equivalently,

$$\mathrm{E}[\mathrm{vec}(X)|Y] = P^T_{B_2\otimes B_1(\Omega_2\otimes\Omega_1)}\mathrm{E}[\mathrm{vec}(X)|Y]. \tag{3.7}$$

Let Γ_1 and Γ_2 be the bases of $\mathrm{span}(\Omega_1 B_1)$ and $\mathrm{span}(\Omega_2 B_2)$ respectively. Then the CTS can be equivalently written as $\mathcal{S}_{Y|\circ X\circ} = (\Omega_1^{-1} \otimes \Omega_1^{-1})\mathrm{span}(\Gamma_1 \otimes \Gamma_2)$. Correspondingly, (3.7) and (3.6) can be reformulated as

$$\mathrm{E}[\mathrm{vec}(X)|Y] = P_{\Gamma_2\otimes\Gamma_1}\mathrm{E}[\mathrm{vec}(X)|Y], \tag{3.8}$$

and

$$\mathrm{E}(X|Y) = P_{\Gamma_1}\mathrm{E}(X|Y)P_{\Gamma_2}. \tag{3.9}$$

This shows that in addition to the relationship $\mathrm{E}[\mathrm{vec}(X)|Y] \in \mathrm{span}(\Gamma_2 \otimes \Gamma_1)$, the two conditions $\mathrm{span}\{\mathrm{E}(X|Y)P_{\Gamma_2}\} \subseteq \mathrm{span}(\Gamma_1)$ and $\mathrm{span}\{\mathrm{E}(X^T|Y)P_{\Gamma_1})\} \subseteq \mathrm{span}(\Gamma_2)$ hold. They suggest that after projecting the row (column) space of $\mathrm{E}(X|Y)$ onto $\mathrm{span}(\Gamma_2)$ ($\mathrm{span}(\Gamma_1)$), the column (row) space of the projected matrix is a subspace of $\mathrm{span}(\Gamma_1)$ ($\mathrm{span}(\Gamma_2)$). Let $\mathrm{cov}_c[A] = \mathrm{E}[AA^T]$ be the column covariance matrix for any random matrix A. Then the column spaces of $\mathrm{cov}_c[\mathrm{E}(X|Y)P_{\Gamma_2}]$ and $\mathrm{cov}_c[\mathrm{E}(X^T|Y)P_{\Gamma_1}]$ are contained in $\mathrm{span}(\Gamma_1)$ and $\mathrm{span}(\Gamma_2)$ respectively. These relationships provide the basic idea for tensor SIR to estimate the CTS and, as stated in the next proposition, they can be derived by minimizing the discrepancy function

$$\mathrm{E}||\mathrm{E}(X|Y) - P_{\Gamma_1}\mathrm{E}(X|Y)P_{\Gamma_2}||_{\mathrm{F}}^2. \tag{3.10}$$

Proposition 3.1. *Let* (Γ_1, Γ_2) *be the minimizers of the objective function*

$$\mathrm{E}||\mathrm{E}(X|Y) - P_{G_1}\mathrm{E}(X|Y)P_{G_2}||_{\mathrm{F}}^2, \tag{3.11}$$

over all semi-orthogonal matrices $G_1 \in \mathbb{R}^{p_1 \times d_1}$ *and* $G_2 \in \mathbb{R}^{p_2 \times d_2}$. *then*
(i) For fixed G_1, *the columns of the minimizer* Γ_2 *over* G_2 *consist of the* d_2 *eigenvectors of the matrix* $\Sigma_R = \mathrm{E}[\mathrm{E}(X^T|Y)P_{G_1}\mathrm{E}(X|Y)]$ *corresponding to its* d_2 *largest nonzero eigenvalues.*
(ii) For fixed G_2, *the columns of the minimizer* Γ_1 *over* G_1 *are given by the* d_1 *eigenvectors of the matrix* $\Sigma_L = \mathrm{E}[\mathrm{E}(X|Y)P_{G_2}\mathrm{E}(X^T|Y)]$, *corresponding to its* d_1 *largest nonzero eigenvalues.*

According to Proposition 3.1, for an iid sample (X_i, Y_i), $i = 1, ..., n$, by slicing the responses into H categories, one can apply the following algorithm to estimate Γ_1, Γ_2 and the CTS.

1. Generate initial values of Γ_{10} and let $\hat{\Gamma}_1 = \Gamma_{10}$.

2. Given fixed $\hat{\Gamma}_1$, for each slice J_s, $s = 1, ..., H$, compute the sample mean within the category by $\bar{X}_s = \frac{\sum_{Y_i \in J_s} X_i}{n_s}$, where n_s is number of observations within category s. Compute the weighted column covariance matrix $\hat{\Sigma}_R = \sum_{s=1}^{H} \frac{n_s}{n} \bar{X}_s^T \hat{\Gamma}_1 \hat{\Gamma}_1^T \bar{X}_s$ and take the d_2 eigenvectors of $\hat{\Sigma}_R$ corresponding to its d_2 largest eigenvalues to form the columns of $\hat{\Gamma}_2$.

3. For fixed $\hat{\Gamma}_2$, compute the weighted column covariance matrix $\hat{\Sigma}_L = \sum_{s=1}^{H} \frac{n_s}{n} \bar{X}_s \hat{\Gamma}_2 \hat{\Gamma}_2^T \bar{X}_s^T$ and take the d_1 eigenvectors of $\hat{\Sigma}_L$ corresponding to its d_1 largest eigenvalues to form the columns of $\hat{\Gamma}_1$.

4. Repeat 2-3 and iterate with the updated estimators until the objective function $\sum_{s=1}^{H} \frac{n_s}{n} \|\bar{X}_s - P_{\hat{\Gamma}_1} \bar{X}_s P_{\hat{\Gamma}_2}\|_{\mathrm{F}}^2$ converges. Then the CTS $\mathcal{S}_{Y|X}$ is estimated by $(\hat{\Omega}_2^{-1} \otimes \hat{\Omega}_1^{-1})\mathrm{span}(\hat{\Gamma}_2 \otimes \hat{\Gamma}_1)$, where $\hat{\Omega}_1 = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T$ and $\hat{\Omega}_2 = \frac{1}{n} \sum_{i=1}^{n} X_i^T X_i$.

Two-tensor SIR is well connected with conventional SIR and is easily interpreted. It can be treated as an adaptive SIR procedure because it performs an adjusted SIR at each iteration step. To reduce the dimension of each mode, one first needs to project the column space of the other mode into its sufficient reduction subspace and then apply SIR for the reduced predictors. Moreover, two-tensor SIR demonstrates good asymptotic properties. We will show the advantages of this method in Section 3.4 and Section 3.5.

3.3 Multiple mode tensor SIR

3.3.1 Methodology

In this section, we develop tensor SIR for a general m-mode tensor predictor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_m}$ and a univariate response $Y \in \mathbb{R}^1$. Let $\mathcal{M} = \{1, 2, \cdots, m\}$. We first review some important tensor operations and properties.

Definition 3.2 (Tensor product)**.** *The product of an m-mode tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_m}$ and a matrix $A_k \in \mathbb{R}^{d_k \times p_k} (k \in \mathcal{M})$, is a $(p_1 \times \cdots \times p_{k-1} \times d_k \times p_{k+1} \times \cdots \times p_m)$-mode tensor, denoted by $\mathcal{X} \times_k A_k$ with*

$$(\mathcal{X} \times_k A_k)_{i_1 \cdots i_{k-1} j_k i_{k+1} \cdots i_m} = \sum_{i_k=1}^{p_k} \mathcal{X}_{i_1 \cdots i_{k-1} i_k i_{k+1} \cdots i_m} A_{j_k i_k}.$$

Definition 3.3 (Unfolding matrix)**.** *The k-th $(k \in \mathcal{M})$ unfolding matrix of an m-mode tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_m}$ is defined as $X_{(k)} \in \mathbb{R}^{p_k \times (p_{k+1} \cdots p_m p_1 \cdots p_{k-1})}$, where row i of $X_{(k)}$ contains all elements of $\mathcal{X}$ that have the k-th index equal to i.*

For example, let $\mathcal{B} \in \mathbb{R}^{3 \times 4 \times 2}$ be a three mode tensor formed as

$$\mathcal{B}[\,,\,,1] = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{pmatrix} \qquad \mathcal{B}[\,,\,,2] = \begin{pmatrix} 13 & 14 & 15 & 16 \\ 17 & 18 & 19 & 20 \\ 21 & 22 & 23 & 24 \end{pmatrix},$$

then the unfolding along the third mode gives

$$B_{(3)} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 & 17 & 18 & 19 & 20 & 21 & 22 & 23 & 24 \end{pmatrix}$$

Based on Definitions 3.2 and 3.2, the following properties hold. For $A_k \in \mathbb{R}^{p_k \times d_k}$ $(k \in \mathcal{M})$,

i) $\mathcal{Y} = \mathcal{X} \times_k A_k^T \Leftrightarrow Y_{(k)} = A_k^T X_{(k)}$;

ii) $\mathcal{Y} = \mathcal{X} \times_1 A_1^T \times_2 A_2^T \times_3 \cdots \times_m A_m^T \Leftrightarrow Y_{(k)} = A_k^T X_{(k)} (A_m \otimes \cdots \otimes A_{k+1} \otimes A_{k-1} \otimes \cdots \otimes A_1)$;

iii) $\mathrm{vec}(\mathcal{Y}) = \mathrm{vec}(Y_{(1)}) = (A_m \otimes \cdots A_1)^T \mathrm{vec}(\mathcal{X}) = (\bigotimes_{j=m}^{1} A_j)^T \mathrm{vec}(\mathcal{X})$.

For further background on tensor operations, see Kolda (2006). The vectorization of a tensor is usually defined by vectorizing its first mode unfolding matrix, that is, $\mathrm{vec}(\mathcal{X}) = \mathrm{vec}(X_{(1)})$. Hence in iii), the index order of A_j $(j \in \mathcal{M})$ is from m to 1. In general, the choice of the unfolding order is not important as one can always convert $\mathrm{vec}(X_{(k)})$ to $\mathrm{vec}(X_{(1)})$ by elementary row exchange. The goal of SDR for an m-mode

tensor predictor $\mathcal{X}$ is to reduce the predictor's multiple dimensions simultaneously so that the reduced m-mode tensor contains full information about the response Y while preserving the tensor structure. The CTS for $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_m}$ is defined as below.

Definition 3.4. *Let $B_1 \in \mathbb{R}^{p_1 \times d_1}$, $B_2 \in \mathbb{R}^{p_2 \times d_2}$, ..., $B_m \in \mathbb{R}^{p_m \times d_m}$ be m semi-orthogonal matrices. If d_1, d_2, ..., d_m reach the minimum column dimensions such that $Y \perp\!\!\!\perp \mathcal{X} | \mathcal{X} \times_1 B_1^T \times_2 B_2^T \cdots \times_m B_m^T$ or equivalently, $Y \perp\!\!\!\perp \text{vec}(\mathcal{X}) | (\bigotimes_{j=m}^{1} B_j)^T \text{vec}(\mathcal{X})$, then $\text{span}(\bigotimes_{j=m}^{1} B_j)$ is the CTS for $\mathcal{X}$, denoted as $\mathcal{S}_{Y|\mathcal{X}\circ_m}$.*

For tensor-valued predictors, we assume that the linearity condition holds along each mode of the predictor. Let α_k ($k \in \mathcal{M}$) be full rank $p_k \times d_k$, $d_k \leq p_k$, matrices. Assume that $\text{E}(X_{(k)} | \alpha_k^T X_{(k)})$ is a linear function of $\alpha_k^T X_{(k)}$, $k \in \mathcal{M}$. That is, there exist matrices $A_k \in \mathbb{R}^{p_k \times d_k}$ such that

$$\text{E}(X_{(k)} | \alpha_k^T X_{(k)}) = A_k \alpha_k^T X_{(k)}, \;\; k \in \mathcal{M}, \tag{3.12}$$

or equivalently,

$$\text{E}(\mathcal{X} | \mathcal{X} \times_k \alpha_k^T) = \mathcal{X} \times_k A_k \alpha_k^T, \;\; k \in \mathcal{M}, \tag{3.13}$$

Then A_k ($k \in \mathcal{M}$) are uniquely determined by the following lemma.

Lemma 3.2. *Let $\Omega_k = \text{E}(X_{(k)} X_{(k)}^T)$ be the k-th mode covariance matrix of $\mathcal{X}$. If condition (3.12) holds for full rank matrices α_k, then $A_k = \Omega_k \alpha_k (\alpha_k^T \Omega_k \alpha_k)^{-1}$, $k \in \mathcal{M}$.*

Suppose that the CTS of $Y|\mathcal{X}$ is $\text{span}(\bigotimes_{j=m}^{1} B_j)$. According to Lemma 3.2, using the same argument in (3.5), we have $\text{E}(X_{(k)}|Y) = P_{B_k(\Omega_k)}^T \text{E}(X_{(k)}|Y)$, or equivalently, $\text{E}(\mathcal{X}|Y) = \text{E}(\mathcal{X}|Y) \times_k P_{B_k(\Omega_k)}^T$, $k \in \mathcal{M}$. Therefore, by continuing operation on $\text{E}(\mathcal{X}|Y)$ over all $k \in \mathcal{M}$, we have

$$\text{E}[\mathcal{X}|Y] = \text{E}[\mathcal{X}|Y] \times_1 P_{B_1(\Omega_1)}^T \times_2 P_{B_2(\Omega_2)}^T \times_3 \cdots \times_m P_{B_m(\Omega_m)}^T, \tag{3.14}$$

which is equivalent to two other versions

$$\text{E}[X_{(k)}|Y] = P_{B_k(\Omega_k)}^T \text{E}[X_{(k)}|Y] (\bigotimes_{j=m, j\neq k}^{1} P_{B_j(\Omega_j)}^T), \;\; k \in \mathcal{M}. \tag{3.15}$$

Let Γ_k be the bases of span$(\Omega_k B_k)$, $k \in \mathcal{M}$. Then $\mathcal{S}_{Y|\mathcal{X}\circ_m} = (\overset{1}{\underset{j=m}{\bigotimes}} \Omega_j^{-1})\text{span}(\overset{1}{\underset{j=m}{\bigotimes}} \Gamma_j)$ and

$$\mathrm{E}[X_{(k)}|Y] = P_{\Gamma_k}\mathrm{E}[X_{(k)}|Y](\overset{1}{\underset{j=m,\, j\neq k}{\bigotimes}} P_{\Gamma_j}),\ k \in \mathcal{M}. \tag{3.16}$$

The basis matrices Γ_k $(k \in \mathcal{M})$ can be solved as follows.

Proposition 3.2. *Let* $(\Gamma_1, \Gamma_2, ..., \Gamma_m)$ *be the minimizers of the objective function*

$$\mathrm{E}||\mathrm{E}[\mathcal{X}|Y] - \mathrm{E}[\mathcal{X}|Y] \times_1 P_{G_1} \times_2 P_{G_2} \times_3 \cdots \times_m P_{G_m}||_{\mathrm{F}}^2, \tag{3.17}$$

over all semi-orthogonal matrices $G_k \in \mathbb{R}^{p_k \times d_k}$ $(k \in \mathcal{M})$. *Then for fixed* G_1, ..., G_{k-1}, $G_{k+1}, ..., G_m$, *the columns of the minimizer* Γ_k *over* G_k *are given by the first* d_k *eigenvectors of the kernel matrix* $\Sigma_k = \mathrm{E}\{\mathrm{E}[X_{(k)}|Y](\overset{1}{\underset{j=m,\, j\neq k}{\bigotimes}} P_{G_j})\mathrm{E}[X_{(k)}^T|Y]\}$, $\Sigma_k \in \mathbb{R}^{p_k \times p_k}$ $(k \in \mathcal{M})$.

Correspondingly, for an i.i.d sample with m-mode tensor predictors and univariate responses $(\mathcal{X}_i, Y_i)$, one can apply the eigen-based iteration algorithm to estimate Γ_k and the CTS. Let $\bar{\mathcal{X}}_s = \sum_{y_i \in J_s} \mathcal{X}_i / n_s$ be the sample mean within slice J_s and let $\bar{X}_{s_{(k)}}$ be the k-th unfolding matrix of $\bar{\mathcal{X}}_s$, $s = 1, ..., H$, where n_s is the number of observations in J_s.

1. Generate initial values of $\hat{\Gamma}_k^{(0)} \in \mathbb{R}^{p_k \times d_k}$, $k = 2, ..., m$, such that the columns of $\hat{\Gamma}_k^{(0)}$ form the dominant eigen-subspace of the sample estimate of $\mathrm{cov}_c[\mathrm{E}(X_{(k)}|Y)]$. For notation convenience, let $\hat{\Gamma}_k = \hat{\Gamma}_k^{(0)}$.

2. Update $\hat{\Gamma}_1$, ..., $\hat{\Gamma}_m$ sequentially by forming the columns of $\hat{\Gamma}_k$ as the first d_k eigenvectors of

$$\hat{\Sigma}_k = \sum_{s=1}^{H} \frac{n_s}{n} \bar{X}_{s_{(k)}}(\overset{1}{\underset{j=m, j\neq k}{\bigotimes}} P_{\hat{\Gamma}_j})\bar{X}_{s_{(k)}}^T,\ k = 1, ..., m,$$

   with updated parameters.

3. Iterate step 2 until the objective function $\sum_{s=1}^{H} \frac{n_s}{n} \|\bar{\mathcal{X}}_s - \bar{\mathcal{X}}_s \times_1 P_{\hat{\Gamma}_1} \times_2 P_{\hat{\Gamma}_2} \times_3 \cdots \times_m P_{\hat{\Gamma}_m}\|_{\mathrm{F}}^2$ converges. The CTS is then estimated by $(\bigotimes_{j=m}^{1} \hat{\Omega}_j^{-1})\mathrm{span}(\bigotimes_{j=m}^{1} \hat{\Gamma}_j)$, where $\hat{\Omega}_j = \frac{1}{n}\sum_{i=1}^{n} X_{(j)_i} X_{(j)_i}^T$ is the sample column covariance matrix of $X_{(j)}$, $j \in \mathcal{M}$.

Similar as discussed in Section 3.2, this algorithm can be treated as an adaptive SIR algorithm for multiple mode tensor predictors.

3.3.2 Kronecker tensor SIR

In the conventional setting $X \in \mathbb{R}^p$, Cook (2007) showed that SIR provides the MLE for the central subspace when $X|Y$ is multivariate normal. It would be interesting to see whether tensor SIR yields the MLE for the CTS. We propose an alternative tensor SIR procedure that requires a special setting on $\mathrm{cov}\{\mathrm{vec}(\mathcal{X})\}$ and leads to the MLE. The procedure is described below. A statistical justification is given in Section 3.5.4.

Assume that $\mathrm{cov}\{\mathrm{vec}(\mathcal{X})\}$ has a Kronecker structure as

$$\mathrm{cov}\{\mathrm{vec}(\mathcal{X})\} = V_m \otimes V_{m-1} \otimes \cdots \otimes V_1 = \bigotimes_{j=1}^{m} V_j, \tag{3.18}$$

where $V_j \in \mathbb{R}^{p_j \times p_j}$, $j \in \mathcal{M}$. It can be shown that each separate covariance matrix V_j corresponds to the j-th unfolding matrix with $V_j = \mathrm{E}\{X_{(j)} X_{(j)}^T\} / \prod_{i=1, i\neq j}^{m} \mathrm{tr}(V_i)$. Then similar to conventional SIR, tensor SIR can be developed based on the standardized scale $\mathcal{Z} = \mathcal{X} \times_1 V_1^{-1/2} \times_2 V_2^{-1/2} \times \cdots \times V_m^{-1/2}$. Suppose that $\mathcal{S}_{Y|\mathcal{Z}\circ_m} = \mathrm{span}(\bigotimes_{j=m}^{1} \beta_j)$. One can apply the algorithm in Section 3.3.1 to estimate βs using the standardized predictor $\mathcal{Z} = \mathcal{X} \times_1 \hat{V}_1^{-1/2} \times_2 \hat{V}_2^{-1/2} \times \cdots \times \hat{V}_m^{-1/2}$, where $\hat{V}_j = \hat{\Omega}_j = n^{-1} \sum_{i=1}^{n} X_{(j)_i} X_{(j)_i}^T$. The scalar $\prod_{i=1, i\neq j}^{m} \mathrm{tr}(V_i)$ is not essential for the CTS estimation. Therefore, by the equivalence property, the CTS of $Y|\mathcal{X}$ is estimated by $(\bigotimes_{j=m}^{1} \hat{\Omega}_j^{-1/2})\hat{\mathcal{S}}_{Y|\mathcal{Z}\circ_m}$. Since this procedure relies on the Kronecker structure on $\mathrm{cov}\{\mathrm{vec}(\mathcal{X})\}$, we call it as Kronecker tensor SIR, shortened as tensor SIR-K.

When (3.12) holds but cov{vec($\mathcal{X}$)} does not have the Kronecker structure, tensor SIR-K tends to provide biased estimation because the transformation $\mathcal{Z} = \mathcal{X} \times_1 V_1^{-1/2} \times_2 V_2^{-1/2} \times \cdots \times V_m^{-1/2}$ does not standardize the predictor properly.

3.4 Large sample properties

In this section, we provide asymptotic properties of the tensor SIR estimates. We first introduce the following notations. Let $(\Gamma_1, \Gamma_2, ..., \Gamma_m)$ be the minimizers of (3.17) and let $\Gamma_1 = [\gamma_{1,1}, \ldots, \gamma_{1,d_1}]$, $\Gamma_2 = [\gamma_{2,1}, \ldots, \gamma_{2,d_2}]$, ..., $\Gamma_m = [\gamma_{m,1}, \ldots, \gamma_{m,d_m}]$ be the column expressions of these semi-orthogonal matrices. Then the bases of the CTS can be represented as $\{ \bigotimes_{k=m}^{1} (\Omega_k^{-1}\gamma_{k,j_k}), j_1 = 1, ..., d_1, \cdots, j_m = 1, ..., d_m\}$, which are the tensor SIR principal directions. Since we can estimate Ω_k at rate $\sqrt{n}$, the rate for estimating the CTS is determined by how well the Γ_k are estimated. Therefore, we first develop asymptotic properties for Γ_k, $k \in \mathcal{M}$.

Let $\zeta_k = \{(i,j) : \text{vec}(X_{(k)}) = \prod_{i,j} T_{i,j}\text{vec}(\mathcal{X})\}$ be a set of indexes to transform vec($X_{(k)}$) to vec($\mathcal{X}$), where $T_{i,j}$ is an elementary matrix produced by exchanging row i and row j of the identity matrix I_u. Denote the transformation matrices $\prod_{(i,j)\in\zeta_k} T_{i,j}$ by T_k, $k \in \mathcal{M}$. It follows that $\text{vec}(X_{(k)}) = T_k\text{vec}(\mathcal{X})$. Then an alternative expression of Σ_k can be obtained.

Lemma 3.3. *The kernel matrix* $\Sigma_k = \text{E}\{\text{E}[X_{(k)}|Y](\bigotimes_{j=m, j\neq k}^{1} P_{\Gamma_j})\text{E}[X_{(k)}|Y]^T\}$ *is equial to* $\sum_{j_1=1}^{d_1} \cdots \sum_{j_{k-1}=1}^{d_{k-1}} \sum_{j_{k+1}=1}^{d_{k+1}} \cdots \sum_{j_m=1}^{d_m} [(\bigotimes_{l=m,l\neq k}^{1} \gamma_{l,j_l}) \otimes I_{p_k}]^T T_k \Omega T_k^T [(\bigotimes_{l=m,l\neq k}^{1} \gamma_{l,j_l}) \otimes I_{p_k}]$, *where* $\Omega = \text{cov}\{\text{E}[\text{vec}(\mathcal{X})|Y]\}$.

Let $\lambda_{k,1} > \lambda_{k,2} > \cdots > \lambda_{k,d_k} \geq 0$ be the first d_k eigenvalues of Σ_k, $k \in \mathcal{M}$. According to Proposition 3.2, the columns of Γ_k consists of the corresponding eigenvectors of Σ_k.

Therefore, the following equation system

$$\{\sum_{j_1=1}^{d_1}\cdots\sum_{j_{k-1}=1}^{d_{k-1}}\sum_{j_{k+1}=1}^{d_{k+1}}\cdots\sum_{j_m=1}^{d_m}[(\bigotimes_{l=m,l\neq k}^{1}\gamma_{l,j_l})\otimes I_{p_k}]^T T_k\Omega T_k^T[(\bigotimes_{l=m,l\neq k}^{1}\gamma_{l,j_l})\otimes I_{p_k}]\}\gamma_{k,j_k}$$
$$=\lambda_{k,j_k}\gamma_{k,j_k} \tag{3.19}$$

holds for $j_k=1,\ldots,d_k,\ k\in\mathcal{M}$.

We establish the asymptotic properties of tensor SIR based on these equations. From (3.19), all of the leading eigenvalues and eigenvectors $\{\lambda_{k,j_k},\ \gamma_{k,j_k},\ j_k=1,\ldots,d_k,\ k\in\mathcal{M}\}$ can be expressed as functions of Ω. Correspondingly, the sample estimates $\hat{\Gamma}_1,\ldots,\hat{\Gamma}_m$ are functions of $\hat{\Omega}$, where $\hat{\Omega}=\sum_{s=1}^{H}\frac{n_s}{n}\mathrm{vec}(\bar{\mathcal{X}}_s)\mathrm{vec}(\bar{\mathcal{X}}_s)^T$. Hence if the asymptotic distribution of $\hat{\Omega}$ is obtainable, the statistical properties of the tensor SIR estimates can be derived based on a delta method. We adopt the idea of Zhu and Ng (1995) to establish the asymptotics for $\hat{\Omega}$.

Let $g(Y)=\mathrm{E}[\mathrm{vec}(\mathcal{X})|Y]$ be the mean inverse regression function of $\mathrm{vec}(\mathcal{X})$ on Y, let $\epsilon=\mathrm{vec}(\mathcal{X})-g(Y)$ be the regression error, let $u=\prod_{i=1}^{m}p_i$ be the vectorized tensor dimension and let $u_{-k}=\prod_{i=1,i\neq k}^{m}p_i$. The function $g(Y)\in\mathbb{R}^u$ is said to have a total variation of order r if for any closed interval $[-\delta,\delta]$ with fixed real number $\delta>0$,

$$\lim_{n\to\infty}\frac{1}{n^r}\sup_{\mathrm{P}^n([-\delta,\delta])}\sum_{i=1}^{n-1}||g(Y^*_{(i+1)})-g(Y^*_{(i)})||_{\mathrm{F}}=0,$$

where $\mathrm{P}^n([-\delta,\delta])=\{(Y^*_{(1)},\cdots,Y^*_{(n)}):-\delta\leq Y^*_{(1)}\leq\cdots\leq Y^*_{(n)}\leq\delta\}$ is the collection of all n-point partitions of $[-\delta,\delta]$. In addition, $g(Y)$ is called non-expansive in the metric of $G(Y)$ in both side of δ_0, if there exist a non-decreasing function $G(Y)\in\mathbb{R}^1$ and a real number $\delta_0>0$ such that for any two points $Y_1,Y_2\in(-\infty,-\delta_0]$ or $Y_1,Y_2\in[\delta_0,\infty]$,

$$||g(Y_1)-g(Y_2)||_{\mathrm{F}}\leq|G(Y_1)-G(Y_2)|.$$

The asymptotic distribution of $\hat{\Omega}$ is established based on the following regularity assumptions.

Assumption 1. Each slice has the same number of observations, c_n.

Assumption 2. $\mathrm{E}(||\mathrm{vec}(\mathcal{X})||^{4+b}) < \infty$ for some nonnegative number b.

Assumption 3. The inverse regression function $g(Y)$ has a total variation of order $r > 0$.

Assumption 4. $g(Y)$ is non-expansive in the metric of $G(Y)$ on both sides of a positive number δ_0, such that $G^{4+b}(t)P(y > t) \to 0$ as $t \to \infty$.

Lemma 3.4. *Given Assumptions 1-4 with $b > 0$, when $c = O(n^\tau)$, where $\tau = 1/2 - \max\{2r, 2/(4+b)\} > 0$, then $\sqrt{n}[\mathrm{vec}(\hat{\Omega} - \Omega)] \xrightarrow{d} W$, as $n \to \infty$, where W follows a multivariate normal distribution with zero means and covariance matrix*

$$\mathrm{cov}[\mathrm{vec}(\mathcal{X}) \otimes \mathrm{vec}(\mathcal{X}) - \epsilon \otimes \epsilon]. \tag{3.20}$$

Based on the relationship between $\hat{\Gamma}_1, \ldots, \hat{\Gamma}_m$ and $\hat{\Omega}$ and the asymptotic results from Lemma 3.4, we obtain the asymptotic distribution of $\hat{\Gamma}_1, \ldots, \hat{\Gamma}_m$ as follows.

Theorem 3.1. *Under the linearity condition (3.12) and the conditions in Lemma 3.4, $\sqrt{n}[\mathrm{vec}(\hat{\Gamma}_1, \cdots, \hat{\Gamma}_m) - \mathrm{vec}(\Gamma_1, \cdots, \Gamma_m)]$ converges in distribution to $J_m W$, where $J_m = [(\partial\gamma_{1,1}/\partial\mathrm{vec}(\Omega))^T, \ldots, (\partial\gamma_{1,d_1}/\partial\mathrm{vec}(\Omega))^T, \cdots, (\partial\gamma_{m,1}/\partial\mathrm{vec}(\Omega))^T, \ldots, (\partial\gamma_{m,d_m}/\partial\mathrm{vec}(\Omega))^T]^T$ and $\partial\gamma_{k,j_k}/\partial\mathrm{vec}(\Omega) = \{\gamma_{k,j_k} \otimes \mathrm{vec}(\bigotimes_{j=m,\, j\neq k}^{1} P_{\Gamma_j}) \otimes \{\lambda_{k,j_k} I_{p_k} - \mathrm{E}[\mathrm{E}(X_{(k)}|Y)(\bigotimes_{j=m,\, j\neq k}^{1} P_{\Gamma_j}) \mathrm{E}(X_{(k)}|Y)^T]\}^+\}^T [(K_{p_k, u_{-k}} T_k) \otimes T_k]$, for $j_k = 1, \ldots, d_k, k \in \mathcal{M}$.*

The important point of Theorem 3.1 is the consistency and the asymptotic normality of the estimates $\hat{\Gamma}_k$, $k \in \mathcal{M}$. This proves that tensor SIR provides $\sqrt{n}$ consistent estimator for the CTS as $\hat{\Omega}_k$ converges to Ω_k ($k \in \mathcal{M}$) at rate $\sqrt{n}$. The next theorem gives the asymptotic normality of the tensor SIR estimator. Let $\Sigma = \mathrm{cov}[\mathrm{vec}(\mathcal{X})]$, $Q = \mathrm{E}\{\mathrm{cov}[\mathrm{vec}(\mathcal{X})|Y]\}$, and let $\mathcal{I}_k^j \in \mathbb{R}^{p_k \times p_k u_{-k}}$ be a block matrix with its j-th column block equal to I_{p_k} and all other column blocks equal to zero, that is, $\mathcal{I}_k^j = [\mathbf{0} \ \ldots \mathbf{0} \ I_{p_k} \ \mathbf{0} \ \ldots \ \mathbf{0}]$.

Theorem 3.2. *Under the linearity condition (3.12) and the conditions in Lemma 3.4,*

$$\sqrt{n}[\mathrm{vec}(\hat{\Omega}_1^{-1}\hat{\Gamma}_1,\cdots,\hat{\Omega}_m^{-1}\hat{\Gamma}_m)-\mathrm{vec}(\Omega_1^{-1}\Gamma_1,\cdots,\Omega_m^{-1}\Gamma_m)]\xrightarrow{d} HW_1, \tag{3.21}$$

where W_1 follows a multivariate normal distribution with mean zero and covariance matrix $N(\mathbf{0},\mathrm{cov}[(\mathrm{vec}(\mathcal{X})^T\otimes\mathrm{vec}(\mathcal{X})^T,\epsilon^T\otimes\epsilon^T)^T])$, and

$$H=\begin{pmatrix}\partial\mathrm{vec}(\Omega_1^{-1}\Gamma_1)/\partial\mathrm{vec}(\Sigma)^T & \partial\mathrm{vec}(\Omega_1^{-1}\Gamma_1)/\partial\mathrm{vec}(Q)^T\\ \cdots\cdots & \cdots\cdots\\ \partial\mathrm{vec}(\Omega_m^{-1}\Gamma_m)/\partial\mathrm{vec}(\Sigma)^T & \partial\mathrm{vec}(\Omega_m^{-1}\Gamma_m)/\partial\mathrm{vec}(Q)^T\end{pmatrix}$$

with $\partial\mathrm{vec}(\Omega_k^{-1}\Gamma_k)/\partial\mathrm{vec}(\Sigma)^T=-\sum_{j=1}^{u_{-k}}(\Gamma_k^T\Omega_k^{-1}\mathcal{I}_k^j T_k^T\otimes\Omega_k^{-1}\mathcal{I}_k^j T_k^T)+(I_{d_k}\otimes\Omega_k^{-1})\partial\mathrm{vec}(\Gamma_k)/\partial\mathrm{vec}(\Omega)^T$ and $\partial\mathrm{vec}(\Omega_m^{-1}\Gamma_m)/\partial\mathrm{vec}(Q)^T=-(I_{d_k}\otimes\Omega_k^{-1})\partial\mathrm{vec}(\Gamma_k)/\partial\mathrm{vec}(\Omega)^T$, where $\partial\mathrm{vec}(\Gamma_k)/\partial\mathrm{vec}(\Omega)^T=[(\partial\gamma_{k,1}/\partial\mathrm{vec}(\Omega))^T,\ldots,(\partial\gamma_{k,d_k}/\partial\mathrm{vec}(\Omega))^T]^T$, $k\in\mathcal{M}$, are given in Theorem 3.1.

Theorem 3.2 shows that tensor SIR not only provides the $\sqrt{n}$ consistent estimator, but also gives asymptotic normal estimation for the CTS. The covariance matrix of the asymptotic normal distributions is not the main emphasis here as it is not commonly used in SDR studies.

3.5 Connections with other higher-order SDR methods

To the best of our knowledge, all of the other higher-order SDR methods were mainly proposed for matrix-valued predictors. Thus, we analyze the relationship between two-tensor SIR and the other methods for $X\in\mathbb{R}^{p_1\times p_2}$.

3.5.1 Comparison of different linearity conditions

In literature, the higher-order SDR methods, such as dimension folding SIR, longitudinal SIR and dimension folding PCA and PFC, require a linearity condition imposed directly on $\mathrm{vec}(X)$. That is, $\mathrm{E}[\mathrm{vec}(X)|\eta^T\mathrm{vec}(X)]$ is linear function of $\eta^T\mathrm{vec}(X)$ for the basis

matrix $\eta \in \mathbb{R}^{p_1 p_2 \times d} (d \leq p_1 p_2)$ of the CTS. Since η is usually unknown, this condition is generally satisfied when the distribution of $\text{vec}(X)$ is elliptical symmetric (Li, 1991). In comparison, two-tensor SIR uses the tensor-formed linearity condition 3.4. It requires elliptically symmetry only along each mode of X but the different modes need not be jointly elliptically symmetric. However, joint elliptical symmetry is requisite when we directly impose the linearity condition on $\text{vec}(X)$. In addition, longitudinal SIR further requires the Kronecker structure (3.18) on $\text{cov}[\text{vec}(X)]$ for $m = 2$. It can be shown that when the linearity condition holds for $\text{vec}(X)$ and $\text{cov}\{\text{vec}(X)\}$ has the Kronecker structure, Condition 3.4 is satisfied. Yet the opposite direction does not necessarily hold. In other words, when the tensor-formed linearity condition is satisfied, $\text{cov}[\text{vec}(X)]$ needs not be Kronecker structured. The simulation in Section 3.6.1 can serve as a counter example. For higher-order tensors, the tensor-formed linearity condition is weaker.

3.5.2 Two-tensor SIR and dimension folding SIR

Dimension folding SIR relies on the linearity condition on $\text{vec}(\mathbf{X})$. Under this condition, Li et al. (2010) shows that $U(Y) = \Sigma^{-1}\text{E}[\text{vec}(\mathbf{X})|Y]$ is contained in a subspace of the CTS, where $\Sigma = \text{cov}[\text{vec}(\mathbf{X})]$. This subspace is called as the Kronecker envelope of the $U(Y)$, denoted as $\mathcal{E}^{\otimes}(U)$. It is the Kronecker product of two smallest subspaces $\mathcal{S}_{\circ U}$ and $\mathcal{S}_{U\circ}$ such that $\text{span}\{U(Y)\} \subseteq \mathcal{S}_{\circ U} \otimes \mathcal{S}_{U\circ}$, for any Y. Dimension folding SIR then estimates $\mathcal{E}^{\otimes}(U)$ by minimizing

$$\text{E} \parallel U(Y) - (b \otimes a)\omega_y \parallel_{\text{F}}^2,$$

over $b \in \mathbb{R}^{p_L \times u_L} (u_L \leq d_L)$, $a \in \mathbb{R}^{p_R \times u_R} (u_R \leq d_R)$ and $\omega_y \in \mathbb{R}^{u_L u_R}$, where $\text{span}(b) = \mathcal{S}_{\circ U}$, $\text{span}(a) = \mathcal{S}_{U\circ}$ and ω_y is a vector-valued latent function of y. Although dimension folding SIR serve to reduce the predictor's row and column dimensions simultaneously, its estimation procedure is still based on $\text{vec}(\mathbf{X})$. In contrast, two-tensor SIR operates on the original tensor-formed predictor with the following novel aspects: (1) It contains only the CTS parameters in estimation, whereas dimension folding

SIR involves high dimensional matrix inversion Σ^{-1} and additional latent vectors ω_i, $i = 1, \ldots, H$. The maximum covariance matrix inversion of two-tensor SIR is in dimension $\max\{p_1, p_2, \ldots, p_m\} \times \max\{p_1, p_2, \ldots, p_m\}$, which is much smaller than the dimension $\prod_{i=1}^{m} p_i \times \prod_{i=1}^{m} p_i$ required for Σ^{-1} in dimension folding SIR. (2) Two-tensor SIR provides eigen-based subspace estimation. Dimension folding SIR yields element-based estimation using iterative least square algorithm. As a result, it is harder to be generalized to m-mode tensor predictors. (3) Two-tensor SIR shows good theoretical properties. It provides $\sqrt{n}$ consistent estimator for the CTS and is asymptotically normal under certain regularity conditions. (4) When the predictor is vector-valued, tensor SIR coincides with conventional SIR. Dimension folding SIR usually does not have this property.

3.5.3 Two-tensor SIR and longitudinal SIR

Longitudinal SIR (Pfeiffer et al. 2012) addresses dimension reduction for data with longitudinal predictors, a special form of matrix-valued predictors. It estimates the SIR directions, or the basis of the CTS, by $(\hat{\Omega}_2^{-\frac{1}{2}} \otimes \hat{\Omega}_1^{-\frac{1}{2}})\text{span}(\hat{\eta}_2 \otimes \hat{\eta}_1)$, where the columns of $\hat{\eta}_1$ and $\hat{\eta}_2$ are formed by the leading d_1 and d_2 eigenvectors of $\hat{\Psi}_1 = \sum_{s=1}^{H} \frac{n_s}{n} \bar{Z}_s \bar{Z}_s^T$ and $\hat{\Psi}_2 = \sum_{s=1}^{H} \frac{n_s}{n} \bar{Z}_s^T \bar{Z}_s$, the sample estimates of $\Psi_1 = \text{E}[\text{E}(Z|Y)\text{E}(Z^T|Y)]$ and $\Psi_2 = \text{E}[\text{E}(Z^T|Y)\text{E}(Z|Y)]$. Longitudinal SIR requires the linearity condition on $\text{vec}(X)$ and the Kronecker structure on $\text{cov}[\text{vec}(X)]$. Thus, it is more restrictive than two-tensor SIR. In comparison to two-tensor SIR-K, the later uses the leading eigenvectors of $\hat{\Sigma}_L$ and $\hat{\Sigma}_R$, the sample estimates of $\Sigma_L = \text{E}[\text{E}(Z|Y)P_{\Gamma_2}\text{E}(Z^T|Y)]$ and $\Sigma_R = \text{E}[\text{E}(Z^T|Y)P_{\Gamma_1}\text{E}(Z|Y)]$ for estimation. Two-tensor SIR-K has more efficiency gains because it projects $\text{E}(Z|X)$ onto each direction before estimating the other direction. The projection removes redundant information from each direction. It provides an intuition regarding the asymptotic efficiency shown in Section 3.5.4.

3.5.4 Two-tensor SIR and dimension folding PFC

Dimension folding PFC (Ding and Cook 2013) is a model-based SDR method for matrix-valued predictors. It gains efficiency by effectively modeling the conditional mean function $\mathrm{E}(X|Y)$. Under the normality assumption for $X|Y$, dimension folding PFC inherits optimal asymptotic properties from maximum likelihood estimation. The model can be built in several ways depending on the relations between the predictors and response. A general coefficient model is formed as:

$$X = \mu + \Gamma_1 \mathrm{vec}^{-1}(\beta g(Y))\Gamma_2^T + e, \tag{3.22}$$

where $\mu \in \mathbb{R}^{p_1 \times p_2}$ is the overall mean of X, $\Gamma_1 \in \mathbb{R}^{p_1 \times d_1}(d_1 \leq p_1)$ and $\Gamma_2 \in \mathbb{R}^{p_2 \times d_2}(d_2 \leq p_2)$ are semi-orthogonal matrices that reduce the column and row dimensions of X, $\beta \in \mathbb{R}^{d_1 d_2 \times r}$ is a coefficient matrix of rank $d_1 d_2$, $f(Y) \in \mathbb{R}^r$ is a known centered vector-valued function of Y and e is a random error independent of Y. According to Proposition 3 in Ding and Cook (2013), when e follows a matrix normal distribution $N_{p_1 \times p_2}(0, M_1, M_2)$, the CTS of $Y|X$ is $(M_2^{-1} \otimes M_1^{-1})\mathrm{span}(\Gamma_2 \otimes \Gamma_1)$. The relationship between tensor SIR and dimension folding PFC is established below.

Proposition 3.3. *Under the matrix normality of $X|Y$, when Y is categorical and* $\mathrm{cov}[\mathrm{vec}(X)]$ *has a Kronecker structure, two-tensor SIR-K is equivalent to dimension folding PFC and thus provides the MLE of the CTS.*

This implies that when $X|Y$ is matrix-normal and the Kronecker condition on $\mathrm{cov}[\mathrm{vec}(X)]$ is satisfied, two-tensor SIR-K provides the optimal estimation for the CTS.

3.6 Simulation studies

In this section, we evaluate the performance of tensor SIR and compare it with other methods numerically. To access the accuracy of the CTS estimation, we used the criterion

$$\|P_{\hat{\mathcal{S}}_{Y|X\circ_m}} - P_{\mathcal{S}_{Y|X\circ_m}}\|_{\mathrm{F}}, \tag{3.23}$$

to measure the distance between the estimated and true projection matrices of $\mathcal{S}_{Y|X\circ_m}$. To evaluate conventional SIR, we used the criterion

$$\|P_{\hat{\mathcal{S}}_{Y|\mathrm{vec}(X)}} - P_{\mathcal{S}_{Y|\mathrm{vec}(X)}}\|_{\mathrm{F}}. \tag{3.24}$$

3.6.1 Two-mode tensor predictors

We first consider the simulation setup from Li et al. (2010). Let $d_1 = d_2 = 2$ and $p_1 = p_2 = p = 5, 10$. The response Y is a binary variable and was generated from the Bernoulli distribution with success probability equal to 0.5. The matrix-valued predictor X was generated based on the conditional distribution of X given Y which is assumed to be multivariate normal with conditional mean

$$\mathrm{E}(X|Y=0) = \mathbf{0}_{p\times p}, \quad \mathrm{E}(X|Y=1) = \begin{pmatrix} a\mathbf{I_2} & \mathbf{0}_{2\times(p-2)} \\ \mathbf{0}_{(p-2)\times 2} & \mathbf{0}_{(p-2)\times(p-2)} \end{pmatrix}$$

and conditional variance

$$\mathrm{var}(X_{ij}|Y=0) = \begin{cases} 0.1 & \text{if } (i,j)\in A, \\ 1 & \text{if } (i,j)\notin A, \end{cases} \quad \mathrm{var}(X_{ij}|Y=1) = \begin{cases} 1.5 & \text{if } (i,j)\in A, \\ 1 & \text{if } (i,j)\notin A, \end{cases}$$

where a is a scalar and A is the index set $\{(1,1),(1,2),(2,1)\}$. Let $e_i \in \mathbb{R}^p$ be the vector with i-th element equal to 1 and all other elements equal to zero. It can be shown that the CTS is $\Gamma_2 \otimes \Gamma_1$, where $\Gamma_1 = \Gamma_2 = (e_1, e_2)$, and the linearity condition (3.4) holds. However, cov[vec(X)] does not exactly have a Kronecker structure.

We applied two-tensor SIR, two-tensor SIR-K, dimension folding SIR, longitudinal SIR and conventional SIR for the simulated data and evaluated their estimation accuracy based on (3.23) and (3.24). The comparison results for $a = 4$ are listed in Table 3.1 with the shortened names 2-T SIR, 2-T SIR-K, DF-SIR and L-SIR, respectively. It can be seen that two-tensor SIR provides the most accurate estimation as it is less restrictive. Two-tensor SIR-K and longitudinal SIR perform similarly. Although the Kronecker structure is not satisfied for cov[vec(X)], they still outperform dimension folding SIR. The reason is that in this example the Kronecker structure is not violated significantly and dimension folding SIR is computed based on vec(X) that requires more parameters in estimation. Conventional SIR omits the matrix structure of the predictors. Thus, it is less accurate than other methods.

Table 3.1: Comparison of the CTS estimation among different higher-order SDR methods for two-mode tensor predictors when $a = 4$. Each entry is the mean of the estimation errors (3.23) over 500 samples.

Method	n=100	n=200	n=300	n=500	n=800
		$p_1 = p_2 = 5$			
2-T SIR	0.4310	0.3048	0.2518	0.1926	0.1524
2-T SIR-K	0.4366	0.3066	0.2528	0.1931	0.1527
DF-SIR	1.0697	0.7212	0.5785	0.4425	0.3433
L-SIR	0.4366	0.3066	0.2528	0.1931	0.1527
SIR	1.6664	1.5547	1.5123	1.5085	1.4964
		$p_1 = p_2 = 10$			
2-T SIR	0.6429	0.4553	0.3717	0.2902	0.2295
2-T SIR-K	0.6527	0.4568	0.3736	0.2910	0.2299
DF-SIR	1.9465	1.2478	0.9816	0.7452	0.5764
L-SIR	0.6527	0.4568	0.3736	0.2910	0.2299
SIR	2.5057	2.0573	1.9386	1.8422	1.7850

We now vary the conditional mean E($X|Y = 1$) by choosing $a = 50$ and keep all other settings the same. and keep all other settings the same. In this case, the signal of

$\text{cov}[\text{E}\{\text{vec}(X) \mid Y\}]$ is strong and thus the Kronecker structure of $\text{cov}[\text{vec}(X)]$ is violated significantly. Table 3.2 shows that the performance of tensor SIR is not affected but the accuracy of tensor SIR-K and longtudinal SIR are dramatically decreased. The latter two methods highly rely on the Kronecker decomposition of $\text{cov}[\text{vec}(X)]$, resulting in less efficiency when this assumption is not well satisfied.

Table 3.2: Comparison of the CTS estimation among different higher-order SDR methods for two-mode tensor predictors when $a = 50$. Each entry is the mean of the estimation errors (3.23) over 500 samples.

Method	n=100	n=200	n=300	n=500	n=800
		$p_1 = p_2 = 5$			
2-T SIR	0.2922	0.2081	0.1707	0.1298	0.1047
2-T SIR-K	2.1731	1.1536	0.5523	0.1826	0.1213
DF-SIR	0.9921	0.6845	0.5510	0.4148	0.3297
L-SIR	2.2038	1.2521	0.6517	0.1928	0.1256
SIR	1.8576	1.8336	1.7792	1.7872	1.7318
		$p_1 = p_2 = 10$			
2-T SIR	0.3518	0.2473	0.2045	0.1591	0.1244
2-T SIR-K	2.6990	0.8116	0.3005	0.1897	0.1371
DF-SIR	1.8507	1.1761	0.9360	0.7069	0.5524
L-SIR	2.7020	0.9182	0.3237	0.1929	0.1387
SIR	2.6409	2.1492	2.0208	1.9231	1.8598

3.6.2 Three-mode tensor predictors

We next evaluated the performance of tensor SIR for three-mode tensor predictor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$. Let $p_1 = p_2 = p = 5, 10$, $p_3 = 2$ and $d_1 = d_2 = 2$, $d_3 = 1$. The response Y is a binary variable and is generated from Bernoulli (0.5). The tensor predictor $\mathcal{X}$ was generated based on the conditional distribution of $\mathcal{X}$ given Y that is multivariate normal with conditional mean

$$\mathrm{E}\{\mathcal{X}[,,1]|Y=0\} = \mathrm{E}\{\mathcal{X}[,,2]|Y=0\} = \mathrm{E}\{\mathcal{X}[,,2]|Y=1\} = \mathbf{0}_{p\times 2p},$$

$$\mathrm{E}\{\mathcal{X}[,,1]|Y=1\} = \begin{pmatrix} a\mathbf{I_2} & \mathbf{0}_{2\times(p-2)} \\ \mathbf{0}_{(p-2)\times 2} & \mathbf{0}_{(p-2)\times(p-2)} \end{pmatrix}$$

and conditional variance

$$\mathrm{var}\{\mathcal{X}[i,j,k]|Y=0\} = \begin{cases} 0.1 & \text{if } (i,j)\in A, \\ 1 & \text{if } (i,j)\notin A, \end{cases}$$

$$\mathrm{var}\{\mathcal{X}[i,j,k]|Y=1\} = \begin{cases} 1.5 & \text{if } (i,j)\in A, \\ 1 & \text{if } (i,j)\notin A, \end{cases}$$

where A is the index set $\{(1,1,1),(1,2,1),(2,1,1),(1,1,2),(1,2,2),(2,1,2)\}$. It can be seen that the CTS of $\mathcal{X}|Y$ is $\Gamma_3\otimes\Gamma_2\otimes\Gamma_1$, where $\Gamma_1=\Gamma_2=(e_1,e_2)$, the same as that in the two-mode case, and $\Gamma_3=(1,0)^T$. Similar as the two-mode example, the tensor linearity condition (3.12) is satisfied for the data. However, $\mathrm{cov}[\mathrm{vec}(\mathcal{X})]$ cannot be exactly decomposed into the Kronecker structure (3.18) ($m=3$). The larger the value of a, the stronger the violation of the Kronecker assumption. As dimension folding SIR and longitudinal SIR were proposed only for matrix-valued predictors, we applied tensor SIR, tensor SIR-K to estimate the CTS and added the results of conventional SIR for comparison. When $p_1=p_2=10$, the sample covariance matrix $\hat{\Sigma}=\widehat{\mathrm{cov}}[\mathrm{vec}(\mathcal{X})]$ is singular and the ridge-regression-type inverse $(\hat{\Sigma}+0.001I_{200})^{-1}$ is used for conventional SIR. The results based on (3.23) and (3.24) were summarized in Table 3.3 for $a=4$.

It shows the similar phenomenon as that in the two-mode case. Tensor SIR provides the most accurate estimation for the CTS over all sample sizes. Tensor SIR-K performs closely to tensor SIR because of the weak violation of the Kronecker structure on $\mathrm{cov}[\mathrm{vec}(\mathcal{X})]$. Both tensor SIR procedures beat conventional SIR considerably.

Table 3.3: Comparison of the CTS (or CS) estimation among different SDR methods for three-mode tensor predictors when $a = 4$. Each entry is the mean of the estimation errors over 500 samples.

Method	n=100	n=200	n=300	n=500	n=800
		$p_1 = p_2 = 5$			
T-SIR	0.3237	0.2290	0.1908	0.1458	0.1160
T-SIR-K	0.3270	0.2305	0.1917	0.1463	0.1161
SIR	2.1782	2.0897	2.0183	1.9315	1.8699
		$p_1 = p_2 = 10$			
T-SIR	0.4669	0.3335	0.2723	0.2103	0.1666
T-SIR-K	0.4750	0.3350	0.2730	0.2107	0.1668
SIR	2.2108	2.2240	2.2132	2.1718	2.1009

We now vary the conditional mean $\mathrm{E}\{\mathcal{X}[,,1]|Y = 1\}$ using $a = 50$, so $\mathrm{cov}[\mathrm{vec}(\mathcal{X})]$ deviates significantly from the Kronecker structure. From Table 3.4, we see that tensor SIR outperforms tensor SIR-K noticeably as it does not impose any constraint on $\mathrm{cov}[\mathrm{vec}(\mathcal{X})]$. Tensor SIR-K highly depends on the Kronecker constraint. It can perform worse than conventional SIR when $\mathrm{cov}\{\mathrm{vec}(\mathcal{X})\}$ deviates strongly from the Kronecker structure.

In application, we recommend tensor SIR since it is less restrictive. When the Kronecker condition holds, tensor SIR and tensor SIR-K perform closely according to our empirical studies.

3.7 Data analysis

We now analyze the EEG data using tensor SIR. Recall that the EEG data contains 122 subjects, which are divided into alcoholic and control groups with 77 subjects and 45 subjects respectively. For each subject, the predictor contains measurements from 64 channels of electrodes placed on the subject's scalp and sampled at 256 times. The

Table 3.4: Comparison of the CTS (or CS) estimation among different SDR methods for three-mode tensor predictors when $a = 50$. Each entry is the mean of the estimation errors over 500 samples.

Method	n=100	n=200	n=300	n=500	n=800
		$p_1 = p_2 = 5$			
T-SIR	0.2181	0.1536	0.1269	0.0998	0.0773
T-SIR-K	2.8284	2.8284	2.8203	2.8203	2.7977
SIR	2.2145	2.2205	2.2194	2.2184	2.2154
		$p_1 = p_2 = 10$			
T-SIR	0.2525	0.1781	0.1461	0.1144	0.0898
T-SIR-K	2.8284	2.8284	2.8184	2.8140	2.7734
SIR	2.2318	2.2294	2.2297	2.2309	2.2312

64 sites were matched among individuals. Thus, the predictor X is formed as a matrix of dimension 256×64, and the response Y is a binary variable indicating groups. Since the row and column dimensions of the predictor are moderately large, it is very likely that only a few row linear combinations and a few column linear combinations are relevant to classify the response. Moreover, as $n \ll p_L \times p_R$, conventional classification methods, such as linear discriminate analysis (LDA) and quadratic discriminate analysis (QDA) cannot be directly applied to the data. Consequently, higher-order SDR tools are desirable to reduce the predictor's row and column dimensions.

Assume that the observations of the subjects are independently and identically distributed. To evaluate the performance of tensor SIR, we compared its classification rate with the other methods. We used leave-one-out cross validation to obtain training datasets, (X_i, Y_i), $i = 1, \ldots, 122,\ i \neq j$, and test datasets (X_j, Y_j) for $j = 1, \ldots, n$. Two-tensor SIR, dimension folding SIR and longitudinal SIR were applied to each training set and then QDA was employed to the reduced training data, $(\hat{\Gamma}_1^{-1}\hat{\Omega}_1^{-1}X_i\hat{\Omega}_2^{-1}\hat{\Gamma}_1, Y_i)$, $i = 1, \ldots, 122,\ i \neq j$, to obtain the classification rule. This classification rule is then used for the corresponding test dataset $(\hat{\Gamma}_1^{-1}\hat{\Omega}_1^{-1}X_j\hat{\Omega}_2^{-1}\hat{\Gamma}_1, Y_j)$.

Since two-tensor SIR circumvents vectorization of the predictors, it can be directly applied to the original EEG data without any prescreening work. In this case, it correctly classified 88 out of 122 subjects with $(d_L, d_R) = (1, 2)$, while longitudinal SIR classified 75 subjects under the same setting. Dimension folding SIR, however, cannot be directly applied due to the high dimension of $\text{vec}(X)$, which is equal to $256 \times 64 = 16,384$. In order to make comparison with dimension folding SIR, we applied the procedure in Li et al. (2010) to prescreen the predictor's row and column dimensions to $(s_L, s_R) = (15, 15)$ first and then performed higher-order SDR and QDA with $(d_1, d_2) = (1, 2)$. As a result, two-tensor SIR correctly classified 97 subjects of total 122 subjects based on the reduced two-dimensional predictor vector $\hat{\Gamma}_1^{-1}\hat{\Omega}_1^{-1}X\hat{\Omega}_2^{-1}\hat{\Gamma}_1 = (x_{11}, x_{12})$. Both dimension folding SIR and longitudinal SIR provided 92 out of 122 correct decisions. We also tried to prescreen the predictors to other different dimensions, such as $(s_L, s_R) = (10, 10), (30, 30), (95, 15)$. In all cases, tensor SIR showed more accurate classification rates than the other two methods. Figure 3.1 demonstrates the separation of the two groups by two-tensor SIR and longitudinal SIR when $(s_L, s_R) = (95, 15)$. Tensor SIR shows better separation.

3.8 Discussion

We proposed a new approach for sufficient dimension reduction of tensor-valued predictors and refer to it as tensor SIR. In comparison to the existing higher-order dimension reduction methods, tensor SIR is asymptotically consistent and normal under certain regularity conditions. It requires the linearity condition on each mode of the tensor-valued predictor, which is less restrictive than the conditions required by the other methods. In addition, the tensor SIR procedure enhances estimation accuracy and improves computation efficiency. It can be treated as an adaptive SIR and is easily implemented.

To determine the reduced dimensions, one can apply cross-validation to select $(d_1, d_2,$

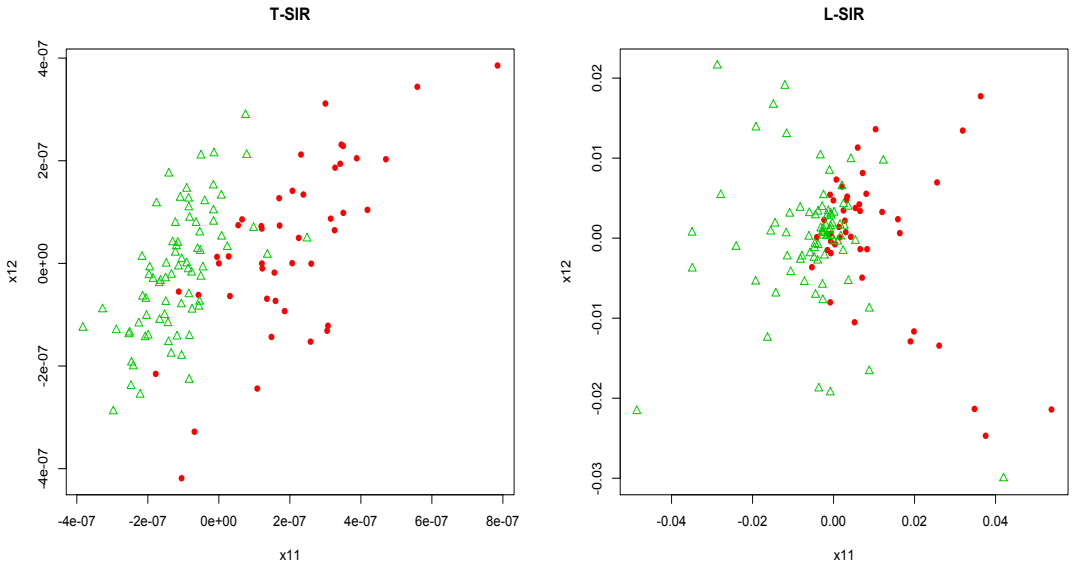


Figure 3.1: Scatter plots by dimension reduced predictors X_{11}, X_{12} with $(s_L, s_R) = (95, 15)$. The triangles indicate alcoholic subjects. The circles represent nonalcoholic subjects.

$\ldots, d_m)$ that provides the smallest prediction or classification error. One can also apply the procedure described in Dong and Li (2010) that adapts the bootstrap idea in Ye and Weiss (2003) to evaluate the multivariate correlation (Hall and Mathiason 1990) between the original estimated principal tensor SIR components $C = (\bigotimes_{j=1}^{m} \hat{\Gamma}_j)\text{vec}(\mathcal{X})$ and the estimated bootstrap tensor SIR components $C_b = (\bigotimes_{j=1}^{m} \hat{\Gamma}_j^b)\text{vec}(\mathcal{X})$. The multivariate correlation is defined as

$$\{\text{var}(C_b)\}^{-\frac{1}{2}}\text{cov}(C_b, C)\{\text{var}(C)\}^{-1}\text{cov}(C, C_b)\{\text{var}(C_b)\}^{-\frac{1}{2}}, \tag{3.25}$$

where $\hat{\Gamma}_j^b$ is the b-th bootstrap sample estimate of Γ_j. Let $\lambda_1, \ldots, \lambda_l$ be the nonzero eigenvalues of (3.25) and let $r^2(C_b, C) = \prod_{i=1}^{l} \lambda_i$ be the eigen-based correlation coefficient. The optimal dimension combination is then selected to maximize the average bootstrap

sample correlation $\bar{r}^2 = \frac{1}{B}\sum_{b=1}^{B} r^2(C_b, C)$.

Although the core of this chapter focuses on extending SIR to tensor-valued predictors, the same logic can be used to study tensor SAVE and other tensor SDR methods based on the tensor-formed linearity condition (3.12). Furthermore, we can relax the linearity condition (3.12) and use the recent results in Ma and Zhu (2012, 2013) to develop semiparametric tensor SDR methods. These extensions are under investigation.

3.9 Appendix

3.9.1 Proof of Lemma 3.1 and Lemma 3.2

Proof of Lemma 3.1. We only show $A = \Omega_1\alpha(\alpha^T\Omega_1\alpha)^{-1}$ since the expression of B can be similarly derived. Consider

$$\mathrm{E}\{\mathrm{E}(X|\alpha^T X)(\alpha^T X)^T\} = \mathrm{E}\{\mathrm{E}(XX^T\alpha|\alpha^T X)\} = \mathrm{E}(XX^T)\alpha = \Omega_1\alpha.$$

Based on the fact that $\mathrm{E}(X|\alpha^T X) = A\alpha^T X$, we have $\mathrm{E}\{\mathrm{E}(X|\alpha^T X)(\alpha^T X)^T\} = \mathrm{E}(A\alpha^T X X^T\alpha) = A\alpha^T\Omega_1\alpha$. Therefore, $A\alpha^T\Omega_1\alpha = \Omega_1\alpha$ and the result is proved.

The proof of Lemma 3.2 can be done based on the same logic shown above and thus is omitted.

3.9.2 Proof of Proposition 3.1 and 3.2

We demonstrate the proof of Proposition 3.1 first. It is easy to see that the objective function (3.11) is equivalent to

$$\begin{aligned}&\mathrm{E}\{\mathrm{tr}\{[\mathrm{E}(X|Y) - P_{G_1}\mathrm{E}(X|Y)P_{G_2}]^T[\mathrm{E}(X|Y) - P_{G_1}\mathrm{E}(X|Y)P_{G_2}]\}\}\\ =&\mathrm{tr}\{\mathrm{E}[\mathrm{E}(X^T|Y)\mathrm{E}(X|Y)]\} - \mathrm{tr}\{\mathrm{E}[P_{G_2}\mathrm{E}(X^T|Y)P_{G_1}\mathrm{E}(X|Y)]\}.\end{aligned}$$

Thus, minimizing (3.11) is the same as maximizing $L = \mathrm{tr}\{\mathrm{E}[P_{G_2}\mathrm{E}(X^T|Y)P_{G_1}\mathrm{E}(X|Y)]\} = \mathrm{tr}\{G_1^T\mathrm{E}[\mathrm{E}(X|Y)P_{G_2}\mathrm{E}(X^T|Y)]G_1\}$. Then for fixed G_2, the minimizer $\hat{\Gamma}_1$ over G_1 is obtained by choosing its columns to be the first d_1 eigenvectors of $\mathrm{E}[\mathrm{E}(X|Y)P_{G_2}\mathrm{E}(X^T|Y)]$.

Similarly, L can be written as $\mathrm{tr}\{G_2^T \mathrm{E}[\mathrm{E}(X^T|Y) P_{G_1} \mathrm{E}(X|Y)] G_2\}$ and thus the minimizer $\hat{\Gamma}_2$ can be similarly proved.

The proof of Proposition 3.2 can be similarly done since the objective function (3.17) is equivalent to

$$\mathrm{E}||\mathrm{E}(X_{(k)}|Y) - P_{\Gamma_k} \mathrm{E}(X_{(k)}|Y)(\bigotimes_{j=m,\, j\neq k}^{1} P_{\Gamma_j})||_{\mathrm{F}}^2, \ k \in \mathcal{M}. \tag{3.26}$$

Treating Γ_j ($j \in \mathcal{M}$, $j \neq k$) fixed, the estimate $\hat{\Gamma}_k$ is obtained.

3.9.3 Proof of Lemma 3.3 and Lemma 3.4

To prove Lemma 3.3, we first consider the column-wise expression of $\bigotimes_{j=m,\, j\neq k}^{1} \Gamma_j$. It is easy to see that

$$\begin{aligned}\bigotimes_{j=m,\, j\neq k}^{1} \Gamma_j &= [\bigotimes_{l=m,\, j\neq k}^{1} \gamma_{l,1},\ (\bigotimes_{l=m,\, j\neq k}^{2} \gamma_{l,1}) \otimes \gamma_{1,2},\ \cdots,\ \bigotimes_{l=m,\, j\neq k}^{1} \gamma_{l,d_l}] \\ &= [\bigotimes_{l=m,\, j\neq k}^{1} \gamma_{l,j_l}]_{\substack{j_1=1,\ldots,d_1 \\ \cdots\cdots \\ j_m=1,\ldots,d_m}}\end{aligned}$$

Then $\bigotimes_{j=m,\, j\neq k}^{1} P_{\Gamma_j} = \bigotimes_{j=m,\, j\neq k}^{1} \Gamma_j \Gamma_j^T = [\bigotimes_{l=m,\, j\neq k}^{1} \gamma_{l,j_l}]_{\substack{j_1=1,\ldots,d_1 \\ \cdots\cdots \\ j_m=1,\ldots,d_m}} [\bigotimes_{l=m,\, j\neq k}^{1} \gamma_{l,j_l}]^T_{\substack{j_1=1,\ldots,d_1 \\ \cdots\cdots \\ j_m=1,\ldots,d_m}}$ and

$$\begin{aligned}&\mathrm{E}(X_{(k)}|Y)(\bigotimes_{j=m,\, j\neq k}^{1} P_{\Gamma_j})\mathrm{E}(X_{(k)}|Y)^T \\ &= \mathrm{E}(X_{(k)}|Y)[\bigotimes_{l=m,\, j\neq k}^{1} \gamma_{l,j_l}]_{\substack{j_1=1,\ldots,d_1 \\ \cdots\cdots \\ j_m=1,\ldots,d_m}} [\bigotimes_{l=m,\, j\neq k}^{1} \gamma_{l,j_l}]^T_{\substack{j_1=1,\ldots,d_1 \\ \cdots\cdots \\ j_m=1,\ldots,d_m}} \mathrm{E}(X_{(k)}|Y)^T.\end{aligned}$$

For any arbitrary j_l ($l = 1, \ldots, m$, $l \neq k$), by taking vectorization operation, we have

$$\begin{aligned}\mathrm{E}(X_{(k)}|Y)[\bigotimes_{l=m,\, j\neq k}^{1} \gamma_{l,j_l}] &= \{[\bigotimes_{l=m,\, j\neq k}^{1} \gamma_{l,j_l}]^T \otimes I_{p_k}\}\mathrm{E}[\mathrm{vec}(X_{(k)})|Y] \\ &= \{[\bigotimes_{l=m,\, j\neq k}^{1} \gamma_{l,j_l}]^T \otimes I_{p_k}\}T_k \mathrm{E}[\mathrm{vec}(\mathcal{X})|Y].\end{aligned}$$

Hence

$$\begin{aligned}
&\mathrm{E}\{\mathrm{E}(X_{(k)}|Y)(\bigotimes_{j=m,\,j\neq k}^{1} P_{\Gamma_j})\mathrm{E}(X_{(k)}|Y)^T\} \\
&= \sum_{j_1=1}^{d_1}\cdots\sum_{j_{k-1}=1}^{d_{k-1}}\sum_{j_{k+1}=1}^{d_{k+1}}\cdots\sum_{j_m=1}^{d_m}[(\bigotimes_{l=m,l\neq k}^{1}\gamma_{l,j_l})\otimes I_{p_k}]^T T_k \mathrm{cov}\{\mathrm{E}[\mathrm{vec}(\mathcal{X})|Y]\}T_k^T\cdot \\
&[(\bigotimes_{l=m,l\neq k}^{1}\gamma_{l,j_l})\otimes I_{p_k}].
\end{aligned}$$

The proof of Lemma 3.4 can be done by following the proofs of Theorem 1 and Theorem 2 in Zhu and Ng (1995). Note that $\sqrt{n}(\hat{\Omega}-\Omega)=\sqrt{n}[(\hat{\Sigma}-\hat{Q})-(\Sigma-Q)]=-M_1-M_2-M_3+M_4^{(1)}-M_4^{(2)}$, where M_1, M_2, M_3 and $M_4^{(2)}$ are the same defined as T_1, T_2, T_3 and $T_4^{(2)}$ in Zhu and Ng (1995), only with the predictor x replaced by $\mathrm{vec}(\mathcal{X})$, and $M_4^{(1)}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mathrm{vec}(\mathcal{X}_i)\mathrm{vec}(\mathcal{X}_i)^T-\varepsilon_i\varepsilon_i^T-\Omega]$. Under the conditions in Lemma 3.4, the elements in M_1, M_2, M_3 and $M_4^{(2)}$ are all equal to $o_p(1)$, as shown in Zhu and Ng (1995), and $\mathrm{vec}(M_4^{(1)})$ converges to $N(\mathbf{0},\mathrm{cov}[\mathrm{vec}(\mathcal{X})\otimes\mathrm{vec}(\mathcal{X})-\epsilon\otimes\epsilon])$.

3.9.4 Proof of Theorem 3.1

The main procedure is to show the gradient matrices $\partial\gamma_{k,j_k}/\partial\mathrm{vec}(\Omega)$. Inspired by Hung et al. (2012), we apply the perturbation method to derive these results. Let Ω be perturbed to $\Omega(\varepsilon)=\Omega+\varepsilon\Omega^*+o(\varepsilon)$. With this perturbation, the eigen equation system becomes

$$\begin{aligned}
&\Sigma_k(\varepsilon)\gamma_{k,j_k}(\varepsilon)=\{\sum_{j_1=1}^{d_1}\cdots\sum_{j_{k-1}=1}^{d_{k-1}}\sum_{j_{k+1}=1}^{d_{k+1}}\cdots\sum_{j_m=1}^{d_m}[(\bigotimes_{l=m,l\neq k}^{1}\gamma_{l,j_l}(\varepsilon))\otimes I_{p_k}]^T T_k\Omega(\varepsilon)T_k^T\cdot \\
&[(\bigotimes_{l=m,l\neq k}^{1}\gamma_{l,j_l}(\varepsilon))\otimes I_{p_k}]\}\gamma_{k,j_k}(\varepsilon)=\lambda_{k,j_k}(\varepsilon)\gamma_{k,j_k}(\varepsilon),\quad j_k=1,\ldots,d_k,\ k\in\mathcal{M},
\end{aligned} \tag{3.27}$$

where $\lambda_{k,j_k}(\varepsilon) = \lambda_{k,j_k} + \varepsilon\lambda^*_{k,j_k} + o(\varepsilon)$ and $\gamma_{k,j_k}(\varepsilon) = \gamma_{k,j_k} + \varepsilon\gamma^*_{k,j_k} + o(\varepsilon)$ satisfying $\gamma_{k,j_k}(\varepsilon)^T\gamma_{k,j_k}(\varepsilon) = 1$ and $\gamma_{k,j_k}(\varepsilon)^T\gamma_{i,j_i}(\varepsilon) = 0$, for $i \neq k$. Therefore,

$$\begin{aligned}\Sigma_k(\varepsilon) = &\sum_{j_1=1}^{d_1}\cdots\sum_{j_{k-1}=1}^{d_{k-1}}\sum_{j_{k+1}=1}^{d_{k+1}}\cdots\sum_{j_m=1}^{d_m}\{[\bigotimes_{l=m,l\neq k}^{1}(\gamma_{l,j_l}+\varepsilon\gamma^*_{l,j_l}+o(\varepsilon))]\otimes I_{p_k}\}^T(T_k\Omega T_k^T+\\ &\varepsilon T_k\Omega^* T_k^T+o(\varepsilon))\{[\bigotimes_{l=m,l\neq k}^{1}(\gamma_{l,j_l}+\varepsilon\gamma^*_{l,j_l}+o(\varepsilon))]\otimes I_{p_k}\}\\ = &\Sigma_k+\varepsilon\Sigma_k^*+o(\varepsilon),\quad k\in\mathcal{M},\end{aligned}$$

where

$$\begin{aligned}\Sigma_k^* = &\sum_{j_1=1}^{d_1}\cdots\sum_{j_{k-1}=1}^{d_{k-1}}\sum_{j_{k+1}=1}^{d_{k+1}}\cdots\sum_{j_m=1}^{d_m}[(\bigotimes_{l=m,l\neq k}^{1}\gamma_{l,j_l})\otimes I_{p_k}]^T T_k\Omega^* T_k^T[(\bigotimes_{l=m,l\neq k}^{1}\gamma_{l,j_l})\otimes I_{p_k}]\\ &+\sum_{j_1=1}^{d_1}\cdots\sum_{j_{k-1}=1}^{d_{k-1}}\sum_{j_{k+1}=1}^{d_{k+1}}\cdots\sum_{j_m=1}^{d_m}[(\bigotimes_{l=m,l\neq k}^{1}\gamma^*_{l,j_l})\otimes I_{p_k}]^T T_k\Omega T_k^T[(\bigotimes_{l=m,l\neq k}^{1}\gamma_{l,j_l})\otimes I_{p_k}]\\ &+\sum_{j_1=1}^{d_1}\cdots\sum_{j_{k-1}=1}^{d_{k-1}}\sum_{j_{k+1}=1}^{d_{k+1}}\cdots\sum_{j_m=1}^{d_m}[(\bigotimes_{l=m,l\neq k}^{1}\gamma_{l,j_l})\otimes I_{p_k}]^T T_k\Omega T_k^T[(\bigotimes_{l=m,l\neq k}^{1}\gamma^*_{l,j_l})\otimes I_{p_k}]\end{aligned}$$

Let $\Lambda = [\bigotimes_{l=m,\, j\neq k}^{1}\gamma_{l,j_l}]_{\substack{j_1=1,\ldots,d_1\\ \cdots\cdots\\ j_m=1,\ldots,d_m}}$ and $\Lambda^* = [\bigotimes_{l=m,\, j\neq k}^{1}\gamma^*_{l,j_l}]_{\substack{j_1=1,\ldots,d_1\\ \cdots\cdots\\ j_m=1,\ldots,d_m}}$ be two matrices with their columns formed by $\bigotimes_{l=m,\, j\neq k}^{1}\gamma_{l,j_l}$ and $\bigotimes_{l=m,\, j\neq k}^{1}\gamma^*_{l,j_l}$, $j_l = 1,\ldots,d_l$ for all $l\in\mathcal{M}$ and $l\neq k$, respectively. Then

$$\begin{aligned}\Sigma_k^* = &\sum_{j_1=1}^{d_1}\cdots\sum_{j_{k-1}=1}^{d_{k-1}}\sum_{j_{k+1}=1}^{d_{k+1}}\cdots\sum_{j_m=1}^{d_m}[(\bigotimes_{l=m,l\neq k}^{1}\gamma_{l,j_l})\otimes I_{p_k}]^T T_k\Omega^* T_k^T[(\bigotimes_{l=m,l\neq k}^{1}\gamma_{l,j_l})\otimes I_{p_k}]\\ &+\mathrm{E}[\mathrm{E}(X_{(k)}\mid Y)(\Lambda^*\Lambda^T+\Lambda{\Lambda^*}^T)\mathrm{E}(X_{(k)}^T\mid Y)].\end{aligned}$$

The expression of Σ_k^* can be further simplified by showing its last term equal to zero. Since tensor SIR can be modeled as

$$X_{(k)}\mid Y = \Gamma_k\nu_y(\bigotimes_{l=m,l\neq k}^{1}\Gamma_l)^T + e,$$

where $\nu_y = \Gamma_k^T \mathrm{E}(X_{(k)} \mid Y)(\bigotimes_{l=m,l\neq k}^{1} \Gamma_l)$ represents a coordinate mean structure and $e \in \mathbb{R}^{p_k \times u_{-k}}$ is a random error with mean zero and constant covariance matrix, it follows

$$\begin{aligned}
&\mathrm{E}[\mathrm{E}(X_{(k)} \mid Y)(\Lambda^*\Lambda^T + \Lambda\Lambda^{*^T})\mathrm{E}(X_{(k)}^T \mid Y)] \\
&= \mathrm{E}\{\mathrm{E}(X_{(k)} \mid Y)[\bigotimes_{l=m,l\neq k}^{1} (\Gamma_l\Gamma_l^{*^T} + \Gamma_l^*\Gamma_l^T)]\mathrm{E}(X_{(k)}^T \mid Y)\} \\
&= \mathrm{E}\{[\Gamma_k\nu_y(\bigotimes_{l=m,l\neq k}^{1} \Gamma_l)^T][\bigotimes_{l=m,l\neq k}^{1} (\Gamma_l\Gamma_l^{*^T} + \Gamma_l^*\Gamma_l^T)][\Gamma_k\nu_y(\bigotimes_{l=m,l\neq k}^{1} \Gamma_l)^T]^T\} \\
&= \mathrm{E}\{\Gamma_k\nu_y[\bigotimes_{l=m,l\neq k}^{1} (\Gamma_l^{*^T}\Gamma_l + \Gamma_l^T\Gamma_l^*)]\nu_y^T\Gamma_k^T\}.
\end{aligned}$$

Now we show that the middle term $\bigotimes_{l=m,l\neq k}^{1} (\Gamma_l^{*^T}\Gamma_l + \Gamma_l^T\Gamma_l^*)$ is equal to zero. Consider the fact that $\gamma_{l,j_l}(\varepsilon)^T\gamma_{l,j_l}(\varepsilon) = 1$ for all $l \in \mathcal{M}$, we have

$$(\gamma_{l,j_l} + \varepsilon\gamma_{l,j_l}^* + o(\varepsilon))^T(\gamma_{l,j_l} + \varepsilon\gamma_{l,j_l}^* + o(\varepsilon)) = \gamma_{l,j_l}^T\gamma_{l,j_l} + \varepsilon(\gamma_{l,j_l}^T\gamma_{l,j_l}^* + \gamma_{l,j_l}^{*^T}\gamma_{l,j_l}) + o(\varepsilon) = 1.$$

Hence $\gamma_{l,j_l}^T\gamma_{l,j_l}^* + \gamma_{l,j_l}^{*^T}\gamma_{l,j_l} = 0$ for all $l \in \mathcal{M}$. Similarly, $\gamma_{l,j_l}^T\gamma_{i,j_i}^* + \gamma_{l,j_l}^{*^T}\gamma_{i,j_i} = 0$ for all $i \neq l$, based on the fact that $\gamma_{l,j_l}(\varepsilon)^T\gamma_{i,j_i}(\varepsilon) = 0$, $i \neq l$. Therefore, for any $l \in \mathcal{M}$,

$$\Gamma_l^{*^T}\Gamma_l + \Gamma_l^T\Gamma_l^* = \begin{pmatrix} \gamma_{l,1}^{*^T} \\ \vdots \\ \gamma_{l,d_l}^{*^T} \end{pmatrix}(\gamma_{l,1}, \cdots, \gamma_{l,d_l}) + \begin{pmatrix} \gamma_{l,1}^T \\ \vdots \\ \gamma_{l,d_l}^T \end{pmatrix}(\gamma_{l,1}^*, \cdots, \gamma_{l,d_l}^*) = 0.$$

Correspondingly, $\mathrm{E}[\mathrm{E}(X_{(k)} \mid Y)(\Lambda^*\Lambda^T + \Lambda\Lambda^{*^T})\mathrm{E}(X_{(k)}^T \mid Y)] = 0$ and

$$\Sigma_k^* = \sum_{j_1=1}^{d_1} \cdots \sum_{j_{k-1}=1}^{d_{k-1}} \sum_{j_{k+1}=1}^{d_{k+1}} \cdots \sum_{j_m=1}^{d_m} [(\bigotimes_{l=m,l\neq k}^{1} \gamma_{l,j_l}) \otimes I_{p_k}]^T T_k \Omega^* T_k^T [(\bigotimes_{l=m,l\neq k}^{1} \gamma_{l,j_l}) \otimes I_{p_k}]. \tag{3.28}$$

From (3.27), using the result of Lemma 2.1 in Sibson (1979), we have

$$\gamma^*_{k,j_k} = \{\lambda_{k,j_k} I_{p_k} - \sum_{j_1=1}^{d_1} \cdots \sum_{j_{k-1}=1}^{d_{k-1}} \sum_{j_{k+1}=1}^{d_{k+1}} \cdots \sum_{j_m=1}^{d_m} [(\bigotimes_{l=m,l\neq k}^{1} \gamma_{l,j_l}) \otimes I_{p_k}]^T T_k \Omega T_k^T \cdot$$
$$[(\bigotimes_{l=m,l\neq k}^{1} \gamma_{l,j_l}) \otimes I_{p_k}]\}^+ \Sigma^*_k \gamma_{k,j_k} \tag{3.29}$$
$$= \left\{ \lambda_{k,j_k} I_{p_k} - \mathrm{E}[\mathrm{E}(X_{(k)} \mid Y)(\bigotimes_{j=m,j\neq k}^{1} P_{\Gamma_j})\mathrm{E}(X^T_{(k)} \mid Y)] \right\}^+ \Sigma^*_k \gamma_{k,j_k}$$

The combination of (3.28) and (3.29) gives

$$\partial\gamma_{k,j_k}/\partial\mathrm{vec}(\Omega) = \{\gamma_{k,j_k} \otimes \mathrm{vec}(\bigotimes_{j=m,\, j\neq k}^{1} P_{\Gamma_j}) \otimes \{\lambda_{k,j_k} I_{p_k} - \mathrm{E}[\mathrm{E}(X_{(k)} \mid Y)(\bigotimes_{j=m,\, j\neq k}^{1} P_{\Gamma_j}) \cdot$$
$$\mathrm{E}(X^T_{(k)} \mid Y)]\}^+\}^T (K_{p_k,u_{-k}} \otimes I_u)(T_k \otimes T_k),$$

for $j_k = 1, \ldots, d_k$, $k =\in \mathcal{M}$. Then by applying the delta method and the result in Lemma 3.4, we finish the proof of Theorem 3.1.

3.9.5 Proof of Theorem 3.2

Theorem 3.2 is established based on the delta method. Since Γ_k, $k \in \mathcal{M}$, are functions of Ω, they are functions of $(\mathrm{vec}(\Sigma)^T, \mathrm{vec}(Q)^T)^T$ as $\Omega = \Sigma - Q$. Moreover, it can be shown Ω_k are functions of Σ. Note that $\Omega_k = \mathrm{E}(X_{(k)}X^T_{(k)}) = \sum_{j=1}^{u_{-k}} \mathrm{E}(X_{(k),j}X^T_{(k),j})$, where $X_{(k),j}$ denotes the j-th column of $X_{(k)}$. On the other hand,

$$\Sigma = \mathrm{E}[\mathrm{vec}(\mathcal{X})\mathrm{vec}(\mathcal{X})^T] = \mathrm{E}[\mathrm{vec}(X_{(1)})\mathrm{vec}(X_{(1)})^T] = T_k \mathrm{E}[\mathrm{vec}(X_{(k)})\mathrm{vec}(X_{(k)})^T] T_k^T$$
$$= T_k \mathrm{E}[(X^T_{(k),1}, \ldots, X^T_{(k),u_{-k}})^T (X^T_{(k),1}, \ldots, X^T_{(k),u_{-k}})] T_k^T.$$

Therefore, $\Omega_k = \sum_{j=1}^{u_{-k}} \mathcal{I}^j_k T_k^T \Sigma T_k \mathcal{I}^{j^T}_k$ where $T_k^T = T_k^{-1}$, $\mathcal{I}^j_k = [\mathbf{0} \ \ldots \mathbf{0} \ I_{p_k} \ \mathbf{0} \ \ldots \ \mathbf{0}] \in \mathbb{R}^{p_k \times p_k u_{-k}}$ is a block matrix with its j-th column block equal to I_{p_k} and all other column blocks equal to zero. This shows that Ω_k, $k \in \mathcal{M}$, are functions of Σ. Thus, $\mathrm{vec}(\Omega_1^{-1}\Gamma_1, \cdots, \Omega_m^{-1}\Gamma_m)$ is a function of $(\mathrm{vec}(\Sigma)^T, \mathrm{vec}(Q)^T)^T$. The sample analogues similarly hold.

Under the conditions in Lemma 3.4, following the proof of Theorem 2 in Zhu and Ng (1995), we have $\sqrt{n}\hat{\Omega} = n^{-\frac{1}{2}}\sum_{i=1}^{n}\epsilon_i\epsilon_i^T$. Along with the fact that $\sqrt{n}\hat{\Sigma} = n^{-\frac{1}{2}}\sum_{i=1}^{n}\text{vec}(\mathcal{X}_i)\text{vec}(\mathcal{X}_i)^T$, we have

$$\sqrt{n}\left[\begin{pmatrix}\text{vec}(\hat{\Sigma})\\ \text{vec}(\hat{\Omega})\end{pmatrix} - \begin{pmatrix}\text{vec}(\Sigma)\\ \text{vec}(\Omega)\end{pmatrix}\right]$$

converges in distribution to a normal random vector W_1 with mean zero and covariance matrix $\text{cov}[(\text{vec}(\mathcal{X})^T \otimes \text{vec}(\mathcal{X})^T, \epsilon^T \otimes \epsilon^T)^T]$. Therefore, applying the delta method, we can conclude that

$$\sqrt{n}[\text{vec}(\hat{\Omega}_1^{-1}\hat{\Gamma}_1, \cdots, \hat{\Omega}_m^{-1}\hat{\Gamma}_m) - \text{vec}(\Omega_1^{-1}\Gamma_1, \cdots, \Omega_m^{-1}\Gamma_m)]$$

converges in distribution to HW_1, where H is the gradient matrix given in Theorem 3.2 with

$$\begin{aligned}\partial\text{vec}(\Omega_k^{-1}\Gamma_k)/\partial\text{vec}(\Sigma)^T =& \partial\text{vec}(\Omega_k^{-1}\Gamma_k)/\partial\text{vec}(\Omega_k)^T \cdot \partial\text{vec}(\Omega_k)/\partial\text{vec}(\Sigma)^T + \\ & \partial\text{vec}(\Omega_k^{-1}\Gamma_k)/\partial\text{vec}(\Gamma_k)^T \cdot \partial\text{vec}(\Gamma_k)/\partial\text{vec}(\Sigma)^T\end{aligned}$$

and $\partial\text{vec}(\Omega_k^{-1}\Gamma_k)/\partial\text{vec}(Q)^T = \partial\text{vec}(\Omega_k^{-1}\Gamma_k)/\partial\text{vec}(\Gamma_k)^T \cdot \partial\text{vec}(\Gamma_k)/\partial\text{vec}(Q)^T$, $k \in \mathcal{M}$. Then the expression of H is given by

$$\frac{\partial\text{vec}(\Omega_k^{-1}\Gamma_k)}{\partial\text{vec}(\Omega_k)^T} = \frac{\partial(\Gamma_k^T \otimes I_{p_k})\text{vec}(\Omega_k^{-1})}{\partial\text{vec}(\Omega_k)^T} = -(\Gamma_k^T \otimes I_{p_k})(\Omega_k^{-1} \otimes \Omega_k^{-1}) = -(\Gamma_k^T\Omega_k^{-1} \otimes \Omega_k^{-1}),$$

$\partial\text{vec}(\Omega_k)/\partial\text{vec}(\Sigma)^T = \partial(\sum_{j=1}^{u_{-k}}\mathcal{I}_k^j T_k^T \Sigma T_k \mathcal{I}_k^{j^T})/\partial\text{vec}(\Sigma)^T = \sum_{j=1}^{u_{-k}}\mathcal{I}_k^j T_k^T \otimes \mathcal{I}_k^j T_k^T$,
$\partial\text{vec}(\Omega_k^{-1}\Gamma_k)/\partial\text{vec}(\Gamma_k)^T = I_{d_k} \otimes \Omega_k^{-1}$ and $\partial\text{vec}(\Gamma_k)/\partial\text{vec}(\Sigma)^T = -\partial\text{vec}(\Gamma_k)/\partial\text{vec}(Q)^T$ $= \partial\text{vec}(\Gamma_k)/\partial\text{vec}(\Omega)^T$, where $\partial\text{vec}(\Gamma_k)/\partial\text{vec}(\Omega)^T$ is shown in Theorem 3.1.

3.9.6 Proof of Proposition 3.3

We consider the dimension folding PFC model (8.3) in Ding and Cook (2013). When the range of the response is divided into h slices, the fitting function $f(Y)$ in (8.3) is

naturally determined as $(I(Y \in H_1)-n_1/n, I(Y \in H_2)-n_2/n), \ldots, I(Y \in H_h)-n_h/n)^T$. Let $\mathcal{S}_d(A)$ be the subspace spanned by the leading d eigenvectors of A and $\mathcal{S}_d(A,B) = A^{-1/2}\mathcal{S}_d(A^{-1/2}BA^{-1/2})$. Based on Corollary 1 in Ding and Cook (2013), the MLE of the CTS is equal to $\mathcal{S}_{d_2}(\hat{\Omega}_2, \hat{\Sigma}_{\text{fit}_R}) \otimes \mathcal{S}_{d_1}(\hat{\Omega}_1, \hat{\Sigma}_{\text{fit}_L})$, where

$$\hat{\Sigma}_{\text{fit}_R} = n^{-1}\sum_{s=1}^{H} n_s \bar{X}_s \hat{M}_2^{-1}\hat{\Gamma}_2\hat{\Gamma}_2^T\hat{M}_2^{-1}\bar{X}_s^T, \;\; \hat{\Sigma}_{\text{fit}_L} = n^{-1}\sum_{s=1}^{H} n_s \bar{X}_s^T \hat{M}_1^{-1}\hat{\Gamma}_1\hat{\Gamma}_1^T\hat{M}_1^{-1}\bar{X}_s.$$

To prove Proposition 3.3, using the results in Section 3.3.2, it is sufficient to show that $\mathcal{S}_{d_1}(\hat{\Omega}_1, \hat{\Sigma}_{\text{fit}_L}) = \text{span}(\hat{\Omega}_1^{-1/2}\hat{\beta}_1)$ and $\mathcal{S}_{d_2}(\hat{\Omega}_2, \hat{\Sigma}_{\text{fit}_R}) = \text{span}(\hat{\Omega}_2^{-1/2}\hat{\beta}_2)$. We only demonstrate the first equation since the second one is satisfied based on the first equation. Since $\mathcal{S}_{d_1}(\hat{\Omega}_1, \hat{\Sigma}_{\text{fit}_L}) = \hat{\Omega}_1^{-1}\text{span}_d\{n^{-1}\sum_{s=1}^{H} n_s \bar{X}_s \hat{M}_2^{-1}\hat{\Gamma}_2\hat{\Gamma}_2^T\hat{M}_2^{-1}\bar{X}_s^T\}$, it is equal to $\hat{\Omega}_1^{-1}\text{span}_d\{n^{-1}\sum_{s=1}^{H} n_s \bar{X}_s \hat{\Omega}_2^{-1/2}\hat{\beta}_2\hat{\beta}_2^T\hat{\Omega}_2^{-1/2}\bar{X}_s^T\}$ based on the equation $\hat{\Omega}_2^{-1/2}\hat{\beta}_2 = \hat{M}_2^{-1}\hat{\Gamma}_2$. This equation holds by initiating $\Gamma_{20}, \beta_{20}, \Omega_{20}$ and M_{20} such that $\Omega_{20}^{-1/2}\beta_{20} = M_{20}^{-1}\Gamma_{20}$.

Chapter 4

Matrix-variate regressions and the envelope models

In the previous chapters, we mainly concern data with matrix- or array-valued predictors. The response variables are still univariate. In this chapter, we broaden our vision to data with matrix-valued responses. The predictors can be either matrix-valued, or vector-valued, or univariate. We propose matrix-variate regressions (matrix regressions) for such data and establish envelope models for matrix-variate regressions to achieve simultaneous dimension reduction and model estimation. The idea of matrix regressions can be naturally extended to array-variate regressions when data have higher-order responses and/or predictors.

4.1 Motivation

Data with a matrix-valued response for each experimental unit are commonly encountered in contemporary statistical applications. For example, a longitudinal multivariate response can be integrally treated as a matrix-valued variable by designing rows and columns to be time and variates. Temporal and spatial data, multivariate growth curve data, image data and data from cross-over designs also generate matrix-valued responses.

In the twin cross-over bioassay of insulin by the rabbit blood sugar method (Vϕlund, 1980), each rabbit received two different treatments on two days. Blood sugar was measured at hourly intervals for six hours each day. In this case, the response for each rabbit is a 2×6 matrix, with rows and columns indicating treatments and time respectively. The EEG data (Li et al. 2010) is another example that contains a temporal-spatial matrix-variate from 77 alcoholic subjects and 45 non-alcoholic subjects. The electrical records of each subject form a matrix of dimensions 256×64 that can be treated as a matrix-valued response variable when we investigate the association between the brain signals and alcoholism.

In both examples, the components of the matrix-variates are dependent among rows and columns. Vectorizing such response, or modeling the row or column vectors separately, typically loses the dependency information and fails to capture the data structure. Tackling matrix-variates directly can circumvent this issue. Research into this topic has gained considerable interest in recent years. Li et al. (2010) proposed a class of sufficient dimension reduction (SDR) methods for data with matrix-valued predictors. Ding and Cook (2013) developed model-based SDR for the same goal. On another track, Hung and Wang (2013) and Zhou et al. (2013) extended generalized linear model (GLM) to matrix- and tensor-valued predictors respectively, for analyzing image data. All these methods, however, address data with matrix or tensor-valued predictors.

Research into matrix-variate analysis, which directly models a matrix-valued response, is still in a nascent stage. Studies on this topic are very limited. Viroli (2012) proposed matrix-variate regression analysis that focuses on special regression cases where either covariates are time-independent or error terms have independent rows. In the twin cross-over assay or the EEG data, however, we encounter dependency among both rows and columns of the matrix responses. In this chapter, we study matrix-variate regressions under the more general framework and propose a class of envelope models for efficient estimation in matrix-variate regressions. The proposed methods directly model matrix-valued responses and use the intrinsic data structure to

reduce the number of parameters in estimation. By applying the idea of enveloping, one can extract material information and eliminate immaterial information in estimation, thus leading to more efficient model estimation.

The remainder of this chapter is organized as follows. Section 4.2 establishes matrix-variate regressions and connects them with conventional regression models. Section 4.3 is devoted to the development of envelope models for matrix-variate regressions. Section 4.4 studies theoretical properties of matrix regression and envelope matrix regression models. Section 4.5 discusses dimension selection procedures. Sections 4.6 and 4.7 provide simulations and real data analysis. Technical proofs are given in Section 4.8.

4.2 Matrix-variate regression

4.2.1 Model formulation

Generally, a two-way measurement layout can be treated integrally as a matrix-valued variable, denoted as $Y \in \mathbb{R}^{r\times m}$. Assume that the matrix-variate Y follows a distribution with matrix mean $\mu \in \mathbb{R}^{r\times m}$, column covariance matrix $\Sigma_1 \in \mathbb{R}^{r\times r}$ and row covariance matrix $\Sigma_2 \in \mathbb{R}^{m\times m}$. Then the simplest model for Y is:

$$Y = \mu + \varepsilon, \tag{4.1}$$

where $\mu \in \mathbb{R}^{r\times m}$ is the overall mean, and ε is the random error with mean zero and covariance matrices Σ_1 and Σ_2. The column and row covariance matrices are defined as $\Sigma_1 = \mathrm{cov}_c(\varepsilon) = \mathrm{E}(\varepsilon\varepsilon^T)/\mathrm{tr}(\Sigma_2)$ and $\Sigma_2 = \mathrm{cov}_r(\varepsilon) = \mathrm{E}(\varepsilon^T\varepsilon)/\mathrm{tr}(\Sigma_1)$. In this setting, it can be shown that $\mathrm{cov}[\mathrm{vec}(\varepsilon)] = \Sigma_2 \otimes \Sigma_1$ (De Waal 1985). The Kronecker covariance structure reveals a relational characteristic of the matrix-variate Y, as the covariances of the column vectors of ε are all proportional to Σ_1 and the covariances of the row vectors of ε are all proportional to Σ_2. Such relationship is usually desirable for matrix-valued variables, especially for multivariate repeated measures and multivariate longitudinal data, because of the intrinsic relationship among elements. For instance, the EEG data

contains measurements of each subject from different time (row) and different scalp locations (column). It is reasonable to assume similar variations among measurements over rows and measurements over columns. When the Kronecker structure does not hold, a general covariance matrix $\text{cov}[\text{vec}(\varepsilon)] = \Sigma$ can be applied. Hypothesis tests for the Kronecker covariance structure can be found in Shitan and Brockwell (1995), Lu and Zimmerman (2005), and Roy and Khattree (2005).

General formulation

Simply modeling the response Y as (4.1) is usually not sufficient in matrix-variate analysis. In applications, one often need to consider covariate effects on $Y \in \mathbb{R}^{r\times m}$. The covariate effects can be matrix-valued, or vector-valued, or univariate depending on the specific data. In the insulin assay data, for example, the covariates are formed as a 2×2 matrix with elements indicating different treatments and dose levels and the goal is to investigate how the treatments and dose levels influence the blood sugar concentration of each rabbit.

Given a matrix-variate predictor $X \in \mathbb{R}^{p_1\times p_2}$, we define the matrix regression of $Y \in \mathbb{R}^{r\times m}$ on X as

$$Y = \mu + \beta_1 X \beta_2^T + \varepsilon, \tag{4.2}$$

or equivalently,

$$\text{vec}(Y) = \text{vec}(\mu) + (\beta_2 \otimes \beta_1)X + \text{vec}(\varepsilon), \tag{4.3}$$

where μ is the overall mean, $\beta_1 \in \mathbb{R}^{r\times p_1}$ and $\beta_2 \in \mathbb{R}^{m\times p_2}$ are the row and column coefficients, and ε is the random error defined in (4.1) and it is independent of X. From (4.2), it is easy to see that $\beta_2 \otimes \beta_1$ is equal to $c\beta_2 \otimes \frac{1}{c}\beta_1$. Therefore, the coefficient matrices might also have other kronecker decompositions and are not be uniquely defined. Lemma 4.1 provides a justification regarding this concern.

Lemma 4.1. *Under (4.2), the coefficient matrices β_1 and β_2 are uniquely defined only up to a constant.*

The matrix regression (4.2) is a new model formulation that has not been discussed in the literature. It incorporates simple regression, multiple regression and multivariate multiple regression as special cases under different settings of X and Y. For instance, when both response and predictor are vector-valued, β_1 and β_2 can be combined and (4.2) coincides with the multivariate multiple regression model.

Compared to the multivariate analysis that regresses vec(Y) on vec(X), the matrix regression (4.2) utilizes the original data structure to explore the intrinsic linear relationship between the response and the predictor. It shows a multiplicative coefficient form as

$$Y_{ij} = \mu_{ij} + \sum_{k=1}^{r}\sum_{l=1}^{m} \beta_{ik}^{(2)}\beta_{lj}^{(1)} X_{kl} + \varepsilon_{ij},$$

where $\beta_{ik}^{(2)}$ is the ik-th element of β_2 and $\beta_{lj}^{(1)}$ is the lj-th element of β_1. For the response elements in the same row, the regression coefficients varies only from β_1, and similarly, for the response elements in the same column, the coefficients varies only from β_2. This reveals the relational characteristic among the elements of the matrix-valued response and the distinctive functions of the row and column coefficient matrices. By capturing the row and column relationships, model (4.2) reduces the number of parameters by $rmp_1p_2 - (rp_1 + mp_2) + rm(rm+1)/2 - r(r+1)/2 - m(m+1)/2$, where $rmp_1p_2 - (rp_1 + mp_2)$ is the parameter reduction from the coefficients and the rest part is the parameter reduction from the covariance matrices. The total number of the reduced parameters could be very large when the matrix dimensions of X and Y are relative high. To validating the coefficient structure of (4.2), one can apply the likelihood ratio test provided in Section 4.2.3. In comparison to the matrix-variate regression proposed in Viroli (2012), model (4.2) is more generalized. It can be used to map the linear relationship between different dimensional response and predictor. In addition, it considers a more general covariance structure. The time-dependent model

in Viroli (2012) is actually a special case of our model (4.4) discussed next, by assuming the row covariance matrix Σ_1 equal to I_m.

Special formulations

Case 1. When the column (row) dimension of X is the same as the corresponding dimension of Y and the columns (rows) of X and Y represent repeated measures or similar characteristics, it is more appropriate to build (4.2) with $\beta_2 = I_m$ ($\beta_1 = I_r$), as each column (row) of Y is usually associated to the corresponding column (row) of X. For instance, if the matrix-variate response and predictor are both formed by multivariate variables measured at time $1, 2, \ldots, m$, then the i-th column of the response corresponds to the i-th column of the predictor only. It is not necessary to regress each column of Y on all elements of X. Therefore, the matrix regression model is simplified to :

$$Y = \mu + \beta_1 X + \varepsilon, \tag{4.4}$$

where all the model terms are the same defined as in (4.2). In applications, one can apply a likelihood ratio test, as shown in Section 4.2.3, to select the proper model between (4.2) and (4.4).

Case 2. When the predictor $X \in \mathbb{R}^1$ is univariate, for example, a binary indicator, a matrix regression can be simply formulated as

$$Y = \mu + \beta X + \varepsilon, \tag{4.5}$$

where $\beta \in \mathbb{R}^{r \times m}$. In this case, (4.2) still has a multiplicative coefficient structure but the dimensions of β_1 and β_2 are not necessarily restrictive:

$$Y = \mu + \beta_1 \beta_2^T X + \varepsilon, \tag{4.6}$$

where $\beta_1 \in \mathbb{R}^{r\times q}$ and $\beta_2 \in \mathbb{R}^{q\times m}$, for some $q \leq \min(r, m)$. It is easy to see that (4.6) is the reduced rank form of (4.5). Again, to select the proper model between (4.5) and (4.6), one can apply a likelihood ratio test similar to that shown in Section 4.2.3.

Case 3. Sometimes the goal of a study may be about the association between matrix-valued responses and group effects. Suppose that there are g groups and no other predictors, then the group effects can be represented by g dummy variables: $U_1, U_2, \ldots, U_g$, where

$$U_j = \begin{cases} 1 & \text{if in the } j\text{th group} \\ 0 & \text{otherwise.} \end{cases}, \quad j = 1, \ldots, g.$$

A parallel form of (4.5) is then

$$Y = \mu + \boldsymbol{\alpha}_1 U_1 + \boldsymbol{\alpha}_2 U_2 + \cdots + \boldsymbol{\alpha}_g U_g + \varepsilon, \tag{4.7}$$

where $\boldsymbol{\alpha}_i \in \mathbb{R}^{r\times m}$ represents the group effects relative to the overall mean μ.

4.2.2 Model estimation

In this section, we mainly focus on the model estimation of (4.2). The estimations of the other special formulations can be similarly derived and are much easier.

Without a specific parametric distribution assumed on the random error, one can simply estimate the coefficient parameters in (4.2) by a loss function, for instance, a squared loss function, and estimate the covariance matrices by moment estimators. However, in order to make statistical inference, like in the conventional linear regressions, we assume that the random error ε follows a matrix normal distribution $N_{r\times m}(0, \Sigma_1, \Sigma_2)$. The detailed information regarding matrix normal distribution can be found in Section 2.8.1. We next describe the maximum likelihood estimation of (4.2). Without loss of generality, assume that the predictor X is centered.

Let (Y_i, X_i), $i = 1, \ldots, n$, be an i.i.d random sample with sample size n. Since the predictor is centered, the MLE of μ is $\bar{Y}$. In matrix-variate analysis, explicit MLEs

of β_1, β_2, Σ_1 and Σ_2 are generally not obtainable. We propose a two-step iterative algorithm to get the numerical solution. Let B_1, B_2, S_1 and S_2 be the MLEs of β_1, β_2, Σ_1 and Σ_2 respectively. The log-likelihood function of (4.2) is given in the appendix, Section 4.8. According to the log-likelihood function, given B_2 and S_2, we have

$$\begin{aligned} B_1 &= [\sum_{i=1}^{n}(Y_i - \bar{Y})S_2^{-1}B_2X_i^T](\sum_{i=1}^{n}X_iB_2^TS_2^{-1}B_2X_i^T)^{-1} \\ S_1 &= (nm)^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y} - B_1X_iB_2^T)S_2^{-1}(Y_i - \bar{Y} - B_1X_iB_2^T)^T. \end{aligned} \tag{4.8}$$

Similarly, given B_1 and S_1, B_2 and S_2 satisfy

$$\begin{aligned} B_2 &= [\sum_{i=1}^{n}(Y_i - \bar{Y})^TS_1^{-1}B_1X_i](\sum_{i=1}^{n}X_i^TB_1^TS_1^{-1}B_1X_i)^{-1} \\ S_2 &= (nr)^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y} - B_1X_iB_2^T)^TS_1^{-1}(Y_i - \bar{Y} - B_1X_iB_2^T). \end{aligned} \tag{4.9}$$

Therefore, by randomly initializing $B_2 \in \mathbb{R}^{m\times q}$ and initializing $S_2 = n^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^T(Y_i - \bar{Y})$, one can obtain the standard MLEs B_1, B_2, S_1 and S_2 based on the following iterative algorithm.

1. Given B_2 and S_2, estimate β_1 and Σ_1 by (4.8).

2. Given B_1 and S_1, estimate β_2 and Σ_2 by (4.9).

3. Iterate 1-2 with updated values until the log-likelihood function of (4.2) converges.

Since the estimation procedure is essentially a convex optimization problem, the convergence of the algorithm is guaranteed if the estimators exist at each iteration step. The existence of B_1, B_2, S_1 and S_2 depends on the existence of the inverse matrices S_1^{-1} and S_2^{-1}, which require $p_1 \leq \min(p_2, m)n$, $p_2 \leq \min(p_1, r)n$ and $n > \max(r/m, m/r) - 1$. These conditions are fairly mild when the row and column dimensions of the response and the predictor are not two far from each other.

The asymptotic distribution of the MLEs in (4.2) is studied in Section 4.4.

4.2.3 Goodness of fit

To measure how well the matrix regression (4.2) fits the observed data, one can perform hypothesis testing to test the goodness of fit of the model compared to an alternative model. For example, to compare (4.2) with the multivariate regression $\text{vec}(Y)$ on $\text{vec}(X)$,

$$\text{vec}(Y) = \gamma + \nu \text{vec}(X) + \epsilon, \tag{4.10}$$

where $\gamma \in \mathbb{R}^{rm}$, $\nu \in \mathbb{R}^{rm \times pq}$, and $\epsilon \in \mathbb{R}^{rm}$ follows a multivariate normal distribution $N_{rm}(0, \Sigma)$, one can test the hypothesis:

$$\begin{aligned} &H_0 : \nu = \beta_2 \otimes \beta_1,\ \Sigma = \Sigma_2 \otimes \Sigma_1 \\ &H_a : \nu \text{ and } \Sigma \text{ cannot be decomposed.} \end{aligned} \tag{4.11}$$

Since (4.2) is a "reduced" model of (4.11), the likelihood ratio test can be applied. Let L_1 and L_2 be likelihood functions of (4.2) and (4.11) respectively. The test statistic

$$T = 2 \log L_2 - 2 \log L_1$$

then follows a chi-squared distribution with degree of freedom $df = rmp_1p_2 - (rp_1 + mp_2) + rm(rm+1)/2 - r(r+1)/2 - m(m+1)/2$, which is equal to the number of reduced parameters from (4.11) to (4.2). The goodness of fit of (4.2) compared to other models, such as to the special formulations in Section 4.2.1, can be similarly derived.

4.3 Envelope models for matrix-variate regressions

4.3.1 Introduction to envelopes

In order to build the envelope models for matrix-variate regressions, we first provide a brief review of the envelope method. Enveloping was introduced by Cook et al. (2010). It is a nascent concept and was proposed originally for multivariate linear models to

gain efficiency. Consider the multivariate model

$$Y = \mu + \beta X + \epsilon, \tag{4.12}$$

where the response vector $Y \in \mathbb{R}^r$, the predictor vector $X \in \mathbb{R}^p$ and the random error $\epsilon \sim N(0, \Sigma)$ is independent of X. Let $\mathcal{S} \subseteq \mathbb{R}^r$ be a subspace such that $P_{\mathcal{S}}Y$, the projection of Y onto $\mathcal{S}$, depends on X only, and $Q_{\mathcal{S}}Y$, the projection of Y onto $\mathcal{S}^{\perp}$, is independent of X. In addition, assume that given the predictor X, $P_{\mathcal{S}}Y$ and $Q_{\mathcal{S}}Y$ are uncorrelated. Then $\mathcal{S}$ satisfies: (a) $\mathcal{B} \subseteq \mathcal{S}$, where $\mathcal{B} = \text{span}(\beta)$, and (b) $\Sigma = P_{\mathcal{S}}\Sigma P_{\mathcal{S}} + Q_{\mathcal{S}}\Sigma Q_{\mathcal{S}}$.

According to Cook et al. (2010), condition (b) is sufficient and necessary for $\mathcal{S}$ to be a reducing subspace of Σ. Therefore, $\mathcal{S}$ is a reducing subspace of Σ that contains $\mathcal{B}$. Such reducing subspaces are not unique. The smallest one that contains $\mathcal{B}$ is called the Σ-envelope of $\mathcal{B}$, denoted as $\mathcal{E}_{\Sigma}(\mathcal{B})$, or $\mathcal{E}$ when used as a subscript. The envelope $\mathcal{E}_{\Sigma}(\mathcal{B})$ distinguishes material and immaterial information of Y to the estimation of β, because $P_{\mathcal{E}}Y$ carries all the information available for estimating β and $Q_{\mathcal{E}}Y$ carries no such information. By imposing conditions (a) and (b) on (4.12), the envelope model extracts the material part $P_{\mathcal{E}}Y$ and removes the immaterial part $Q_{\mathcal{E}}Y$ in estimation and thus produces efficiency gains. Cook et al. (2010) demonstrated that the efficiency gains will be massive when the largest eigenvalue of $\text{var}(Q_{\mathcal{E}}Y)$ is substantially larger than that of $\text{var}(P_{\mathcal{E}}Y|X)$.

4.3.2 Envelope formulation

In (4.2), one can potentially gain estimation efficiency by enveloping the coefficients. More specifically, if only a few row and column linear combinations of Y are relevant in estimating β_1 and β_2, then an envelope structure on (4.2) is desirable. Suppose that there exist subspaces $\mathcal{S}_L \subseteq \mathbb{R}^r$ and $\mathcal{S}_R \subseteq \mathbb{R}^m$ so that:

$$(i)\ \mathrm{Q}_{\mathcal{S}_L}Y \mid X \sim \mathrm{Q}_{\mathcal{S}_L}Y \ \text{ and } \ \mathrm{cov}_c(\mathrm{P}_{\mathcal{S}_L}Y, \mathrm{Q}_{\mathcal{S}_L}Y \mid X) = 0;$$

$$(ii)\ Y\mathrm{Q}_{\mathcal{S}_R} \mid X \sim Y\mathrm{Q}_{\mathcal{S}_R} \ \text{ and } \ \mathrm{cov}_r(Y\mathrm{P}_{\mathcal{S}_R}, Y\mathrm{Q}_{\mathcal{S}_R} \mid X) = 0.$$

Then $\mathrm{Q}_{\mathcal{S}_L} Y$ and $Y\mathrm{Q}_{\mathcal{S}_R}$ are immaterial to the estimation of β_1 and β_2 as they are independent of the covariates. This implies that the response projection $\mathrm{P}_{\mathcal{S}_L} Y \mathrm{P}_{\mathcal{S}_R}$ contains full information for estimating the coefficients. Let $\mathcal{B}_1 = \mathrm{span}(\beta_1)$ and $\mathcal{B}_2 = \mathrm{span}(\beta_2)$. The first condition in (i) implies that the distribution of $\mathrm{Q}_{\mathcal{S}_L} Y = \mathrm{Q}_{\mathcal{S}_L}\alpha + \mathrm{Q}_{\mathcal{S}_L}\beta_1 X \beta_2 + \mathrm{Q}_{\mathcal{S}_L}\varepsilon$ does not depend on X, which equivalent to the condition $\mathcal{B}_1 \subseteq \mathcal{S}_L$. The second condition in (i) holds if and only if $\mathrm{P}_{\mathcal{S}_L}\Sigma_1\mathrm{Q}_{\mathcal{S}_L} = 0$, which is equivalent to require $\Sigma_1 = \mathrm{P}_{\mathcal{S}_L}\Sigma_1\mathrm{P}_{\mathcal{S}_L} + \mathrm{Q}_{\mathcal{S}_L}\Sigma_1\mathrm{Q}_{\mathcal{S}_L}$. The similar results hold for (ii). Hence conditions (i) and (ii) can be rewritten as

$$\mathcal{B}_1 \subseteq \mathcal{S}_L, \;\; \Sigma_1 = \mathrm{P}_{\mathcal{S}_L}\Sigma_1\mathrm{P}_{\mathcal{S}_L} + \mathrm{Q}_{\mathcal{S}_L}\Sigma_1\mathrm{Q}_{\mathcal{S}_L}, \tag{4.13}$$

$$\mathcal{B}_2 \subseteq \mathcal{S}_R, \;\; \Sigma_2 = \mathrm{P}_{\mathcal{S}_R}\Sigma_2\mathrm{P}_{\mathcal{S}_R} + \mathrm{Q}_{\mathcal{S}_R}\Sigma_2\mathrm{Q}_{\mathcal{S}_R}. \tag{4.14}$$

These two conditions imply that $\mathcal{S}_L$ is the reducing subspaces of Σ_1 that contains $\mathcal{B}_1$ and that $\mathcal{S}_R$ is the reducing subspaces of Σ_2 that contains $\mathcal{B}_2$. The Σ_1-envelope of $\mathcal{B}_1$ and Σ_2-envelope of $\mathcal{B}_2$ are the smallest reducing subspaces that contain β_1 and β_2 respectively, denoted as $\mathcal{E}_{\Sigma_1}(\mathcal{B}_1)$ and $\mathcal{E}_{\Sigma_2}(\mathcal{B}_2)$, or $\mathcal{E}_1$ and $\mathcal{E}_2$. The minimality guarantees that $P_{\mathcal{E}_1} Y P_{\mathcal{E}_2}$ contains only the material information in estimating β_1 and β_2. The rest, $\mathrm{P}_{\mathcal{E}_1} Y \mathrm{Q}_{\mathcal{E}_2}$, $\mathrm{Q}_{\mathcal{E}_1} Y \mathrm{P}_{\mathcal{E}_2}$ and $\mathrm{Q}_{\mathcal{E}_1} Y \mathrm{Q}_{\mathcal{E}_2}$, are all immaterial.

By distinguishing such material and immaterial information, envelopes can reduce estimation variation. To gain some intuition, assume that $\mathcal{E}_1$, β_2, Σ_1 and Σ_2 are known. We illustrate the variance reduction of $\hat{\beta}_1$, which is the MLE of β_1 under the envelope structure. As $\mathcal{B}_1 \subseteq \mathcal{E}_{\Sigma_1}(\mathcal{B}_1)$, we have $\beta_1 = \mathrm{P}_{\mathcal{E}_1}\beta_1$. Then the envelope MLE $\hat{\beta}_1$ is $\mathrm{P}_{\mathcal{E}_1} B_1$, where B_1 is the standard MLE in (4.2). From Proposition 4.1 in Section 4.3.3, we know $\mathrm{var}[\mathrm{vec}(B_1)] = N_1^{-1} \otimes \Sigma_1 + (N_1^{-1} \otimes \Sigma_1) D (N_1^{-1} \otimes \Sigma_1)$, where $N_1 = \mathrm{E}(X\beta_2^T \Sigma_2^{-1} \beta_2 X^T) > 0$ and $D = N_2[N_3 \otimes \Sigma_2^{-1} - N_2^T (N_1^{-1} \otimes \Sigma_1) N_2]^{-1} N_2^T > 0$. The variance of $\hat{\beta}_1$ is then $\mathrm{var}[\mathrm{vec}(\mathrm{P}_{\mathcal{E}_1} B_1)] = N_1^{-1} \otimes \mathrm{P}_{\mathcal{E}_1}\Sigma_1\mathrm{P}_{\mathcal{E}_1} + (N_1^{-1} \otimes \mathrm{P}_{\mathcal{E}_1}\Sigma_1\mathrm{P}_{\mathcal{E}_1}) D (N_1^{-1} \otimes \mathrm{P}_{\mathcal{E}_1}\Sigma_1\mathrm{P}_{\mathcal{E}_1})$. As $\Sigma_1 =$

$\mathrm{P}_{\mathcal{E}_1}\Sigma_1\mathrm{P}_{\mathcal{E}_1} + \mathrm{Q}_{\mathcal{E}_1}\Sigma_1\mathrm{Q}_{\mathcal{E}_1}$, the envelope MLE reduces variation by

$$\begin{aligned}
&\mathrm{var}[\mathrm{vec}(B_1)] - \mathrm{var}[\mathrm{vec}(\mathrm{P}_{\mathcal{E}_1}B_1)] \\
&= N_1^{-1} \otimes \mathrm{Q}_{\mathcal{E}_1}\Sigma_1\mathrm{Q}_{\mathcal{E}_1} + (N_1^{-1} \otimes \mathrm{Q}_{\mathcal{E}_1}\Sigma_1\mathrm{Q}_{\mathcal{E}_1})D(N_1^{-1} \otimes \mathrm{P}_{\mathcal{E}_1}\Sigma_1\mathrm{P}_{\mathcal{E}_1}) \\
&+ (N_1^{-1} \otimes \mathrm{P}_{\mathcal{E}_1}\Sigma_1\mathrm{P}_{\mathcal{E}_1})D(N_1^{-1} \otimes \mathrm{Q}_{\mathcal{E}_1}\Sigma_1\mathrm{Q}_{\mathcal{E}_1}) + (N_1^{-1} \otimes \mathrm{Q}_{\mathcal{E}_1}\Sigma_1\mathrm{Q}_{\mathcal{E}_1})D(N_1^{-1} \otimes \mathrm{Q}_{\mathcal{E}_1}\Sigma_1\mathrm{Q}_{\mathcal{E}_1}) \\
&\geq 0,
\end{aligned}$$

which is the variance of the immaterial information. Therefore, the efficiency gains from the envelope model can be substantial if the variance of the immaterial part of Y, $\mathrm{var}(\mathrm{Q}_{\mathcal{E}_1}Y)$, is relatively large to $\mathrm{var}(\mathrm{P}_{\mathcal{E}_1}Y)$.

To build the envelope model, let $L \in \mathbb{R}^{r\times u_1}(u_1 \leq r)$ and $R \in \mathbb{R}^{m\times u_2}(u_2 \leq m)$ be semi-orthogonal bases of $\mathcal{E}_{\Sigma_1}(\mathcal{B}_1)$ and $\mathcal{E}_{\Sigma_2}(\mathcal{B}_2)$ respectively. Then there exist coordinate matrices $\eta_1 \in \mathbb{R}^{u_1\times p_1}$ and $\eta_2 \in \mathbb{R}^{u_2\times p_2}$ that satisfy $\beta_1 = L\eta_1$ and $\beta_2 = R\eta_2$. Under (4.13) and (4.14), the envelope model of (4.2) is reparameterized as

$$\begin{aligned}
Y &= \mu + L\eta_1 X\eta_2^T R^T + \varepsilon \\
\Sigma_1 &= \mathrm{P}_L\Sigma_1\mathrm{P}_L + \mathrm{P}_{L_0}\Sigma_1\mathrm{P}_{L_0} = L\Omega_1 L^T + L_0\Omega_{10}L_0^T \\
\Sigma_2 &= \mathrm{P}_R\Sigma_2\mathrm{P}_R + \mathrm{P}_{R_0}\Sigma_2\mathrm{P}_{R_0} = R\Omega_2 R^T + R_0\Omega_{20}R_0^T,
\end{aligned} \tag{4.15}$$

where $\Omega_1 = L^T\Sigma_1 L$, $\Omega_{10} = L_0^T\Sigma_1 L_0$, $\Omega_2 = R^T\Sigma_2 R$ and $\Omega_{20} = R_0^T\Sigma_2 R_0$.As $L\eta_1$ and $R\eta_2$ are overparameterized, the matrices L and R themselves are not identified but their column spaces $\mathrm{span}(L)$ and $\mathrm{span}(R)$ are identified. The two parameter spaces are Grassmannians of dimension u_1 and u_2 in $\mathbb{R}^r$ and $\mathbb{R}^m$ with the numbers of unknown real parameters $u_1(r-u_1)$ and $u_2(m-u_2)$ respectively. Therefore, the total number of real parameters in (4.15) is $rm + u_1p_1 + u_2p_2 + r(r+1)/2 + m(m+1)/2$, which is equal to the sum of the numbers of parameters rm in μ, u_1p_1 in η_1, u_2p_2 in η_2, $u_1(r-u_1)$ in L, $u_2(m-u_2)$ in R, $u_1(u_1+1)/2$ in Ω_1, $(r-u_1)(r-u_1+1)/2$ in Ω_{10}, $u_2(u_2+1)/2$ in Ω_2 and $(m-u_2)(m-u_2+1)/2$ in Ω_{20}, while (4.2) has $rm+rp_1+mp_2+r(r+1)/2+m(m+1)/2$ parameters. The envelope model reduces $p_1(r-u_1)+p_2(m-u_2)$ parameters.

4.3.3 Maximum likelihood estimation

In (4.15), the MLE of μ is still $\bar{Y}$. The remaining parameters have no explicit MLEs. We develop a two-step iteration algorithm to obtain the numerical MLEs. The log-likelihood function of (4.15) is given in Section 4.8.1. Let L, R, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ be the envelope MLEs of L, R, β_1, β_2, Σ_1 and Σ_2 respectively. As shown in Section 4.8.1, given $\hat{\beta}_2$ and $\hat{\Sigma}_2$, we have

$$\hat{\mathcal{E}}_{\Sigma_1}(\mathcal{B}_1) = \underset{\mathcal{T}\in\mathcal{G}(u_1,r)}{\operatorname{argmin}} \log \mid \mathrm{P}_{\mathcal{T}}\hat{\Sigma}_{\mathrm{res}}\mathrm{P}_{\mathcal{T}} \mid_0 + \log \mid \mathrm{Q}_{\mathcal{T}}\hat{\Sigma}_Y\mathrm{Q}_{\mathcal{T}} \mid_0,$$

where $\mathcal{G}(u_1, r)$ means the Grassman manifold of dimension u_1 in $\mathbb{R}^r$ defined in Section 1.3, $\hat{\Sigma}_{\mathrm{res}} = (nm)^{-1}\sum_{i=1}^n(Y_i - \bar{Y} - \tilde{B}_1X_i\hat{\beta}_2^T)\hat{\Sigma}_2^{-1}(Y_i - \bar{Y} - \tilde{B}_1X_i\hat{\beta}_2^T)^T$ and $\hat{\Sigma}_Y = (nm)^{-1}\sum_{i=1}^n(Y_i-\bar{Y})\hat{\Sigma}_2^{-1}(Y_i-\bar{Y})^T$ are the sample column covariance matrix of the residuals and the second sample column moment of Y, with $\tilde{B}_1 = [\sum_{i=1}^n(Y_i-\bar{Y})\hat{\Sigma}_2^{-1}\hat{\beta}_2X_i^T](\sum_{i=1}^n X_i\hat{\beta}_2^T\hat{\Sigma}_2^{-1}\hat{\beta}_2X_i^T)^{-1}$, which is the MLE of β_1 from the standard model (4.2) given $\hat{\beta}_2$ and $\hat{\Sigma}_2$.

Let $\hat{L}$ be the semi-orthogonal basis of $\hat{\mathcal{E}}_{\Sigma_1}(\mathcal{B}_1)$, then $\hat{L}$ can be obtained as

$$\hat{L} = \underset{B}{\operatorname{argmin}} \log \mid B^T\hat{\Sigma}_{\mathrm{res}}B \mid + \log \mid B^T\hat{\Sigma}_Y^{-1}B \mid, \tag{4.16}$$

where argmin_B is taken over all semi-orthogonal matrices $B \in \mathbb{R}^{r\times u_1}$. Based on $\hat{L}$, the envelope MLE of β_1 is $\hat{\beta}_1 = \mathrm{P}_{\hat{L}}\tilde{B}_1$, and the envelope MLE of Σ_1 is $\hat{\Sigma}_1 = \mathrm{P}_{\hat{L}}\hat{\Sigma}_{\mathrm{res}}\mathrm{P}_{\hat{L}} + \mathrm{P}_{\hat{L}_0}\hat{\Sigma}_Y\mathrm{P}_{\hat{L}_0}$.

Similarly, given $\hat{\beta}_1$ and $\hat{\Sigma}_1$, $\hat{\mathcal{E}}_{\Sigma_2}(\mathcal{B}_2)$ satisfies

$$\hat{\mathcal{E}}_{\Sigma_2}(\mathcal{B}_2) = \underset{\mathcal{U}\in\mathcal{G}(u_2,m)}{\operatorname{argmin}} \log \mid \mathrm{P}_{\mathcal{U}}\hat{S}_{\mathrm{res}}\mathrm{P}_{\mathcal{U}} \mid_0 + \log \mid \mathrm{Q}_{\mathcal{U}}\hat{S}_Y\mathrm{Q}_{\mathcal{U}} \mid_0,$$

where $\mathcal{G}(u_2, m)$ is the Grassman manifold of dimension u_2 in $\mathbb{R}^m$, $\hat{S}_{\mathrm{res}} = (nr)^{-1}\sum_{i=1}^n(Y_i^T - \bar{Y}^T - \tilde{B}_2X_i^T\hat{\beta}_1^T)\hat{\Sigma}_1^{-1}(Y_i^T - \bar{Y}^T - \tilde{B}_2X_i^T\hat{\beta}_1^T)^T$ and $\hat{S}_Y = (nr)^{-1}\sum_{i=1}^n(Y_i-\bar{Y})^T\hat{\Sigma}_1^{-1}(Y_i-\bar{Y})$ are the sample row covariance matrix of the residuals and the second sample row moment of Y, with $\tilde{B}_2 = [\sum_{i=1}^n(Y_i - \bar{Y})^T\hat{\Sigma}_1^{-1}\hat{\beta}_1X_i](\sum_{i=1}^n X_i^T\hat{\beta}_1^T\hat{\Sigma}_1^{-1}\hat{\beta}_1X_i)^{-1}$ as the MLE of β_2 from the standard model (4.2) given $\hat{\beta}_1$ and $\hat{\Sigma}_1$.

Let $\hat{R}$ be the semi-orthogonal basis of $\hat{\mathcal{E}}_{\Sigma_2}(\mathcal{B}_2)$, then $\hat{R}$ can be obtained as

$$\hat{R} = \underset{U}{\operatorname{argmin}} \log \mid U^T \hat{S}_{\text{res}} U \mid + \log \mid U^T \hat{S}_Y^{-1} U \mid, \tag{4.17}$$

where argmin_U is taken over all semi-orthogonal matrices $U \in \mathbb{R}^{m \times u_2}$. Correspondingly, $\hat{\beta}_2 = \mathrm{P}_{\hat{R}} \tilde{B}_2$ and $\hat{\Sigma}_2 = \mathrm{P}_{\hat{R}} \hat{S}_{\text{res}} \mathrm{P}_{\hat{R}} + \mathrm{P}_{\hat{R}_0} \hat{S}_Y \mathrm{P}_{\hat{R}_0}$.

Therefore, the envelope parameters can be estimated according to the following two-step iterative algorithm:

1. Initialize β_2 as the standard MLE in (4.2), denoted as β_{20}, and initialize Σ_2 as $\Sigma_{20} = n^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^T (Y_i - \bar{Y})$. Let $\hat{\beta}_2 = \beta_{20}$ and let $\hat{\Sigma}_2 = \Sigma_{20}$.

2. Given $\hat{\beta}_2$ and $\hat{\Sigma}_2$, estimate L by (4.16). Then obtain $\hat{\beta}_1 = \mathrm{P}_{\hat{L}} \tilde{B}_1$, $\hat{\Omega}_1 = \hat{L}^T \hat{\Sigma}_{\text{res}} \hat{L}$, $\hat{\Omega}_{10} = \hat{L}_0^T \hat{\Sigma}_Y \hat{L}_0$ and $\hat{\Sigma}_1 = \mathrm{P}_{\hat{L}} \hat{\Sigma}_{\text{res}} \mathrm{P}_{\hat{L}} + \mathrm{P}_{\hat{L}_0} \hat{\Sigma}_Y \mathrm{P}_{\hat{L}_0}$.

3. Given $\hat{\beta}_1$ and $\hat{\Sigma}_1$, estimate R by (4.17). Then obtain $\hat{\beta}_2 = \mathrm{P}_{\hat{R}} \tilde{B}_2$, $\hat{\Omega}_2 = \hat{R}^T \hat{S}_{\text{res}} \hat{R}$, $\hat{\Omega}_{20} = \hat{R}_0^T \hat{S}_Y \hat{R}_0$ and $\hat{\Sigma}_2 = \mathrm{P}_{\hat{R}} \hat{S}_{\text{res}} \mathrm{P}_{\hat{R}} + \mathrm{P}_{\hat{R}_0} \hat{S}_Y \mathrm{P}_{\hat{R}_0}$.

4. Iterate 2-3 until the log-likelihood function of (4.15) converges.

See Section 4.8.1 for detailed derivations. The envelope estimators of β_1 and β_2 are the row and column projections of their corresponding standard MLEs onto the envelope subspaces $\hat{\mathcal{E}}_{\Sigma_1}(\mathcal{A})$ and $\hat{\mathcal{E}}_{\Sigma_2}(\mathcal{B})$. In addition, $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ are both partitioned into the estimated second moments that are material and immaterial to Y. As shown in Section 4.3.2, these formulations lead to more efficient estimation of the model parameters.

When the response Y and the predictor X are both vector-valued, the envelope model of (4.2) coincides with the envelope model proposed by Cook et al. (2010) for the multivariate regression model. Thus, it is a generalization of the coventional envelope model.

4.3.4 Special cases

One-sided enveloping

We now consider the envelope model for (4.4). Since only β_1 is of the primary interest, we use Σ_1-envelope of β_1 to formulate the envelope model of (4.4) as:

$$\begin{aligned} Y &= \mu + L\eta_1 X + \varepsilon \\ \Sigma_1 &= \Sigma_{\mathcal{E}_1} + \Sigma_{\mathcal{E}_1^\perp} = L\Omega_1 L^T + L_0\Omega_{10}L_0^T, \end{aligned} \tag{4.18}$$

where the parameters are the same defined as in (4.15).

To estimate the model parameters, one can apply the algorithm in Section 4.3.3 with some simplification: (1) as $\beta_2 = I_m$, only one initial value Σ_{20} is needed in Step 1; (2) In Step 3, given $\hat{\beta}_1$ and $\hat{\Sigma}_1$, one only needs to estimate Σ_2 as the standard MLE $\hat{\Sigma}_2 = (nr)^{-1}\sum_{i=1}^n (Y_i - \bar{Y} - \hat{\beta}_1 X_i)^T \hat{\Sigma}_1 (Y_i - \bar{Y} - \hat{\beta}_1 X_i)$, since Σ_2 is not enveloped.

Group effects enveloping

let $Y_{ij} \in \mathbb{R}^{r\times m}$ denote the matrix-valued responses of the j-th subject in the i-th group. Then (4.11) can be equivalently written in the sample version:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{(i)j}, \quad i = 1, \ldots, g,\ j = 1, \ldots, n_i, \tag{4.19}$$

where $\varepsilon_{(i)j} \in \mathbb{R}^{r\times m}$ are iid random errors and n_i is the sample size of the i-th group. The total sample size is $n = \sum_{i=1}^g n_i$. Under the model setting, it is easy to see that $\sum_{i=1}^g n_i\alpha_i = 0$. The MLEs of μ and α_i are $\bar{Y}$ and $\bar{Y}_i - \bar{Y}$ respectively.

Let $\text{span}(\alpha_1, \ldots, \alpha_g) = \mathcal{A}$ and $\text{span}\{(\alpha_1, \ldots, \alpha_g)^T\} = \mathcal{B}$. Assume that there exist subspaces $\mathcal{S}_L \subseteq \mathbb{R}^r$ and $\mathcal{S}_R \subseteq \mathbb{R}^m$ so that:

$$\mathcal{A} \subseteq \mathcal{S}_L,\ \ \Sigma_1 = \Sigma_{\mathcal{S}_L} + \Sigma_{\mathcal{S}_L^\perp}, \tag{4.20}$$

$$\mathcal{B} \subseteq \mathcal{S}_R,\ \ \Sigma_2 = \Sigma_{\mathcal{S}_R} + \Sigma_{\mathcal{S}_R^\perp}, \tag{4.21}$$

where $\Sigma_{\mathcal{S}_L} = \mathrm{P}_{\mathcal{S}_L}\Sigma_1\mathrm{P}_{\mathcal{S}_L}$, $\Sigma_{\mathcal{S}_L^\perp} = \mathrm{Q}_{\mathcal{S}_L}\Sigma_1\mathrm{Q}_{\mathcal{S}_L}$, $\Sigma_{\mathcal{S}_R} = \mathrm{P}_{\mathcal{S}_R}\Sigma_2\mathrm{P}_{\mathcal{S}_R}$ and $\Sigma_{\mathcal{S}_R^\perp} = \mathrm{Q}_{\mathcal{S}_R}\Sigma_2\mathrm{Q}_{\mathcal{S}_R}$. The smallest subspaces that satisfy (4.20) and (4.21) are the Σ_1-envelope of $\mathcal{A}$ and Σ_2-envelope of $\mathcal{B}$, denoted as $\mathcal{E}_{\Sigma_1}(\mathcal{A})$ and $\mathcal{E}_{\Sigma_2}(\mathcal{B})$, or $\mathcal{E}_1$ and $\mathcal{E}_2$. Let $L \in \mathbb{R}^{r\times u_1}(u_1 \le r)$ and $R \in \mathbb{R}^{m\times u_2}(u_2 \le m)$ be semi-orthogonal bases of $\mathcal{E}_{\Sigma_1}(\mathcal{A})$ and $\mathcal{E}_{\Sigma_2}(\mathcal{B})$ respectively. Under (4.20) and (4.21), the envelope model of (4.1) is

$$\begin{aligned} Y_{ij} &= \mu + L\eta_i R^T + \varepsilon_{(i)j}, \;\; i = 1,\ldots,g, j = 1,\ldots,n_i, \\ \Sigma_1 &= \Sigma_{\mathcal{E}_1} + \Sigma_{\mathcal{E}_1^\perp} = L\Omega_1 L^T + L_0\Omega_{10}L_0^T \\ \Sigma_2 &= \Sigma_{\mathcal{E}_2} + \Sigma_{\mathcal{E}_2^\perp} = R\Omega_2 R^T + R_0\Omega_{20}R_0^T, \end{aligned} \tag{4.22}$$

where $\eta_i \in \mathbb{R}^{u_1\times u_2}$, $i = 1,\ldots,g$, are the coordinates of α_i with respect to L and R.

The envelope parameters in (4.22) can be estimated by the following two-step iteration algorithm (see derivations in Section 4.8.1):

1. Initialize R by $\hat{R}$ to be any semi-orthogonal matrice of rank u_2 and initialize Σ_2 as $\hat{\Sigma}_2 = (nr)^{-1}\sum_{i=1}^g\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^T(Y_{ij}-\bar{Y}_i)$.

2. Given $\hat{R}$ and $\hat{\Sigma}_2$, estimate L by (4.28). Accordingly, obtain $\hat{\Omega}_1 = \hat{L}^T\hat{C}_{\text{res}}\hat{L}$, $\hat{\Omega}_{10} = \hat{L}_0^T\hat{C}_Y\hat{L}_0$ and $\hat{\Sigma}_1 = \mathrm{P}_{\hat{L}}\hat{C}_{\text{res}}\mathrm{P}_{\hat{L}} + \mathrm{P}_{\hat{L}_0}\hat{C}_Y\mathrm{P}_{\hat{L}_0}$.

3. Given $\hat{L}$ and $\hat{\Sigma}_1$, estimate R by (4.29). Then obtain $\hat{\Omega}_2 = \hat{R}^T\hat{R}_{\text{res}}\hat{R}$, $\hat{\Omega}_{20} = \hat{R}_0^T\hat{R}_Y\hat{R}_0$ and $\hat{\Sigma}_2 = \mathrm{P}_{\hat{R}}\hat{R}_{\text{res}}\mathrm{P}_{\hat{R}} + \mathrm{P}_{\hat{R}_0}\hat{R}_Y\mathrm{P}_{\hat{R}_0}$.

4. Iterate 2-3 with updated parameters until the log-likelihood function of (4.22) converges. The matrix-variate group effects α_i are then estimated by $\hat{\alpha}_i = \mathrm{P}_{\hat{L}}(\bar{Y}_i - \bar{Y})\mathrm{P}_{\hat{R}}$, $i = 1,\ldots,g$.

4.4 Theoretical properties

In this section, we investigate the asymptotic properties of the model estimators in (4.2) and (4.15). We show that the envelope estimators are asymptotically more efficient than the standard MLEs under the model assumptions. We use 'vec' and 'vech' to denote

the vectorization and half-vectorization operators that satisfy $\text{vech}(B) = C_r\text{vec}(B)$ and $\text{vec}(B) = E_r\text{vech}(B)$ for any symmetric $B \in \mathbb{R}^{r\times r}$ (Henderson and Searle, 1979). The terms C_r and E_r are uniquely defined. Let $K_{pq} \in \mathbb{R}^{pq\times pq}$ be the communication matrix that satisfies $\text{vec}(G^T) = K_{pq}\text{vec}(G)$, for $G \in \mathbb{R}^{p\times q}$. Proposition 4.1 provides the Fisher information matrix of the parameters in (4.2).

Proposition 4.1. *Under (4.2), let* $\Psi = \text{vec}(\Sigma_1)\text{vec}(\Sigma_2)^T$, $N_1 = \text{E}(X\beta_2^T\Sigma_2^{-1}\beta_2X^T)$, $N_2 = \text{E}(X\beta_2^T\Sigma_2^{-1}\otimes\Sigma_1^{-1}\beta_1X)K_{mp_2}$, $N_3 = \text{E}(X^T\beta_1^T\Sigma_1^{-1}\beta_1X)$, $\Pi_1 = \Sigma_1^{-1}\otimes\Sigma_1^{-1}$, *and* $\Pi_2 = \Sigma_2^{-1}\otimes\Sigma_2^{-1}$. *Then the Fisher information of the parameter vector* $\Theta = (\text{vec}(\mu), \text{vec}(\beta_1)^T, \text{vec}(\beta_2)^T, \text{vech}(\Sigma_1)^T, \text{vech}(\Sigma_2)^T)^T$ *is* $J_{\text{reg}} =$

$$\begin{pmatrix} \Sigma_2^{-1}\otimes\Sigma_1^{-1} & 0 & 0 & 0 & 0 \\ 0 & N_1\otimes\Sigma_1^{-1} & N_2 & 0 & 0 \\ 0 & N_2^T & N_3\otimes\Sigma_2^{-1} & 0 & 0 \\ 0 & 0 & 0 & \frac{m}{2}E_r^T\Pi_1E_r & \frac{1}{2}E_r^T\Pi_1\Psi\Pi_2E_m \\ 0 & 0 & 0 & \frac{1}{2}E_m^T\Pi_2\Psi^T\Pi_1E_r & \frac{r}{2}E_m^T\Pi_2E_m \end{pmatrix}.$$

Therefore, under certain regularity conditions, the MLE $\hat{\Theta} = (\text{vec}(\hat{\mu}), \text{vec}(\hat{B}_1)^T, \text{vec}(\hat{B}_2)^T, \text{vech}(\hat{S}_1)^T, \text{vech}(\hat{S}_2)^T)^T$ of Θ, obtained from (4.2), converges to in distribution to a normal random vector with mean zero and covariance matrix J_{reg}^{-1}.

For (4.15), since the envelope matrix regression model is over-parameterized, we apply Proposition 4.1 in Shapiro (1986) to derive the asymptotic distribution of the envelop estimators. Let $\text{vec}(\eta_1)$, $\text{vec}(L)$, $\text{vec}(\eta_2)$, $\text{vec}(R)$, $\text{vech}(\Omega_1)$, $\text{vech}(\Omega_{10})$, $\text{vech}(\Omega_2)$ and $\text{vech}(\Omega_{20})$ in (4.15) be denoted as ζ_1, ζ_2, $\ldots$, ζ_8 respectively, and let ζ be the combined parameter vector $\zeta = (\zeta_1^T, \zeta_2^T, \ldots, \zeta_8^T)^T$. Therefore, $\text{vec}(\beta_1)$, $\text{vec}(\beta_2)$, $\text{vech}(\Sigma_1)$

and $\text{vech}(\Sigma_2)$ can be expressed as functions of ζ:

$$g(\zeta) = \begin{pmatrix} \text{vec}(\beta_1) \\ \text{vec}(\beta_2) \\ \text{vech}(\Sigma_1) \\ \text{vech}(\Sigma_2) \end{pmatrix} = \begin{pmatrix} \text{vec}(L\eta_1) \\ \text{vec}(R\eta_2) \\ \text{vech}(L\Omega_1 L^T + L_0\Omega_{10}L_0^T) \\ \text{vech}(R\Omega_2 R^T + R_0\Omega_{20}R_0^T) \end{pmatrix} = \begin{pmatrix} g_1(\zeta) \\ g_2(\zeta) \\ g_3(\zeta) \\ g_4(\zeta) \end{pmatrix}. \tag{4.23}$$

Based on (4.23) and the Fisher information in Proposition 4.1, the asymptotic distribution of the envelope estimators is given below.

Proposition 4.2. *Under (4.15), $\sqrt{n}(g(\hat{\zeta}) - g(\zeta))$ converges to in distribution to a normal random vector with mean zero and covariance matrix $\Lambda_{\text{reg}} = G(G^T J_{\text{reg}} G)^\dagger G^T$, where G is equal to*

$$\begin{pmatrix} I_{p_1} \otimes L & \eta_1^T \otimes L_0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I_{p_2} \otimes R & \eta_2^T \otimes R_0 & 0 & 0 & 0 & 0 \\ 0 & G_{32} & 0 & 0 & C_r(L \otimes L)E_{u_1} & C_r(L_0 \otimes L_0)E_{r-u_1} & 0 & 0 \\ 0 & 0 & 0 & G_{44} & 0 & 0 & G_{47} & G_{48} \end{pmatrix},$$

with $G_{32} = 2C_r(L\Omega_1 \otimes L_0 - L \otimes L_0\Omega_{10})$, $G_{44} = 2C_m(R\Omega_2 \otimes R_0 - R \otimes R_0\Omega_{20})$, $G_{47} = C_m(R \otimes R)E_{u_2}$ and $G_{48} = C_m(R_0 \otimes R_0)E_{m-u_2}$. Let V_{reg} be the asymptotic variance of $(\text{vec}(\beta_1)^T, \text{vec}(\beta_2)^T, \text{vech}(\Sigma_1)^T, \text{vech}(\Sigma_2)^T)^T$ under (4.2). Then the envelop estimators in (4.15) are asymptotically more efficient than the standard MLEs in (4.2), that is, $V_{\text{reg}} - \Lambda_{\text{reg}} \geq 0$.

Proposition 4.2 indicates that when the envelope conditions (4.13) and (4.14) hold, one can gain potential efficiency in estimation by applying the envelope matrix regression model.

Since the coefficient matrices are often of interest, we next show the asymptotic variances of the coefficient estimators. For notation convenience, we denote the asymptotic variance of $\sqrt{n}(T_n - \theta)$ as $\text{avar}(\sqrt{n}T_n)$. From Proposition 4.2, we see that G contains eight block columns, denoted as $G = (G_{.1}, G_{.2}, \ldots, G_{.8})$, where $G_{.i}$ represents the i-th

block column of G. Let

$$A_1 = \begin{pmatrix} G_{.1}^T J_{\text{reg}} G_{.1} & 0 \\ 0 & G_{.2}^T J_{\text{reg}} G_{.2} \end{pmatrix}, \qquad A_2 = \begin{pmatrix} G_{.1}^T J_{\text{reg}} G_{.3} & G_{.1}^T J_{\text{reg}} G_{.4} \\ G_{.2}^T J_{\text{reg}} G_{.3} & G_{.2}^T J_{\text{reg}} G_{.4} \end{pmatrix},$$

and

$$A_3 = \begin{pmatrix} G_{.3}^T J_{\text{reg}} G_{.3} & 0 \\ 0 & G_{.4}^T J_{\text{reg}} G_{.4} \end{pmatrix}.$$

Proposition 4.3. *Under (4.15),* $\sqrt{n}[\text{vec}(\hat{\beta}_1) - \text{vec}(\beta_1)]$ *and* $\sqrt{n}[\text{vec}(\hat{\beta}_2) - \text{vec}(\beta_2)]$ *converge in distribution to normal random vectors with mean zeros and covariance matrices*

$$\begin{aligned} \text{avar}[\sqrt{n}\text{vec}(\hat{\beta}_1)] &= (I_{p_1} \otimes L, \eta_1^T \otimes L_0)(A_1 - A_2 A_3^{\dagger} A_2^T)^{\dagger}(I_{p_1} \otimes L, \eta_1^T \otimes L_0)^T, \\ \text{avar}[\sqrt{n}\text{vec}(\hat{\beta}_2)] &= (I_{p_2} \otimes R, \eta_2^T \otimes R_0)(A_3 - A_2^T A_1^{\dagger} A_2)^{\dagger}(I_{p_2} \otimes R, \eta_2^T \otimes R_0)^T. \end{aligned}$$

The explicit expressions of A_1, A_2 and A_3 are given in Section 4.8.4.

4.5 Dimension selection

To determine the envelope dimensions u_1 and u_2, one can apply an information criterion, say AIC or BIC, to select optimal dimensions (u_1, u_2) by minimizing the objective function $-2\hat{l}(u_1, u_2) + h(n)t(u_1, u_2)$. Here $\hat{l}(u_1, u_2)$ is the estimated log-likelihood function of the envelope model, $h(n)$ is $\log(n)$ for BIC and is 2 for AIC, and $t(u_1, u_2)$ is the number of parameters in the model. For example, in (4.15),

$$\begin{aligned} \hat{l}(u_1, u_2) = &-\frac{nrm}{2}(1 + \log 2\pi) \\ &- \frac{nm}{2}\log \mid \hat{L}^T \hat{\Sigma}_{\text{res}} \hat{L} \mid - \frac{nm}{2}\log \mid \hat{L}^T \hat{\Sigma}_Y \hat{L} \mid - \frac{nr}{2}\log \mid \hat{R}^T \hat{S}_{\text{res}} \hat{R} \mid - \frac{nr}{2}\log \mid \hat{R}^T \hat{S}_Y \hat{R} \mid, \end{aligned}$$

and

$$t(u_1, u_2) = rm + u_1 p_1 + u_2 p_2 + r(r+1)/2 + m(m+1)/2.$$

The dimensions (u_1, u_2) could also be selected by testing the hypothesis $u_1 = d_1$ $(d_1 < r)$ and $u_2 = d_2$ $(d_2 < m)$ sequentially using the likelihood ratio test statistic $\Lambda(d_1, d_2) =$

$2[l(r,m) - l(d_1, d_2)]$. In (4.15), when $u_1 = r$ and $u_2 = m$, the envelope model reduces to the standard model (4.1). Therefore,

$$l(r,m) = -\frac{nrm}{2}(1 + \log 2\pi) - \frac{nm}{2}\log \mid S_1 \mid -\frac{nr}{2}\log \mid S_2 \mid$$

and $t(r,m) = rm + +rp_1 + mp_2 + r(r+1)/2 + m(m+1)/2$. The degree of freedom associated to $\Lambda(d_1, d_2)$ is then equal to $p_1(r-u_1)+p_2(m-u_2)$. The dimension selection procedure for other envelope models can be derived analogously. In addition, cross validation can also be applied to envelope dimension selection.

4.6 Simulation studies

In this section, we demonstrate the performance of the envelope models numerically and compare it with the corresponding standard models. We first simulated data based on model (4.22) with $g = 4$, $r = 10$, $m = 5$, $u_1 = u_2 = 2$, $\Omega_1 = \sigma^2 I_{u_1}$, $\Omega_{10} = \sigma_0^2 I_{r-u_1}$, $\Omega_2 = \sigma^2 I_{u_2}$ and $\Omega_{20} = \sigma_0^2 I_{m-u_2}$. Here $\sigma^2 = .5$ and $\sigma_0^2 = 2$. The semi-orthogonal matrices L and R were generated by orthogonalizing matrices of independent uniform (0,1) random variables. The elements of η were selected from standard random normal variables. We evaluated the estimation accuracy of the envelope model according to the criterion

$$|| \hat{\alpha}_i - \alpha ||_{\mathrm{F}} . \tag{4.24}$$

The average estimation errors were computed based on the criterion (4.24) for each group over 200 random samples under the standard model and envelope model respectively. The envelope dimensions were chosen as the true dimensions $u_1 = u_2 = 2$. Figure 4.6 shows that the envelope model provides much smaller estimation error than the standard model does because the standard model cannot remove immaterial information from estimation, thus resulting in less accurate results.

Figure 4.6 demonstrates the substantial efficiency gains of the envelope model by comparing the asymptotic, bootstrap and actual standard errors of the first elements

of the group effects between the envelope and standard model, under the true envelope dimension $u_1 = u_2 = 2$. The bootstrap standard errors were obtained based on 200 residual bootstrap samples from one original sample. The actual standard errors were estimated by the sample standard error of $\hat{\alpha}_i$ over the 200 simulated data. These results similarly hold for other elements.

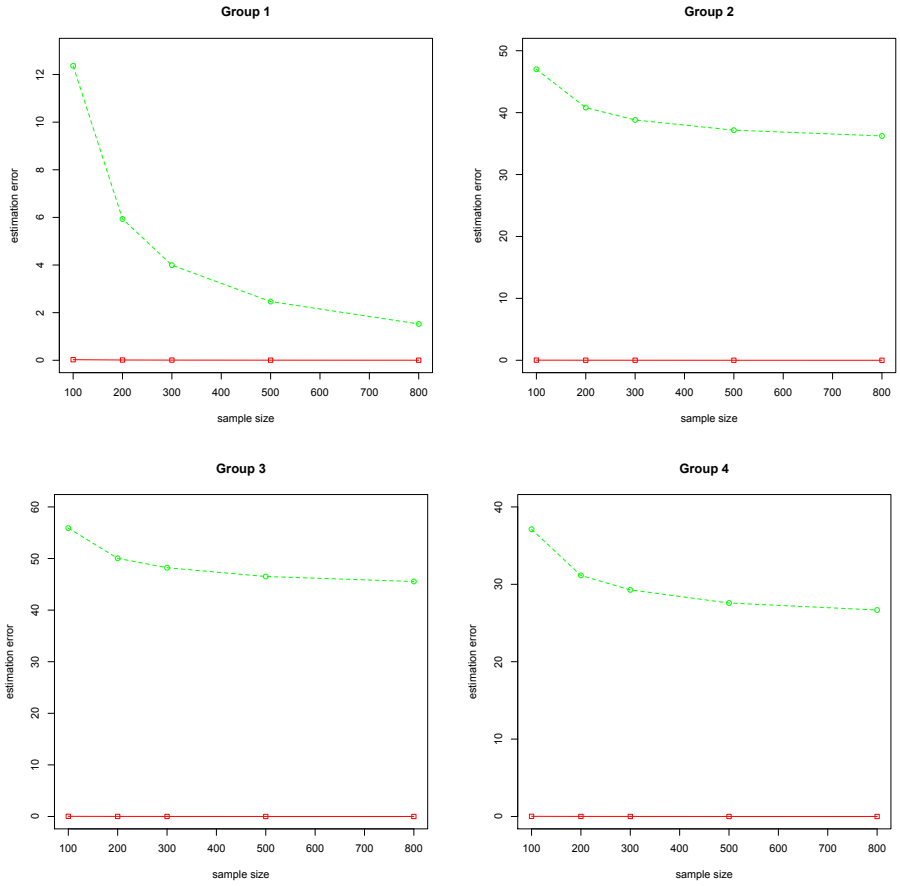


Figure 4.1: The average estimation errors for the four group effects. The solid line indicates the average estimation errors of the envelope models. The dashed line indicates the average estimation errors of the standard models.

It can be seen that the asymptotic standard errors of the envelope model are accurate as both bootstrap standard errors and actual standard errors are close to their corresponding asymptotic standard errors. As expected, the envelope model shows asymptotic efficiency in comparison to the standard model. The ratios of the standard errors between the standard and envelope estimators are in the range of 5.06 and 6.25.

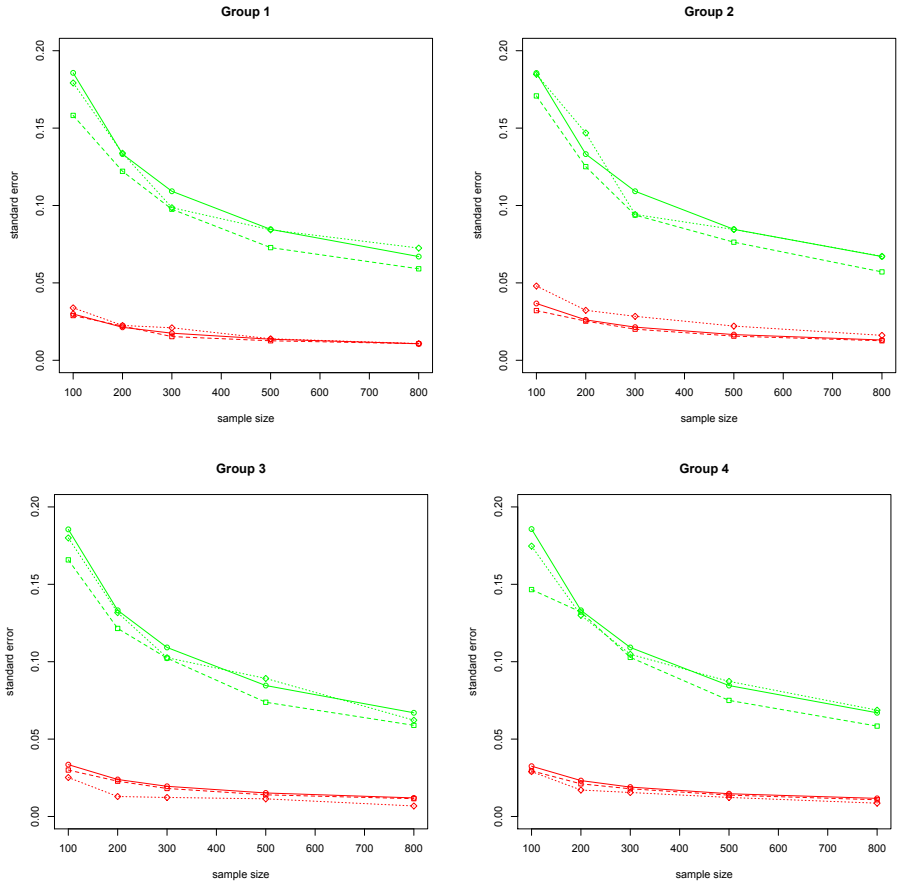


Figure 4.2: The standard errors of the first elements in $\hat{\alpha}_i$ obtained by the envelope model and the standard model. The top three lines indicate the standard errors of the envelope models. The bottom three lines indicate the standard errors of the standard models. The solid lines marks the asymptotic standard errors; the thin dashed lines marks the bootstrap standard errors; and the heavy dashed lines marks the actual standard errors.

4.7 Applications

4.7.1 Multivariate bioassay data

We applied the matrix envelope model to a twin cross-over assay of insulin based on Rabbit blood sugar concentration (Vϕlund, 1980). The design partitioned the animals into four groups of nine rabbits each. Four different treatment combinations were assigned to the rabbits in the four groups. Let K_1 and K_2 denote the low and high dose levels, 0.75 units and 1.5 units, of the standard treatment, and T_1 and T_2 denote the same two dose levels of the test treatment. The treatment assignment is shown in Table 4.1.

Table 4.1: Treatment assignment

Group	1st day	2nd day
1	K_1	T_2
2	K_2	T_1
3	T_1	K_2
4	T_2	K_1

After injection of the insulin dose each day, the blood sugar concentration of each rabbit was measured at 0, 1, 2, 3, 4, and 5 hours. We consider the percentage decreases of the blood sugar concentrations at 1, 2, 3, 4, and 5 hours relative to the initial concentration at 0 hours. Therefore, for each rabbit, the measurements of the percentage falls form a matrix of dimension 2×5. Its rows and columns indicate treatments and hours respectively.

Let $Y_{ij} \in \mathbb{R}^{2\times 5}$ denote the matrix-variate measurements of the j-th rabbit in the i-th group. We first modeled the group effects by (4.22). In this case, the number of groups g is equal to 4. The sample sizes among groups are the same with $n_i = 9$, $i = 1, \ldots, 4$. We applied BIC and LRT to select the envelope dimensions and obtained $u_1 = 1$ and $u_2 = 2$. We then estimated the group effects α_i by both envelope model and standard

model. The corresponding asymptotic standard errors (asySE) are computed for the elements in vec($\hat{\alpha}_i$) and the results for vec($\hat{\alpha}_1$) are summarized in the following Table. The results for other vec($\hat{\alpha}_i$)'s are similar.

Table 4.2: Comparison of the standard errors of vec($\hat{\alpha}_1$) from the envelope and standard fits

					Envelope model					
asySE of vec($\hat{\alpha}_1$)	4.8	4.5	6.3	6.1	8.0	7.7	7.8	7.6	5.8	5.5
					standard model					
asySE of vec($\hat{\alpha}_1$)	7.3	7.3	8.9	8.8	10.6	10.4	11.6	11.5	9.5	9.4

Table 4.2 shows that the envelope model provides more efficient estimators than the standard model does. We further explored the sum of squared prediction error (SSPE) of each model by partition the data into training sets of seven subjects within each group and testing sets of two subjects within each group. We fitted models based on each training set and then computed the SSPE from the corresponding testing set based on the fitted model. As the response is matrix-variate, the prediction error was defined by $||\hat{Y}_{\text{test}} - Y||_F$. The SSPEs over different choices of u_1 and u_2 are shown in Figure 4.3.

It can be seen that the envelope models provide smaller prediction errors over all different choices of u_1 and u_2, except for the case that $u_1 = 2$ and $u_2 = 5$. When $u_1 = 2$ and $u_2 = 5$, there is no reduction in the response, thus the envelope model is the same as the standard model. The smallest prediction error is achieved at the optimal dimension choice $u_1 = 1$ and $u_2 = 2$.

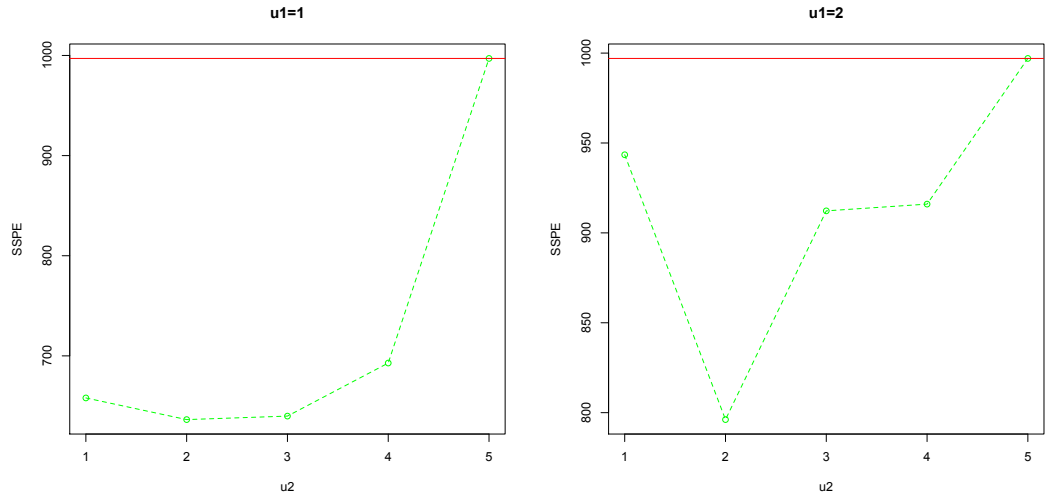


Figure 4.3: The SSPEs over different u_1 and u_2. The solid line indicates the SSPE of the standard model. The dashed line indicates the SSPEs of the envelope models for different choices of u_1 and u_2.

We next modeled the insulin assay data with treatment effects as covariates.

Case I:

Since two treatments (standard and test) at two different dose levels were provided to the four groups of rabbits each day, it can be considered as totally four different treatments were given each day. Therefore, for the twin cross-over design, the covariates of each rabbit forms a matrix $X \in \mathbb{R}^{4\times 2}$, where the rows of X represent the four treatments and the columns of X represent the treatments received in the two days. Each element of X takes values 0 or 1, indicating whether the corresponding treatment is received or not for the corresponding day. Since each rabbit received only one treatment per day, each column of X has only one element equal to one and all others equal to zero. The relationship between the response $Y \in \mathbb{R}^{2\times 5}$ and the predictor $X \in \mathbb{R}^{4\times 2}$ can be modeled as

$$Y^T = \boldsymbol{\beta} X + \varepsilon,$$

where $\boldsymbol{\beta} \in \mathbb{R}^{5\times 4}$ is the coefficient matrix and ε is the random error that follows $N(0, \Sigma_1, \Sigma_2)$. Applying the estimation procedures for (4.4) and (4.18) to the aforementioned training and testing sets, we computed the SSPEs for both envelope and standard models. The results are summarized in Figure 4.4.

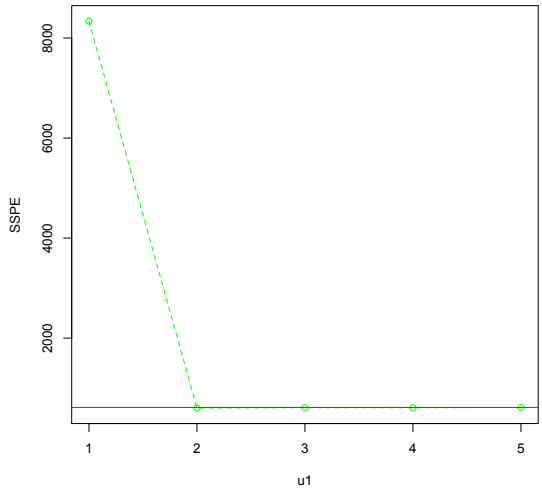


Figure 4.4: The SSPEs of the matrix regression models over different u_1. The solid line represents the SSPE of the standard matrix regression model. The dashed line represents the SSPEs of the envelope matrix regression models for different choices of u_1.

The optimal envelope dimension selected by the LRT is $u_1 = 3$, at which the SSPE for the envelope model is 608.87. The SSPE provided by the standard model is 613.57 that is slightly higher. At the underestimated envelope dimension $u_1 = 1$, the SSPE is 8338.1, which is extremely larger than the others. The relative ratios between the asymptotic standard errors of the standard MLE of vec(β) and the asymptotic standard errors of the envelope MLE of vec(β) range from 1.01 to 1.03. This indicates that the envelope model does not provide much efficiency gains. Therefore, underestimation of the envelope dimension could lose important information in model estimation and

prediction, resulting in high prediction errors.

Case II:

We now consider the dose level as a continuous variable and form the covariates of each rabbit as a matrix $X \in \mathbb{R}^{2\times 2}$, where the two rows of X represent the two treatments: standard treatment and test treatment, respectively, and the columns of X represent the treatments received in the two days. Each element of X is a positive continuous variable with its value indicating the dose level provided for the corresponding treatment.

For example, for the rabbits in group 1, the covariate matrix X is

$$\begin{pmatrix} 0.75 & 0 \\ 0 & 1.5 \end{pmatrix},$$

because the rabbits received the standard treatment at the low dose level 0.75 in day 1, and received the test treatment at the high dose level 1.5 in day 2. Similarly, for the rabbits in groups 2, 3, and 4, the covariate matrices are formed as

$$\begin{pmatrix} 1.5 & 0 \\ 0 & 0.75 \end{pmatrix}, \begin{pmatrix} 0 & 1.5 \\ 0.75 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0.75 \\ 1.5 & 0 \end{pmatrix},$$

respectively. Therefore, the relationship between the response $Y \in \mathbb{R}^{2\times 5}$ and the predictor $X \in \mathbb{R}^{2\times 2}$ can be modeled as

$$Y^T = \mu + \boldsymbol{\beta} X + \varepsilon,$$

where $\mu \in \mathbb{R}^{5\times 2}$ is the intercept matrix, $\boldsymbol{\beta} \in \mathbb{R}^{5\times 2}$ is the coefficient matrix, and ε is the random error that follows $N(0, \Sigma_1, \Sigma_2)$. For estimation convenience, we centered the sample predictors X_i's. Thus the MLE of μ is the overall sample mean $\bar{Y}$. Applying the estimation procedures for (4.4) and (4.18), one can obtain the standard MLE and the envelope MLE of $\boldsymbol{\beta}$. The LRT suggested the envelope dimension $u_1 = 1$. The corresponding asymptotic standard errors of the two estimators are given in Table 4.3.

Table 4.3: Comparison of the standard errors of vec($\hat{\boldsymbol{\beta}}$) from the envelope and standard fits

	Envelope model									
asySE of vec($\hat{\boldsymbol{\beta}}$)	10.7	14.9	19.3	18.7	13.0	10.4	14.7	19.2	18.5	12.7
	standard model									
asySE of vec($\hat{\boldsymbol{\beta}}$)	13.3	16.6	20.2	20.1	15.9	13.3	16.6	20.2	20.1	15.9

It can be seen that the envelope model demonstrates more efficiency gains in estimation compared to the standard model. We further evaluated the SSPEs for both envelope and standard models based on the the aforementioned training and testing sets. The results are summarized in Figure 4.5.

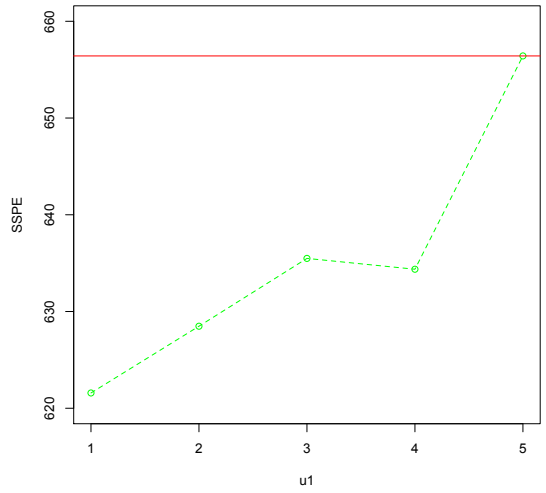


Figure 4.5: The SSPEs of the matrix regression models over different u_1. The solid line represents the SSPE of the standard matrix regression model. The dashed line represents the SSPEs of the envelope matrix regression models for different choices of u_1.

We can see that the envelope regression models outperform the standard model for all choices of the envelope dimension u_1. It reaches the smallest prediction error at the optimal dimension $u_1 = 1$. When $u_1 = 5$, the envelope model coincides with the standard model. Thus, the prediction errors of the two models are identical.

4.7.2 EEG data

The EEG data was briefly introduced in the introduction section. It contains two groups of data: 77 subjects in the alcoholic group and 45 subjects in the control group. Each subject has measurements from the scalp electrical activity, which form a 256×64 matrix. We applied the group effect model (4.19) to explore the influence of the alcoholism on the brain activity. For simple illustration, we prescreened the measurements to a 15 matrix by applying a PCA (principal component analysis) type of dimension reduction and taking $Y_{ij}^* = UY_{ij}V^T$, where the columns of U and V are formed by the leading 15 eigenvectors of $\mathrm{E}_n[(Y_{ij} - \bar{Y}_i)(Y_{ij} - \bar{Y}_i)^T]$ and the leading 15 eigenvectors of $\mathrm{E}_n[(Y_{ij} - \bar{Y}_i)^T(Y_{ij} - \bar{Y}_i)]$ respectively. The illustration of the efficiency of the envelope model is then based on the new responses Y_{ij}^* for $i = 1, 2$, $j = 1, \ldots, n_i$, where $n_1 = 77$ and $n_2 = 45$. By applying the envelope model (4.22) and the standard model (4.19) to Y_{ij}^*, we obtain the standard errors of $\mathrm{vec}(\hat{\alpha}_i)$ from the two models and show the comparison results for the first ten elements in Table 4.4. The likelihood ratio test suggests the envelope dimensions to be $u_1 = u_2 = 1$.

The envelope model shows massive gains in estimating the group effects. The ratios of the standard errors between the standard fit and the envelope fit fall in the range of 5.9 and 169.6, indicating that the envelope model likely suggests more significant influence of the alcoholism on the subjects. Figure 4.6 shows the prediction performance of the two models by computing the SSPEs under different envelope dimensions. The training set was selected to be 20% of the original data. The envelope model outperforms the standard model under all different choices of u_1 and u_2 and it provides the minimum prediction error at the optimal envelope dimensions.

Table 4.4: The standard error (SE) comparison of the first ten elements in $\text{vec}(\hat{\alpha}_i)$ from the envelope and standard fits.

					Envelope model					
$\text{vec}(\hat{\alpha}_1)$	5.6	1.6	4.5	3.4	8.2	6.6	6.9	49.2	24.8	3.4
$\text{vec}(\hat{\alpha}_2)$	9.1	2.6	7.5	5.3	13.5	10.5	10.6	81.0	40.7	5.6
					standard model					
$\text{vec}(\hat{\alpha}_1)$	269.6	263.8	307.5	313.8	323.7	342.3	373.2	368.6	390.0	417.2
$\text{vec}(\hat{\alpha}_2)$	353.1	345.1	402.3	410.5	423.5	447.7	488.2	482.1	510.1	545.7

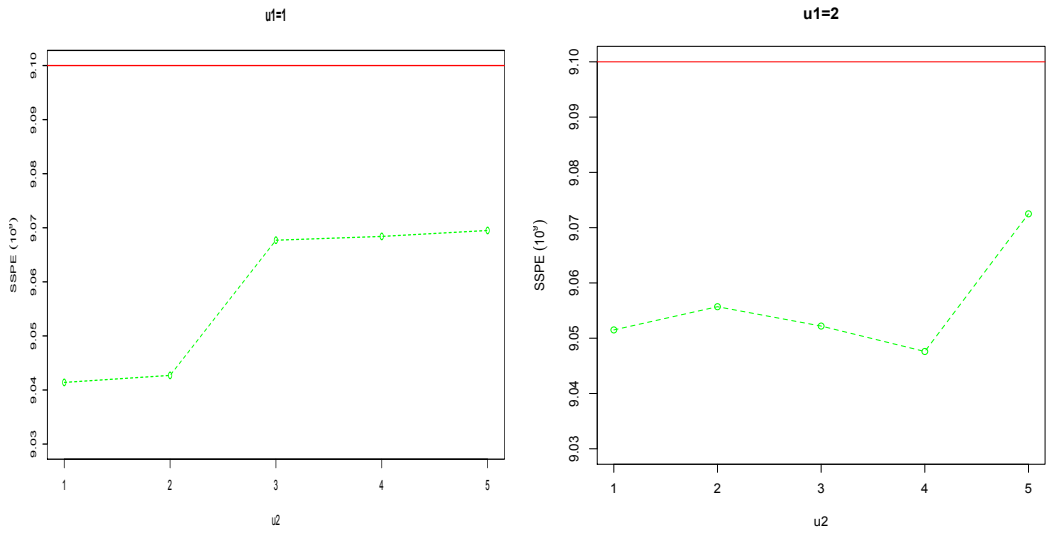


Figure 4.6: The SSPEs over different u_1 and u_2. The solid line indicates the SSPE of the standard model. The dashed line indicates the SSPEs of the envelope models for different choices of u_1 and u_2.

4.8 Appendix

4.8.1 Maximum likelihood estimation

We first derive the MLEs of (4.15). Since the predictors are centered, the MLE $\hat{\mu} = \bar{Y}$. Given β_2 and Σ_2, the log-likelihood function is $l(L, \eta_1, , \Omega_1, \Omega_{10}) =$

$$c - \frac{nr}{2}\log \mid \Sigma_2 \mid -\frac{nm}{2}\log \mid \Omega_{10} \mid -\frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}\{\Sigma_2^{-1}(Y_i - \bar{Y})^T L_0 \Omega_{10}^{-1} L_0^T (Y_i - \bar{Y})\}$$
$$- \frac{nm}{2}\log \mid \Omega_1 \mid -\frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}\{\Sigma_2^{-1}(L^T Y_i - L^T\bar{Y} - \eta_1 X\beta_2^T)\Omega_1^{-1}(L^T Y_i - L^T\bar{Y} - \eta_1 X\beta_2^T)^T\}. \tag{4.25}$$

We now assume that the MLEs $\hat{\beta}_2$ and $\hat{\Sigma}_2$ are known. Then for fixed L, we have $\hat{\eta}_1 = L^T[\sum_{i=1}^n (Y_i - \bar{Y})\hat{\Sigma}_2^{-1}\hat{\beta}_2 X_i^T](\sum_{i=1}^n X_i\hat{\beta}_2^T\hat{\Sigma}_2^{-1}\hat{\beta}_2 X_i^T)^{-1} = L^T B_1$. Substituting $\hat{\eta}_1$ back to (4.25), we find $\hat{\Omega}_1 = L^T\hat{\Sigma}_{\mathrm{res}}L$ and $\hat{\Omega}_{10} = L^T\hat{\Sigma}_Y L$, where $\hat{\Sigma}_{\mathrm{res}}$ and $\hat{\Sigma}_Y$ are defined in Section 4.3.2. These results lead to

$$\hat{l}(L) = c - \frac{nr}{2}\log \mid \hat{\Sigma}_2 \mid -\frac{nm}{2}\log \mid L^T\hat{\Sigma}_Y L \mid -\frac{nm}{2}\log \mid L^T\hat{\Sigma}_{\mathrm{res}}L \mid -\frac{nmr}{2}.$$

Therefore, given $\hat{\beta}_2$ and $\hat{\Sigma}_2$, the MLE of L is $\hat{L} = \underset{\mathrm{span}(B)\in\mathcal{G}(u_1,r)}{\mathrm{argmin}} \log \mid B^T\hat{\Sigma}_{\mathrm{res}}B \mid + \log \mid B^T\hat{\Sigma}_Y^{-1}B \mid$. Correspondingly, $\hat{\beta}_1 = \hat{L}\hat{\eta}_1 = \mathrm{P}_{\hat{L}}B_1$ and $\hat{\Sigma}_1 = \mathrm{P}_{\hat{L}}\hat{\Sigma}_{\mathrm{res}}\mathrm{P}_{\hat{L}} + \mathrm{P}_{\hat{L}_0}\hat{\Sigma}_Y\mathrm{P}_{\hat{L}_0}$. By taking transpose of (4.15), the MLEs of β_2 and Σ_2 can be derived in the same way.

The log-likelihood function of (4.22) is $l(\mu, \eta_i, \Sigma_1, \Sigma_2) =$

$$c - \frac{nm}{2}\log \mid \Sigma_1 \mid -\frac{nr}{2}\log \mid \Sigma_2 \mid -\frac{1}{2}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\mathrm{tr}\{\Sigma_1^{-1}(Y_{ij} - \bar{Y} - L\eta_i R^T)\Sigma_2^{-1}(Y_{ij} - \bar{Y}L\eta_i R^T)^T\}, \tag{4.26}$$

where $c = -\frac{nmr}{2}\log 2\pi$. The last term of (4.26) is equivalent to

$$-\frac{1}{2}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\mathrm{tr}\{[\mathrm{vec}(Y_{ij}) - \mathrm{vec}(\bar{Y}) - (R\otimes L)\mathrm{vec}(\eta_i)](\Sigma_2^{-1}\otimes\Sigma_1^{-1})[\mathrm{vec}(Y_{ij}) - \mathrm{vec}(\bar{Y}) -$$
$$(R\otimes L)\mathrm{vec}(\eta_i)]^T\}.$$

Then for fixed L and R, the MLE of $\text{vec}(\eta_i)$ is $(R^T \otimes L^T)[\text{vec}(\bar{Y}_i - \text{vec}(\bar{Y})]$, or equivalently, $\hat{\eta} = L^T(\bar{Y}_i - \bar{Y})R$. Substituting $\hat{\eta}_i$ back to (4.26), along with the fact that $\Sigma_1 = L\Omega_1 L^T + L_0\Omega_{10}L_0^T$, we have

$$\begin{aligned} l = c &- \frac{nr}{2}\log \mid \Sigma_2 \mid - \frac{nm}{2}\log \mid \Omega_1 \mid - \frac{nm}{2}\log \mid \Omega_{10} \mid \\ &- \frac{1}{2}\sum_{i=1}^{g}\sum_{j=1}^{n_i} \text{tr}\{\Omega_{10}^{-1}L_0^T(Y_{ij} - \bar{Y})\Sigma_2^{-1}(Y_{ij} - \bar{Y})^T L_0\} \\ &- \frac{1}{2}\sum_{i=1}^{g}\sum_{j=1}^{n_i} \text{tr}\{\Omega_1^{-1}L^T[Y_{ij} - \bar{Y} - (\bar{Y}_i - \bar{Y})RR^T]\Sigma_2^{-1}[Y_{ij} - \bar{Y} - (\bar{Y}_i - \bar{Y})RR^T]^T L\}. \end{aligned} \tag{4.27}$$

Let $\hat{C}_{\text{res}} = (nm)^{-1}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}P_{\hat{R}_0} - \bar{Y}_i\text{P}_{\hat{R}})\hat{\Sigma}_2^{-1}(Y_{ij} - \bar{Y}P_{\hat{R}_0} - \bar{Y}_i\text{P}_{\hat{R}})^T$ and $\hat{C}_Y = (nm)^{-1}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})\hat{\Sigma}_2^{-1}(Y_{ij} - \bar{Y})^T$ be the sample column covariance matrix relative to P_R and the second sample column moment. Based on (4.27), assume that the MLEs $\hat{R}$ and $\hat{\Sigma}_2$ are known, then for fixed L, we have $\hat{\Omega}_1 = \frac{1}{nm}\sum_{i=1}^{g}\sum_{j=1}^{n_i} K_{ij} = L^T\hat{C}_{\text{res}}L$ and $\hat{\Omega}_{10} = \frac{1}{nm}\sum_{i=1}^{g}\sum_{j=1}^{n_i} G_{ij} = L_0^T\hat{C}_Y L_0$, where $K_{ij} = L^T(Y_{ij} - \bar{Y}P_{\hat{R}_0} - \bar{Y}_i P_{\hat{R}})\hat{\Sigma}_2^{-1}(Y_{ij} - \bar{Y}P_{\hat{R}_0} - \bar{Y}_i P_{\hat{R}})^T L$ and $G_{ij} = L_0^T(Y_{ij} - \bar{Y})\hat{\Sigma}_2^{-1}(Y_{ij} - \bar{Y})^T L_0$. Substituting $\hat{\Omega}_1$ and $\hat{\Omega}_{10}$ back to (4.27), we obtain

$$\hat{l}(L) = c - \frac{nmr}{2} - \frac{nr}{2}\log \mid \hat{\Sigma}_2 \mid - \frac{nm}{2}\log \mid L^T\hat{C}_{\text{res}}L \mid - \frac{nm}{2}\log \mid L^T\hat{C}_Y^{-1}L \mid - \frac{nm}{2}\log \mid \hat{C}_Y \mid .$$

Therefore, given $\hat{R}$ and $\hat{\Sigma}_2$,

$$\hat{L} = \underset{\text{span}(B)\in\mathcal{G}(u_1,r)}{\text{argmin}} \log \mid B^T\hat{C}_{\text{res}}B \mid + \log \mid B^T\hat{C}_Y^{-1}B \mid \tag{4.28}$$

and $\hat{\Sigma}_1 = \text{P}_{\hat{L}}\hat{C}_{\text{res}}\text{P}_{\hat{L}} + \text{P}_{\hat{L}_0}\hat{C}_Y\text{P}_{\hat{L}_0}$. Define $\hat{R}_{\text{res}} = (nr)^{-1}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij}^T - \bar{Y}^T P_{L_0} - \bar{Y}_i^T\text{P}_L)\Sigma_1^{-1}(Y_{ij}^T - \bar{Y}^T P_{L_0} - \bar{Y}_i^T\text{P}_L)^T$ and $\hat{R}_Y = (nr)^{-1}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^T\Sigma_1^{-1}(Y_{ij} - \bar{Y})$ as the sample row counterparts. The expressions of $\hat{R}$ and $\hat{\Sigma}_2$ can be similarly derived as

$$\hat{R} = \underset{\text{span}(U)\in\mathcal{G}(u_2,m)}{\text{argmin}} \log \mid U^T\hat{R}_{\text{res}}U \mid + \log \mid U^T\hat{R}_Y^{-1}U \mid, \tag{4.29}$$

and $\hat{\Sigma}_2 = \text{P}_{\hat{R}}\hat{R}_{\text{res}}\text{P}_{\hat{R}} + \text{P}_{\hat{R}_0}\hat{R}_Y\text{P}_{\hat{R}_0}$, where $\hat{R}_0$ is the orthogonal complement of $\hat{R}$. given $\hat{L}$ and $\hat{\Sigma}_1$. The derivations are omitted.

4.8.2 Proof of Lemma 4.1

Suppose that there exist two matrices $\gamma_1 \in \mathbb{R}^{r\times p}$ and $\gamma_2 \in \mathbb{R}^{m\times q}$ such that $\beta_2 \otimes \beta_1 = \gamma_2 \otimes \gamma_1$. Let $\beta_{ij}^{(1)}$ denote the ij-th element of β_1 and $\gamma_{ij}^{(1)}$ denote the ij-th element of γ_1. Then $\beta_{ij}^{(1)}\beta_2 = \gamma_{ij}^{(1)}\gamma_2$ for all i and j. This implies that $\gamma_2 = c\beta_2$, where $c = \beta_{ij}^{(1)}/\gamma_{ij}^{(1)}$ is a constant, for all i and j. Since $\beta_{ij}^{(1)}/\gamma_{ij}^{(1)} = c$ for all i and j, we have $\gamma_1 = \beta_1/c$. This proves the lemma.

4.8.3 Proof of Proposition 4.1

Consider the log-likelihood function of (4.2):

$$\begin{aligned} l(\beta_1,\beta_2,\Sigma_1,\Sigma_2) = c - \frac{nm}{2}\log \mid \Sigma_1 \mid - \frac{nr}{2}\log \mid \Sigma_2 \mid \\ - \frac{1}{2}\mathrm{tr}\{\Sigma_1^{-1}(Y-\mu-\beta_1 X\beta_2^T)\Sigma_2^{-1}(Y-\mu-\beta_1 X\beta_2^T)^T\}. \end{aligned}$$

Similar to the proof of Lemma 4.2, let $\theta = (\mathrm{vec}(\beta_1)^T, \mathrm{vec}(\beta_2)^T, \mathrm{vech}(\Sigma_1^{-1})^T, \mathrm{vech}(\Sigma_2^{-1})^T)^T$ $= (\theta_1^T, \theta_2^T, \theta_3^T, \theta_4^T)^T$. The Fisher information I_θ can be partitioned into

$$I_\theta = \begin{pmatrix} I_{\theta_1\theta_1} & I_{\theta_1\theta_2} & I_{\theta_1\theta_3} & I_{\theta_1\theta_4} \\ I_{\theta_2\theta_1} & I_{\theta_2\theta_2} & I_{\theta_2\theta_3} & I_{\theta_2\theta_4} \\ I_{\theta_3\theta_1} & I_{\theta_3\theta_2} & I_{\theta_3\theta_3} & I_{\theta_3\theta_4} \\ I_{\theta_4\theta_1} & I_{\theta_4\theta_2} & I_{\theta_4\theta_3} & I_{\theta_4\theta_4} \end{pmatrix},$$

where $I_{\theta_i\theta_j} = -\mathrm{E}[\partial^2 l/\partial\theta_i\theta_j^T]$. Without loss of generality, we assume $\mu = 0$. Since the last part of $l(\beta_1,\beta_2,\Sigma_1,\Sigma_2)$ is equal to

$$-\frac{1}{2}\{\mathrm{vec}(Y) - (\beta_2 X^T \otimes I_r)\mathrm{vec}(\beta_1)]^T(\Sigma_2^{-1}\otimes\Sigma_1^{-1})[\mathrm{vec}(Y) - (\beta_2 X_i^T \otimes I_r)\mathrm{vec}(\beta_1)\},$$

we have $\partial^2 l/\partial\theta_1\theta_1^T = -X\beta_2^T\Sigma_2^{-1}\beta_2 X^T \otimes \Sigma_1^{-1}$,

$$\begin{aligned} &\partial^2 l/\partial\theta_1\theta_2^T \\ &= \partial(X\otimes\Sigma_1^{-1}Y\Sigma_2^{-1})\mathrm{vec}(\beta_2)/\partial\mathrm{vec}(\beta_2)^T - \partial(X\otimes\Sigma_1^{-1}\beta_1 X)\mathrm{vec}(\beta_2^T\Sigma_2^{-1}\beta_2)/\partial\mathrm{vec}(\beta_2)^T \\ &= X\otimes\Sigma_1^{-1}Y\Sigma_2^{-1} - (X\otimes\Sigma_1^{-1}\beta_1 X)[I_{p_2}\otimes\beta_2^T\Sigma_2^{-1} + (\beta_2^T\Sigma_2^{-1}\otimes I_m)K_{mp_2}] \\ &= X\otimes\Sigma_1^{-1}\varepsilon\Sigma_2^{-1} - (X\beta_2^T\Sigma_2^{-1}\otimes\Sigma_1^{-1}\beta_1 X)K_{mp_2}, \end{aligned}$$

$\partial^2 l/\partial\theta_1\theta_3^T = \partial \text{vec}(\Sigma_1^{-1}\varepsilon\Sigma_2^{-1}\beta_2 X^T)/\partial \text{vech}(\Sigma_1^{-1})^T = (X\beta_2^T\Sigma_2^{-1}\varepsilon^T \otimes I_r)E_r$ and $\partial^2 l/\partial\theta_1\theta_4^T = \partial \text{vec}(\Sigma_1^{-1}\varepsilon\Sigma_2^{-1}\beta_2 X^T)/\partial \text{vech}(\Sigma_2^{-1})^T = (X\beta_2^T \otimes \Sigma_1^{-1}\varepsilon)E_m$. As X is centered and it is independent of ε, it follows that $I_{\theta_1\theta_1} = N_1 \otimes \Sigma_1^{-1}$, $I_{\theta_1\theta_2} = N_2$ and $I_{\theta_1\theta_3}$ and $I_{\theta_1\theta_4}$ are both zero matrices. Similarly, we can show that $I_{\theta_2\theta_2} = N_3 \otimes \Sigma_2^{-1}$ and $I_{\theta_2\theta_3}$ and $I_{\theta_2\theta_4}$ are both zero matrices. The derivations of $I_{\theta_3\theta_3}$, $I_{\theta_3\theta_4}$ and $I_{\theta_4\theta_4}$ can be found in Lemma 7.3 of Pan and Fang (2000). We omit them here.

Because $g(\phi)$ is a function of θ, denoted as $f(\theta)$, by the delta method, we have $J_{\text{reg}} = \{[f'(\theta)]^T I_\theta^{-1} f'(\theta)\}^{-1}$, where $f'(\theta) = \text{diag}[I_{rp_1}, I_{mp_2}, -E_r^T(\Sigma_1 \otimes \Sigma_1)C_r^T, -E_m^T(\Sigma_2 \otimes \Sigma_2)C_m^T]$. This gives the result in Proposition 4.1 after some matrix multiplications.

4.8.4 Proof of Propositions 4.2 and 4.3

As (4.15) is over-parameterized, according to Proposition 4.1 in Shapiro (1986), $\sqrt{n}(g(\hat{\theta}) - g(\theta))$ converges to in distribution to a normal random vector with mean zero and covariance matrix $\Lambda_{\text{reg}} = G_1(G_1^T J_{\text{reg}} G_1)^\dagger G_1^T$, where $G_1 = (\partial g_i/\partial\theta_j^T)_{i,j}$ is the gradient matrix that is equal to

$$\begin{pmatrix} I_{p_1} \otimes L & \eta_1^T \otimes I_r & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I_{p_2} \otimes R & \eta_2^T \otimes I_m & 0 & 0 & 0 & 0 \\ 0 & G_{1,32} & 0 & 0 & C_r(L \otimes L)E_{u_1} & C_r(L_0 \otimes L_0)E_{r-u_1} & 0 & 0 \\ 0 & 0 & 0 & G_{1,44} & 0 & 0 & G_{47} & G_{48} \end{pmatrix},$$

where $G_{1,32} = 2C_r(L\Omega_1 \otimes I_r - L \otimes L_0\Omega_{10}L_0^T)$, $G_{1,44} = 2C_m(R\Omega_2 \otimes I_m - R \otimes R_0\Omega_{20}R_0^T)$, G_{47} and G_{48} are given in Proposition 4.2. Since G is a non-singular column transformation matrix of G_1 and Λ_{reg} depends on G_1 only through its column space span(G_1), then $\Lambda_{\text{reg}} = G(G^T J_{\text{reg}} G)^\dagger G^T$.

To show the asymptotic efficiency of the envelope MLEs, consider $V_{\text{reg}} - \Lambda_{\text{reg}} = J_{\text{reg}}^{-1} - G(G^T J_{\text{reg}} G)^\dagger G^T = J_{\text{reg}}^{-1/2} Q_{J_{\text{reg}}^{1/2} G} J_{\text{reg}}^{-1/2}$, where $Q_{J_{\text{reg}}^{1/2} G}$ is the projection matrix onto the orthogonal compliment of span$(J_{\text{reg}}^{1/2} G)$. Therefore, $V_{\text{reg}} - \Lambda_{\text{reg}} \geq 0$. We complete the proof.

Proof of Proposition 4.3. Let $G_{.1} = (I_{p_1} \otimes L^T, 0^T, 0^T, 0^T)^T$, $G_{.2} = (\eta_1 \otimes L_0^T, 0^T, G_{32}^T, 0^T)^T$, $G_{.3} = (0^T, I_{p_2} \otimes R^T, 0^T, 0^T)^T$ and $G_{.4} = (0^T, \eta_2 \otimes R_0^T, 0^T, G_{44}^T)^T$ be the first four block columns of G, and $G^* = (G_{.5}, G_{.6}, G_{.7}, G_{.8})$ be the matrix containing the rest block columns of G. After some matrix operations, it can be shown that

$$G^T J_{\text{reg}} G = \begin{pmatrix} K^* & 0 \\ 0 & {G^*}^T J_{\text{reg}} G^* \end{pmatrix},$$

where

$$K^* = (G_{.1}, G_{.2}, G_{.3}, G_{.4})^T J_{\text{reg}} (G_{.1}, G_{.2}, G_{.3}, G_{.4}) = \begin{pmatrix} A_1 & A_2 \\ A_2^T & A_3 \end{pmatrix}.$$

Let $G_{1.} = (I_{p_1} \otimes L, \eta_1^T \otimes L_0, 0, 0, 0, 0, 0, 0)$ and $G_{2.} = (0, 0, I_{p_2} \otimes R, \eta_2^T \otimes R_0, 0, 0, 0, 0)$ be the first two block rows of G. Then

$$\begin{aligned}
&\text{avar}[\sqrt{n}\text{vec}(\hat{\beta}_1)] \\
&= G_{1.}(G^T J_{\text{reg}} G)^\dagger G_{1.}^T = (I_{p_1} \otimes L, \eta_1^T \otimes L_0, 0, 0) {K^*}^T (I_{p_1} \otimes L, \eta_1^T \otimes L_0, 0, 0)^T \\
&= (I_{p_1} \otimes L, \eta_1^T \otimes L_0)(A_1 - A_2 A_3^\dagger A_2^T)^\dagger (I_{p_1} \otimes L, \eta_1^T \otimes L_0)^T,
\end{aligned}$$

$$\begin{aligned}
&\text{avar}[\sqrt{n}\text{vec}(\hat{\beta}_2)] \\
&= G_{2.}(G^T J_{\text{reg}} G)^\dagger G_{2.}^T = (0, 0, I_{p_2} \otimes R, \eta_2^T \otimes R_0) {K^*}^T (0, 0, I_{p_2} \otimes R, \eta_2^T \otimes R_0)^T \\
&= (I_{p_2} \otimes R, \eta_2^T \otimes R_0)(A_3 - A_2^T A_1^\dagger A_2)^\dagger (I_{p_2} \otimes R, \eta_2^T \otimes R_0)^T.
\end{aligned}$$

The expressions of A_1, A_2 and A_3 are given by: $G_{.1}^T J_{\text{reg}} G_{.1} = N_1 \otimes \Omega_1^{-1}$, $G_{.2}^T J_{\text{reg}} G_{.2} = \eta_1 N_1 \eta_1^T \otimes \Omega_{10}^{-1} + m(\Omega_1 \otimes \Omega_{10}^{-1} + \Omega_1^{-1} \otimes \Omega_{10} - 2I_{u_1} \otimes I_{r-u_1})$, $G_{.3}^T J_{\text{reg}} G_{.3} = N_3 \otimes \Omega_2^{-1}$, $G_{.4}^T J_{\text{reg}} G_{.4} = \eta_2 N_3 \eta_2^T \otimes \Omega_{20}^{-1} + r(\Omega_2 \otimes \Omega_{20}^{-1} + \Omega_2^{-1} \otimes \Omega_{20} - 2I_{u_2} \otimes I_{m-u_2})$, $G_{.1}^T J_{\text{reg}} G_{.3} = (I_{p_1} \otimes L^T) N_2 (I_{p_2} \otimes R)$, $G_{.1}^T J_{\text{reg}} G_{.4} = (I_{p_1} \otimes L^T) N_2 (\eta_2^T \otimes R_0)$, $G_{.2}^T J_{\text{reg}} G_{.3} = (\eta_1 \otimes L_0^T) N_2 (I_{p_2} \otimes R)$ and $G_{.2}^T J_{\text{reg}} G_{.4} = (\eta_1 \otimes L_0^T) N_2 (\eta_2^T \otimes R_0) + 2(L^T \otimes \Omega_{10}^{-1} L_0^T - \Omega_1^{-1} L^T \otimes L_0^T) \Psi (R \otimes R_0 \Omega_{20}^{-1} - R\Omega_2^{-1} \otimes R_0) K_{u_2(m-u_2)}$.

4.8.5 Asymptotic properties of (4.22)

For simplicity, let $\text{vec}(\eta_i)$, $\text{vec}(L)$, $\text{vec}(R)$, $\text{vech}(\Omega_1)$, $\text{vech}(\Omega_{10})$, $\text{vech}(\Omega_2)$ and $\text{vech}(\Omega_{20})$ be denoted as ϕ_{1i}, ϕ_2, $\ldots$, ϕ_7 respectively, for $i = 1, \ldots, g$. Let ϕ_i be the combined

parameter vector $\phi_i = (\phi_{1i}^T, \phi_2^T, \ldots, \phi_7^T)^T$. Therefore, $\mathrm{vec}(\alpha_i)$, $\mathrm{vech}(\Sigma_1)$ and $\mathrm{vech}(\Sigma_2)$ can be expressed as functions of ϕ_i

$$h(\phi_i) = \begin{pmatrix} \mathrm{vec}(\alpha_i) \\ \mathrm{vech}(\Sigma_1) \\ \mathrm{vech}(\Sigma_2) \end{pmatrix} = \begin{pmatrix} \mathrm{vec}(L\eta_i R^T) \\ \mathrm{vech}(L\Omega_1 L^T + L_0\Omega_{10}L_0^T) \\ \mathrm{vech}(R\Omega_2 R^T + R_0\Omega_{20}R_0^T) \end{pmatrix} = \begin{pmatrix} h_1(\phi_i) \\ h_2(\phi_i) \\ h_3(\phi_i) \end{pmatrix}. \tag{4.30}$$

Because of the overparameterization of the second equation in (4.30), we apply Proposition 4.1 in Shapiro (1986) to derive the asymptotic distribution for the envelop estimators. Assume that $n_i/n = c_i$ are fixed for all n, $i = 1, \ldots, g-1$.

Lemma 4.2. *Suppose that* $Y_{(i)} \sim N_{r\times m}(\mu + \alpha_i, \Sigma_1, \Sigma_2)$, $i = 1, \ldots, g-1$. *Then the Fisher information of the parameter vector* $(\mathrm{vec}(\alpha_i)^T, \mathrm{vech}(\Sigma_1)^T, \mathrm{vech}(\Sigma_2)^T)^T$ *is*

$$J_i = \begin{pmatrix} c_i\Sigma_2^{-1}\otimes\Sigma_1^{-1} & 0 & 0 \\ 0 & \frac{m}{2}E_r^T(\Sigma_1^{-1}\otimes\Sigma_1^{-1})E_r & \frac{1}{2}E_r^T\mathrm{vec}(\Sigma_1^{-1})\mathrm{vec}(\Sigma_2^{-1})^T E_m \\ 0 & \frac{1}{2}E_m^T\mathrm{vec}(\Sigma_2^{-1})\mathrm{vec}(\Sigma_1^{-1})^T E_r & \frac{r}{2}E_m^T(\Sigma_2^{-1}\otimes\Sigma_2^{-1})E_m \end{pmatrix}.$$

Proposition 4.4. *Under (4.22),* $\sqrt{n}(h(\hat{\phi}_i) - h(\phi_i))$ *converges in distribution to a normal random vector with mean zero and covariance matrix* $\Lambda_i = H_i(H_i^T J_i H_i)^\dagger H_i^T$, $i = 1, \ldots, g-1$, *where* H_i *is given by*

$$\begin{pmatrix} R\otimes L & R\eta_i^T\otimes L_0 & R_0\otimes L\eta_i & 0 & 0 & 0 & 0 \\ 0 & H_{i,22} & 0 & C_r(L\otimes L)E_{u_1} & C_r(L_0\otimes L_0)E_{r-u_1} & 0 & 0 \\ 0 & 0 & H_{i,33} & 0 & 0 & H_{i,36} & H_{i,37} \end{pmatrix},$$

with $H_{i,22} = 2C_r(L\Omega_1\otimes L_0 - L\otimes L_0\Omega_{10})$, $H_{i,33} = 2C_m(R\Omega_2\otimes R_0 - R\otimes R_0\Omega_{20})K_{u_2(m-u_2)}$, $H_{i,36} = C_m(R\otimes R)E_{u_2}$ *and* $H_{i,37} = C_m(R_0\otimes R_0)E_{m-u_2}$. *Let* $V_i = J_i^{-1}$ *be the asymptotic variance of* $(\mathrm{vec}(\alpha_i)^T, \mathrm{vech}(\Sigma_1)^T, \mathrm{vech}(\Sigma_2)^T)^T$ *under (4.19). Then the envelop estimators are asymptotically more efficient than the standard MLEs, that is,* $V_i - \Lambda_i \geq 0$.

Proposition 4.4 shows that if conditions (4.20) and (4.21) hold, the envelope model always tends to give less variant estimators compared to the standard model. Therefore, one will benefit from using the envelope model. In most cases the group effects α_i, $i = 1, \ldots, g$, are of particular interest. We next provide the limiting distribution for $\hat{\alpha}_i$. Let $H_{i,.2}$ and $H_{i,.3}$ be the second and third columns of H_i. Let $\hat{\alpha}_{i(L,R)}$ $\hat{\alpha}_{i(\eta_i,R)}$, $\hat{\alpha}_{i(L,\eta_i)}$ be the maximum likelihood estimators of α_i when L and R, η_i and R, and L and η_i are known, respectively.

Proposition 4.5. *Under (4.22),* $\sqrt{n}[\mathrm{vec}(\hat{\alpha}_i) - \mathrm{vec}(\alpha_i)]$ *converges in distribution to a normal random vector with mean zero and covariance matrix*

$$\begin{aligned} &\mathrm{avar}[\sqrt{n}\mathrm{vec}(\hat{\alpha}_i)] = R\Omega_2 R^T \otimes L\Omega_1 L^T + \\ &(R\eta_i^T \otimes L_0)(H_2^T J H_2)^\dagger(\eta_i R^T \otimes L_0^T) + (R_0 \otimes L\eta_i)(H_3^T J H_3)^\dagger(R_0^T \otimes \eta_i^T L^T) \\ &= \mathrm{avar}[\sqrt{n}\mathrm{vec}(\hat{\alpha}_{i(L,R)})] + \mathrm{avar}[\sqrt{n}\mathrm{vec}(Q_{\mathcal{E}_1}\hat{\alpha}_{i(\eta_i,R)})] + \mathrm{avar}[\sqrt{n}\mathrm{vec}(\hat{\alpha}_{i(L,\eta_i)}Q_{\mathcal{E}_2})], \end{aligned} \tag{4.31}$$

where $H_2^T J H_2 = \eta_i \Omega_2^{-1} \eta_i^T \otimes \Omega_{10}^{-1} + m(\Omega_1 \otimes \Omega_{10}^{-1} + \Omega_1^{-1} \otimes \Omega_{10} - 2I_{u_1} \otimes I_{r-u_1})$ *and* $H_3^T J H_3 = \Omega_{20}^{-1} \otimes \eta_i^T \Omega_1^{-1} \eta_i + r K_{u_2(m-u_2)}^T (\Omega_2 \otimes \Omega_{20}^{-1} + \Omega_2^{-1} \otimes \Omega_{20} - 2I_{u_2} \otimes I_{m-u_2}) K_{u_2(m-u_2)}$.

In the last equation of (4.31), the asymptotic variance of $\sqrt{n}\mathrm{vec}(\hat{\alpha}_i)$ can be partitioned into three additive parts and each of them corresponds to asymptotic variance of the MLE when two of L, η_i, R are known. The terms $Q_{\mathcal{E}_1}$ and $Q_{\mathcal{E}_2}$ are the projection matrices onto the orthogonal subspaces of span(L) and span(R). They serve to orthogonalize their corresponding random vectors and make the asymptotic variance additive.

4.8.6 Proof of Lemma 4.2

Let $\theta_i = (\text{vec}(\alpha_i)^T, \text{vech}(\Sigma_1^{-1})^T, \text{vech}(\Sigma_2^{-1})^T)^T$, $i = 1, \ldots, g-1$. By Lemma 7.3 in Pan and Fang (2000), the Fisher information of θ_i is

$$I_{\theta_i} = \begin{pmatrix} c_i\Sigma_2^{-1} \times \Sigma_1^{-1} & 0 & 0 \\ 0 & \frac{m}{2}E_r^T(\Sigma_1 \otimes \Sigma_1)E_r & \frac{1}{2}E_r^T\text{vec}(\Sigma_1)\text{vec}(\Sigma_2)^T E_m \\ 0 & \frac{1}{2}E_m^T\text{vec}(\Sigma_2)\text{vec}(\Sigma_1)^T E_r & \frac{r}{2}E_m^T(\Sigma_2 \otimes \Sigma_2)E_m \end{pmatrix}.$$

The term $h(\phi_i)$ is a function of θ_i, denoted as $g(\theta_i)$. The Fisher information of $g(\theta_i)$ is $J_i = I_{g(\theta_i)} = \{\text{avar}[\sqrt{n}g(\hat{\theta}_i)]\}^{-1} = \{[g'(\theta_i)]^T I_{\theta_i}^{-1} g'(\theta_i)\}^{-1}$. It can shown that $g'(\theta_i) = \text{diag}[I_{rm}, -E_r^T(\Sigma_1 \otimes \Sigma_1)C_r^T, -E_m^T(\Sigma_2 \otimes \Sigma_2)C_m^T]$. Then $J_i = [g'(\theta_i)]^{-1} I_{\theta_i} \{[g'(\theta_i)]^{-1}\}^T$. Since $[E_r^T(\Sigma_1 \otimes \Sigma_1)C_r^T][E_r^T(\Sigma_1^{-1} \otimes \Sigma_1^{-1})C_r^T] = I$, it follows that $[E_r^T(\Sigma_1 \otimes \Sigma_1)C_r^T]^{-1} = E_r^T(\Sigma_1^{-1} \otimes \Sigma_1^{-1})C_r^T$. Correspondingly,

$$[g'(\theta_i)]^{-1} = \begin{pmatrix} I_{rm} & 0 & 0 \\ 0 & -E_r^T(\Sigma_1^{-1} \otimes \Sigma_1^{-1})C_r^T & 0 \\ 0 & 0 & -E_m^T(\Sigma_2^{-1} \otimes \Sigma_2^{-1})C_m^T \end{pmatrix}.$$

Therefore, by multiplying $[g'(\theta_i)]^{-1}$ with I_{θ_i} and $\{[g'(\theta_i)]^{-1}\}^T$, we obtain the fomular of J in Lemma 4.2.

4.8.7 Proof of Proposition 4.4

We first derive the expression of H. Since (4.22) is an over-parameterized model, by Proposition 4.1 in Shapiro (1986), we have $\sqrt{n}(h(\hat{\phi}_i) - h(\phi_i))$ converges in distribution to a normal random vector with mean zero and covariance matrix $\Lambda_i = W_i(W_i^T J_i W_i)^\dagger W_i^T$, where $W_i = (\partial h_k / \partial \phi_{il}^T)_{k,l}$ is the gradient matrix and ϕ_{il} denotes the l-th element in ϕ_i. According to (4.30), it can be shown that W_i is equal to

$$\begin{pmatrix} R \otimes L & R\eta_i^T \otimes I_r & (I_m \otimes L\eta_i)K_{mu_2} & 0 & 0 & 0 & 0 \\ 0 & W_{i,22} & 0 & W_{i,24} & W_{i,25} & 0 & 0 \\ 0 & 0 & W_{i,33} & 0 & 0 & W_{i,36} & W_{i,37} \end{pmatrix},$$

where $W_{i,22} = 2C_r(L\Omega_1 \otimes I_r - L \otimes L_0\Omega_{10}L_0^T)$, $W_{i,33} = 2C_m(R\Omega_2 \otimes I_m - R \otimes R_0\Omega_{20}R_0^T)$, $W_{i,24} = C_r(L \otimes L)E_{u_1}$, $W_{i,25} = C_r(L_0 \otimes L_0)E_{r-u_1}$, $W_{i,36} = C_m(R \otimes R)E_{u_2}$ and $W_{i,37} = C_m(R_0 \otimes R_0)E_{m-u_2}$.

Since Λ_i is invariant in W_i with any full rank column transformations, then $\Lambda_i = H_i(H_i^T J_i H_i)^\dagger H_i^T$, where $H_i = W_i D_i^{-1}$ has the same column space as W_i, with $D_i =$

$$\begin{pmatrix} I_{rm} & \eta_i^T \otimes L^T & (R^T \otimes \eta_i)K_{mu_2} & 0 & 0 & 0 & 0 \\ 0 & I_{u_1} \otimes L_0^T & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & (R_0^T \otimes I_{u_2})K_{mu_2} & 0 & 0 & 0 & 0 \\ 0 & 2C_{u_1}(\Omega_1 \otimes L^T) & 0 & I_{\frac{u_1(u_1+1)}{2}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I_{\frac{(r-u_1)(r-u_1+1)}{2}} & 0 & 0 \\ 0 & 0 & 2C_{u_2}(\Omega_2 \otimes R^T) & 0 & 0 & I_{\frac{u_2(u_2+1)}{2}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & I_k \end{pmatrix},$$

where $k = (m-u_2)(m-u_2+1)/2$. The matrix multiplication $W_i D_i^{-1}$ gives the expression of H_i in Proposition 4.4.

We next show the efficiency of the envelope estimation. Note that $V_i - \Lambda_i = J_i^{-1} - H_i(H_i^T J_i H_i)^\dagger H_i^T = J_i^{-\frac{1}{2}}(I - \mathrm{P}_{J_i^{\frac{1}{2}}H_i})J_i^{-\frac{1}{2}} = J_i^{-\frac{1}{2}} Q_{J_i^{\frac{1}{2}}H_i} J_i^{-\frac{1}{2}}$. As $Q_{J_i^{\frac{1}{2}}H_i}$ is the projection matrix onto the orthogonal compliment of $\mathrm{span}(J_i^{\frac{1}{2}}H_i)$, it is positive semi-definite. Hence $V_i - \Lambda_i \geq 0$.

4.8.8 Proof of Proposition 4.5

Let $H_{i,.1} = (R^T \otimes L^T, 0^T, 0^T)^T$, $H_{i,.2} = (\eta_i R^T \otimes L_0^T, H_{i,22}^T, 0^T)^T$ and $H_{i,.3} = (R_0^T \otimes \eta_i^T L^T, 0^T, H_{i,33}^T)^T$ be the first three block columns of H_i, and H_i^* denote the matrix that

contains the rest block columns of H_i. After some matrix derivations, we have

$$H_i^T J_i H_i = (H_{i,.1}, H_{i,.2}, H_{i,.3}, H_i^*)^T J_i (H_{i,.1}, H_{i,.2}, H_{i,.3}, H_i^*)$$
$$= \begin{pmatrix} H_{i,.1}^T J_i H_{i,.1} & 0 & 0 & 0 \\ 0 & H_{i,.2}^T J_i H_{i,.2} & 0 & 0 \\ 0 & 0 & H_{i,.3}^T J_i H_{i,.3} & 0 \\ 0 & 0 & 0 & H_i^{*T} J_i H_i^* \end{pmatrix}$$

which is block-diagonal with $H_{i,.1}^T J_i H_{i,.1} = c_i \Omega_2^{-1} \otimes \Omega_1^{-1}$, $H_{i,.2}^T J_i H_{i,.2} = c_i \eta_i \Omega_2^{-1} \eta_i^T \otimes \Omega_{10}^{-1} + m\Omega_1 \otimes \Omega_{10}^{-1} + m\Omega_1^{-1} \otimes \Omega_{10} - 2mI_{u_1} \otimes I_{r-u_1}$ and $H_{i,.3}^T J_i H_{i,.3} = c_i \Omega_{20}^{-1} \otimes \eta_i^T \Omega_1^{-1} \eta_i + rK_{u_2(m-u_2)}^T (\Omega_2 \otimes \Omega_{20}^{-1} + \Omega_2^{-1} \otimes \Omega_{20} - 2I_{u_2} \otimes I_{m-u_2}) K_{u_2(m-u_2)}$. The expression of $H_i^{*T} J_i H_i^*$ is not needed for our proof. Therefore,

$$\Lambda_i = H_i (H_i^T J_i H_i)^\dagger H_i^T = \sum_{j=1}^{3} H_{i,.j} (H_{i,.j}^T J_i H_{i,.j})^\dagger H_{i,.j}^T + H_i^* (H_i^{*T} J_i H_i^*)^\dagger H_i^{*T}.$$

After some algebra work, we obtain

$$\begin{aligned} &\text{avar}[\sqrt{n}\text{vec}(\hat{\alpha}_i)] \\ &= c_i R\Omega_2 R^T \otimes L\Omega_1 L^T + (R\eta_i^T \otimes L_0)(H_{i,.2}^T J_i H_{i,.2})^\dagger (\eta_i R^T \otimes L_0^T) + \\ &(R_0 \otimes L\eta_i)(H_{i,.3}^T J_i H_{i,.3})^\dagger (R_0^T \otimes \eta_i^T L^T). \end{aligned}$$

According to the delta method, we have $\text{avar}[\sqrt{n}h(\phi_i)] = W_i \text{avar}(\sqrt{n}\phi_i) W_i^T$. Hence the Fisher information of $\hat{\phi}_i$ is $W_i^T J_i W_i$, which can be represented as

$$\begin{pmatrix} J_{\eta_i\eta_i} & J_{\eta_i L} & J_{\eta_i R} & 0 & 0 & 0 & 0 \\ J_{L\eta_i} & J_{LL} & J_{LR} & J_{L\Omega_1} 0 & 0 & 0 & \\ J_{R\eta_i} & J_{RL} & J_{RR} & 0 & 0 & J_{R\Omega_2} & 0 \\ 0 & J_{\Omega_1 L} & 0 & J_{\Omega_1\Omega_1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & J_{\Omega_{10}\Omega_{10}} & 0 & 0 \\ 0 & 0 & J_{\Omega_2 R} & 0 & 0 & J_{\Omega_2\Omega_2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & J_{\Omega_{20}\Omega_{20}} \end{pmatrix},$$

where $J_{\eta_i\eta_i} = -\mathrm{E}[\partial^2 l/\partial\phi_{i1}\phi_{i1}^T]$, $J_{\eta_i L} = -\mathrm{E}[\partial^2 l/\partial\phi_{i1}\phi_{i2}^T]$, $J_{\eta_i R} = -\mathrm{E}[\partial^2 l/\partial\phi_{i1}\phi_{i3}^T]$, and the other terms are analogically defined. These matrices can be computed based on the definitions of W_i and J_i.

If L and R are known, the asymptotic variance of the MLE of η_i, denotes as $\hat{\eta}_{i(L,R)}$ is $J_{\eta_i\eta_i}^{-1}$, because when we cross out the second and third rows and columns of the matrix $H_g^T J_i H_g$, the remaining matrix is block-diagonal with blocks $J_{\eta_i\eta_i}$, $J_{\Omega_1\Omega_1}$, $J_{\Omega_{10}\Omega_{10}}$, $J_{\Omega_2\Omega_2}$ and $J_{\Omega_{20}\Omega_{20}}$. Similarly, if η_i and L are known, the asymptotic variance of the MLE of R, denotes as $\hat{R}_{\eta_i,L}$ is $(J_{RR} - J_{R\Omega_2} J_{\Omega_2\Omega_2}^{-1} J_{R\Omega_2}^T)^{-1}$; and if η_i and R are known, the asymptotic variance of the MLE of L, denoted as $\hat{L}_{\eta_i,R}$, is $(J_{LL} - J_{L\Omega_1} J_{\Omega_1\Omega_1}^{-1} J_{L\Omega_1}^T)^{-1}$. After some matrix operations, we obtain

$$
\begin{aligned}
&\mathrm{avar}[\sqrt{n}\mathrm{vec}(\hat{\eta}_{i(L,R)})] = (c_i^{-1}\Omega_2^{-1} \otimes \Omega_1^{-1})^{-1} = c_i\Omega_2 \otimes \Omega_1 \\
&\mathrm{avar}[\sqrt{n}\mathrm{vec}(\hat{L}_{\eta_i,R})] \\
&= [c_i\eta_i\Omega_2^{-1}\eta_i^T \otimes \Sigma_1^{-1} + m(\Omega_1 \otimes L_0\Omega_{10}^{-1}L_0^T + \Omega_1^{-1} \otimes L_0\Omega_{10}L_0^T - 2I_{u_1} \otimes L_0L_0^T)]^{-1} \\
&\mathrm{avar}[\sqrt{n}\mathrm{vec}(\hat{R}_{\eta_i,L})] = [c_i\Sigma_2^{-1} \otimes \eta_i^T\Omega_1^{-1}\eta_i + \\
&rK_{u_2m}^T(\Omega_2 \otimes R_0\Omega_{20}^{-1}R_0^T + \Omega_2^{-1} \otimes R_0\Omega_{20}R_0^T - 2I_{u_2} \otimes R_0R_0^T)K_{u_2m}]^{-1}.
\end{aligned}
$$

It can be seen that $H_{.2}^T J H_{.2} = (I \otimes L_0^T)\{\mathrm{avar}[\sqrt{n}\mathrm{vec}(\hat{L}_{\eta_i,R})]\}^{-1}(I \otimes L_0)$ and $H_{.3}^T J H_{.3} = (R_0^T \otimes I)\{\mathrm{avar}[\sqrt{n}\mathrm{vec}(\hat{R}_{\eta_i,L})]\}^{-1}(R_0 \otimes I)$. Therefore, by using Corollary E.1 (Cook, Li and Chiaromonte 2010), we have

$$
\begin{aligned}
&\mathrm{avar}[\sqrt{n}\mathrm{vec}(\hat{\alpha}_i)] = (R \otimes L)\mathrm{avar}[\sqrt{n}\mathrm{vec}(\hat{\eta}_{i(L,R)})](R^T \otimes L^T) + \\
&(R\eta_i^T \otimes \mathrm{P}_{L_0})\mathrm{avar}[\sqrt{n}\mathrm{vec}(\hat{L}_{\eta_i,R})](\eta_i R^T \otimes \mathrm{P}_{L_0}) + (\mathrm{P}_{R_0} \otimes L\eta_i)\mathrm{avar}[\sqrt{n}\mathrm{vec}(\hat{R}_{\eta_i,L})](\mathrm{P}_{R_0} \otimes \eta_i) \\
&= \mathrm{avar}[\sqrt{n}\mathrm{vec}(L\hat{\eta}_{i(L,R)}R^T)] + \mathrm{avar}[\sqrt{n}\mathrm{vec}(\mathrm{P}_{R_0}\hat{L}_{\eta_i,R}\eta_i R^T)] + \mathrm{avar}[\sqrt{n}\mathrm{vec}(L\eta_i\hat{R}_{\eta_i,L}\mathrm{P}_{R_0})] \\
&= \mathrm{avar}[\sqrt{n}\mathrm{vec}(\hat{\alpha}_{i(L,R)})] + \mathrm{avar}[\sqrt{n}\mathrm{vec}(Q_{\mathcal{E}_1}\hat{\alpha}_{i(\eta_i,R)})] + \mathrm{avar}[\sqrt{n}\mathrm{vec}(\hat{\alpha}_{i(L,\eta_i)}Q_{\mathcal{E}_2})].
\end{aligned}
$$

Chapter 5

Future works

5.1 Semiparametric higher-order sufficient dimension reduction

In the literature, dimension reduction methods usually rely on linearity and constant variance assumptions on the predictors. These assumptions are similarly required by our proposed tensor SDR approaches, although in a tensor formulation. Inspired by Ma and Zhu (2012), we plan to establish semiparametric tensor SDR that is free of both conditions and yields consistent estimation of the central dimension reduction subspace. By incorporating tensor SDR into the semiparametric framework, the higher-order dimension reduction problems become semiparametric estimation problems. Therefore, statistical inference can be established by powerful semiparametric tools.

In semiparametric tensor SDR, we can use the geometric approach in Bickel et al. (1993) and Tsiatis (2006) to derive a class of influence functions and construct the estimation for the sufficient reduction. In this way, we can relax the linearity and constant variance condition, and further extend the methods to data with discrete or categorical covariates.

5.2 SDR for longitudinal data with ramdom effects

Consider model-based SDR for longitudinal data. By incorporating random effects, the conventional PCA model can be modified as

$$X_{ij} = \mu + \alpha_i + \Gamma\nu_{ij} + \varepsilon_{ij}, \tag{5.1}$$

where $i = 1, \cdots, n$, $j = 1, \cdots, m_i$, X_{ij} is the p dimensional covariate vector for subject i at time j, μ is the grand mean $\mathrm{E}(X_{ij})$, α_i is the random subject effect and is assumed to be identically and independently normal distributed as $N(0, D)$ for all i. Γ is a standardized orthogonal $p \times d$ matrix, that is $\Gamma^T\Gamma = I_d$. Here ν_{ij} is defined as $\Gamma^T[\mathrm{E}(X_{ij}|Y_{ij}) - \mathrm{E}(X_{ij})]$, then ν_{ij} is fixed. The random error ε_{ij} denotes the within group variation and is identically and independently distributed as $N(0, \sigma^2 I_p)$, for all i and j. In addition, ε_{ij} is independent with α_i and ν_{ij} for all i, j. The key idea of the this method is to incorporate the random subject effect into the model thus data with correlated observations can achieve dimension reduction. For example, in a study of high school academic performance, ten schools are random selected from a city and fifty students are chosen from each school, whose academic stores and other possible influential factors on their scores are collected. Here the school variable is considered as a random effect since the student performances within each school are usually correlated. Conventional PCA and PFC model cannot handle such complex data structure, but model (5.1) can capture the subject effect with both α_i and ν_{ij}.

The central subspace (CS) $\mathcal{S}_{v|X}$ based on model (5.1) is span(Γ), since

$$\nu \perp\!\!\!\perp X \mid \Gamma^T X.$$

One can apply maximum likelihood method to estimate Γ. Considering the inverse model for each subject,

$$X_i = (1_m \otimes I_p)\mu + (1_m \otimes I_p)\alpha_i + (1_m \otimes I_p)\Gamma\nu_i + \varepsilon_i, \tag{5.2}$$

where X_i is the subject covariate vector equal to $(X_{i1}^T, \cdots, X_{im_i}^T)^T$, ε_i is the subject

within group error vector $(\varepsilon_{i1}^T, \cdots, \varepsilon_{im_i}^T)^T$, ν_i is equal to $(\nu_{i1}, \cdots, \nu_{im_i})$ and $(1_m \otimes I_p) = (I_p, I_p, \cdots, I_p)^T$ with m identity matrices of dimension p as components. The CS can be estimated by maximum likelihood estimation.

By replacing ν_{ij} with $\beta_i f(Y_{ij})$, the PFC random effect model is established as

$$X_{ij} = \mu + \alpha_i + \Gamma\beta_i f(Y_{ij}) + \varepsilon_{ij}, \;\; i = 1, ..., n, \; j = 1, ..., m, \tag{5.3}$$

where $Y_{ij} \in \mathbb{R}^1$ is the response variable for subject i at time j, $f(Y_{ij}) \in \mathbb{R}^r$ is a known vector-valued function and can be determined by inverse response plot or a sequence of basis functions (Cook and Forzani 2008), and $\beta_i \in \mathbb{R}^{d\times r}$ is the coordinates of $f(Y_{ij})$. The coeffiicent β_i can be either fixed or random depending on the data structure. The random error ε_{ij} is independent of β_i and has a normal distribution $N(0, \Delta)$. Under (5.3), it can be shown that the central subspace $\mathcal{S}_{Y|X}$ is span(Γ). One can estimate the CS by maximum likelihood estimation. In this case, the response information is contained in the estimation of the CS.

References

[1] Adragni, K. P. and Cook, R. D. (2009). Sufficient dimension reduction and prediction in regression. *Phil. Trans. Royal Soc. A* **367**, 4385–4405.

[2] Adragni, K. P. and Cook, R. D. (2009). Sufficient dimension reduction and prediction in regression. *Phil. Trans. Royal Soc. A* **367**, 4385–4405.

[3] Blondin, D. (2007). Rates of strong uniform consistency for local least squares kernel regression estimators. *Statist. & Prob. Letters* **77**, 1526–1534.

[4] Bura, E., and Cook, R.D. (2001). Extending sliced inverse regression: the weighted chi-squared test. *J. Amer. Statist. Assoc.* **96**, 996-1003.

[5] Christensen, R. (2001). *Advanced Linear Modeling.* Springer, New York.

[6] Conway, J. (1990). *A Course in Functional Analysis.* Springer, New York.

[7] Cook, R. D. (1994). On the interpretation of regression plots. *J. Amer. Statist. Assoc.* **89**, 177–190.

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

[9] Cook, R. D. (2007). Fisher Lecture: Dimension reduction in regression (with discussion). *Statist. Sci.* **22**, 1–26.

[10] Cook, R. D. and Forzani, L. (2008). Principal fitted components for dimension reduction in regression. *Statist. Sci.* **23**, 485–501.

[11] Cook, R. D. and Forzani, L. (2009). Likelihood-Based Sufficient Dimension Reduction. *J. Amer. Statist. Assoc.* **104**, 197–208.

[12] Cook, R. D., Li, B. and Chiaromonte, F. (2007). Dimension reduction without matrix inversion. *Biometrika* **94**, 569–584.

[13] Cook, R. D., Li, B. and Chiaromonte, F. (2010). Envelope Models for Parsimonious and Efficient Multivariate Linear Regression. *Statist. Sin.* **20**, 927–1010.

[14] Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. *J. Amer. Statist. Assoc.* **100**, 410–428.

[15] Cook, R. D. and Weisberg, S. (1991). Discussion of Sliced inverse regression for dimension reduction, by K.-C. Li. *J. Amer. Statist. Assoc.* **86**, 328–332.

[16] Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68**, 265–274.

[17] De Waal, D.J. (1985). Matrix-valued distributions. *Encycl. Statist. Sci.* **5**, (Ed., Kotz, S. and Johnson, N.L.), 326–333. Wiley, New York.

[18] Ding, S. and Cook, R. D. (2013). Dimension folding PCA and PFC for matrix-valued predictors. *Statist. Sin.* **24**, 463–492.

[19] Dong, Y. and Li, B. (2010). Dimension reduction for non-elliptically distributed predictors: second-order methods. *Biometrika*, **97**, 279–294.

[20] Dutilleul, P. (1999). The MLE algorithm for the matrix normal distribution. *J. Stat. Comput. Sim.* **64**, 105–123.

[21] Eaton, M. L. (1983). *Multivariate statistics : a vector space approach.* New York: Wiley.

[22] Ferré, L. (1998). Determining the dimension in sliced inverse regression and related methods. *J. Amer. Statist. Assoc.* **93**, 132-140.

[23] Gasser, T., Bächer, P., and Möcks, J. (1982). Transformations towards the normal distribution of broad band spectral parameters of the EEG. *Electroencephalogr. Clin. Neurophysiol.* **53**, 119-124.

[24] Hall, W. J. and Mathiason, D. J. (1990). On large-sample estimation and testing in parametric models. *Int. Statist. Rev.*, **58**, 77-97.

[25] Henderson, H. V., and Searle, S. R. (1979). Vec and Vech operators for matrices, with some uses in Jacobians and multivariate statistics. *Canadian J. Statist.* **7**, 65–81.

[26] Hung, H. and Wang, C. (2013) Matrix variate logistic regression model with application to EEG data. *Biostatistics* **14**, 189–202.

[27] Hung, H., Wu, P., Tu, I., and Huang, S. (2012). On multilinear principal component analysis of order-two tensors. *Biometrika* **99**, 569-583.

[28] Kolda, T. G. (2006). *Multilinear operators for higher-order decompositions.* Tech. Report SAND2006-2081, Sandia National Laboratories, Albuquerque, New Mexico and Livermore, California.

[29] Kolda, T. G. and Bader, B. W. (2009). Tensor decomposition and application. *SIAM Review* **51**, 455–500.

[30] Kroonenberg, P. M. and Leeuw, J. D. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithm. *Psychometrika* **45**, 69–97.

[31] Lathauwer, L. D., Moor, B. D. and Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM J. Matrix Anal. & Appl.*, **21**, 1253–1278.

[32] Lathauwer, L. D., Moor, B. D. and Vandewalle, J. (2000). On the best rank-1 and rank-$R_1, R_2, ..., R_N$ approximation of higher-order tensors. *SIAM J. Matrix Anal. & Appl.*, **21**, 1324–1342.

[33] Leibovici, D. (1998). A singular value decomposition of a k-way array for a principal component analysis of multiway data, PTA-k. *Linear Algebra Appl.* **269**, 307–329.

[34] Li, B. and Dong, Y. (2009). Dimension reduction for non-elliptically distributed predictors. *Ann. Statist.* **37**, 1272–1298.

[35] Li, B., Kim, K.M., and Altman, N. (2010). On dimension folding of matrix or array-valued statistical objects. *Ann. Statist.* **38**, 1094–1121.

[36] Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc.* **102**, 997–1008.

[37] Li, K. C. (1991). Sliced inverse regression for dimension reduction with discussion. *J. Amer. Statist. Assoc.* **86**, 316–327.

[38] Li, K. -C., Aragon, Y., Shedden, K., and Agnan, C. T. (2003). Dimension reduction for multivariate response data. *J. Am. Statist. Assoc.* **98** , 99–109.

[39] Lu, N. and Zimmerman, D.L. (2005). The likelihood ratio test for a sep- arable covariance matrix. *Statist. & Prob. Letters* **73**, 449–457.

[40] Ma, Y. and Zhu, L. (2012). A Semiparametric Approach to Dimension Reduction. *J. Amer. Statist. Assoc.*, **107**, 168–179.

[41] Ma, Y. and Zhu, L. (2013). Efficient Estimation in Sufficient Dimension Reduction. *Ann. Statist.*, **41**, 250–268.

[42] Pan, J. and Fang K. (2000). *Growth curve models and statistical diagnostics.* Springer, New York.

[43] Pfeiffer, R. M., Forzani, L. and Bura, E. (2012). Sufficient dimension reduction for longitudinally measured predictors. *Statist. Med.*, **31**, 2414–2427.

[44] Roy, A. and Khattree, R. (2005). On implementation of a test for Kro- necker product covariance structure for multivariate repeated measures data. *Statist. Mthd.* **2**, 297–306.

[45] Seber, G. A. F. (1984). *Multivariate Observations.* Wiley, New York.

[46] Shan, S., Cao, B., Su, Y., Qing, L., Chen, X., and G, W. (2008). Unified Principal Component Analysis with generalized Covariance Matrix for face recognition. *IEEE Conf. on Comp. Vis. and Pat. Recog.* **13**, 1–7.

[47] Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *J. Amer. Statist. Assoc.* **81**, 142–149.

[48] Shitan, M. and Brockwell, P.J. (1995). An asymptotic test for separability of a spatial autoregressive model. *Commun. Statist. -Theor. M.* **24**, 2027–2040.

[49] Sibson, R. (1979). Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling. *J. R. Statist. Soc. B*, **41**, 217–229.

[50] Su, Z. and Cook, R. D. (2011). Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika* **98**, 133–146.

[51] Su, Z. and Cook, R. D. (2012). Inner envelopes: efficient estimation in multivariate linear regression. *Biometrika* **9**, 687–702. .

[52] Su, Z. and Cook, R. D. (2013). Estimation of multivariate means with heteroscedastic error using envelope models. *Statist. Sin.* **23**, 213–230.

[53] Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *J. Royal Statist. Soc. B* **61**, 611-622.

[54] Viroli, Cinzia (2012). On matrix-variate regression analysis. *J. Multivariate Anal.* **111**, 296–309.

[55] Yang, J., Zhang, D., Frangi, A. F., and Yang, J. (2004). Two dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Trans. on Pat. Anal. and Mac. Intel.* **26**, 131–137.

[56] Ye, J. (2005). Generalized low rank approximation of matrices. *Mac. Learn.* **61**, 167–191.

[57] Ye, Z. and Weiss, R. E. (2003). Using the Bootstrap to select one of a new class of dimension reduction methods. *J. Amer. Statist. Assoc.* **98**, 968–979.

[58] Yu, S., Bi, J., and Ye, J. (2011). Matrix-variate and higher-order probabilistic projections. *Data Min. and Knowl. Disc.* **22**, 372–392.

[59] Zhang, D. and Zhou, Z. (2005). (2D)2PCA: Two-directional two-dimensional PCA for efficient face representation and recognition. *Neurocomp.* **69**, 224–231.

[60] Zhou, H., Li, L. and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *J. Am. Statist. Assoc.* **108**, 540–552.

[61] Zhu, L. and Fang, K. (1996). Asymptotics for kernel estimate of sliced inverse regression. *Ann. Statist.*, **24**, 1053–1068.

[62] Zhu, L. and Ng, K. W. (1995). Asymptotics of sliced inverse regression. *Statist. Sin.* **5**, 727–736.