Sufficient Dimension Folding, Variable Selection and its Inference

by

Yuanwen Wang

(Under the direction of Professor Xiangrong Yin)

Abstract

At the era of big data where 2.5 quintillion bytes of data are produced daily, effective reduction on predictor variables to maintain important information about response variable is required. Sufficient dimension folding (SDF) in particular defines a powerful framework for compressing dimensions of matrix/array predictor variable while preserving its inner structure. This dissertation is composed with three studies based on SDF. In the first study, we introduce sufficient dimension folding with categorical variables which simultaneously brings matrix predictors and categorical variables into consideration during the reduction. In order to improve interpretation of SDF methods, we propose model-free variable selection techniques in the second study by reformulating SDF methods as least square estimations and adapt regularized regression methods to regularized SDF methods. In the third study, three hypothesis testing methods including marginal dimension test, conditional and marginal coordinate tests are presented as future study to assist evaluating matrix predictor's contributions.

Index words: Central folding subspace, marginal coordinate hypothesis testing, partial folding subspace, sufficient dimension folding.

Sufficient Dimension Folding, Variable Selection, and its Inference

by

Yuanwen Wang

B.S., Southwest Jiaotong University, 2012

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2016

Sufficient Dimension Folding, Variable Selection, and its Inference

by

Yuanwen Wang

Approved:

Major Professor: Xiangrong Yin

Committee: Pengsheng Ji
William McCormick
T.N. Sriram
Wenxuan Zhong

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2016

Acknowledgments

I would like to express my sincere gratitudes to my major advisor Professor Xiangrong Yin. As Tang dynasty poet Yu Han once described the common characteristics of remarkable teachers as "one who could propagate the doctrine, impart professional knowledge, and resolve doubts", Dr. Yin surely demonstrates what it takes to be an extraordinary teacher. His resolution to tackle challenging problems, his profound understanding on statistical theories and his infinite enthusiasm in the mysterious world of statistics is always inspiring me to strive for excellence in my research.

I would also like to thank Dr. McCormick, Dr. Ji, Dr Sriram and Dr. Zhong for serving as my committee members and offering me numerous constructive comments and kind encouragement along my research. I am honored to study in our department for the entire four years, and being able to "steal" some of the great qualities from excellent educators like Prof. Franklin and Mr. Morse.

My sincere appreciation is also extended to my family members including my father Xiaoping Wang and mother Xiaolan Nie for their endless and continuous support over the last twenty six years, my twin brother Bowen Wang for motivating me to move beyond good, and last but not least, my girlfriend Xiaoshi Zeng for giving me the best time of my life.

I am always thankful to God for "The Lord is merciful and gracious, slow to anger and plenteous in mercy."

Contents

Acknowledgments **iv**

List of Figures **vii**

List of Tables **viii**

1 Overview **1**

2 Sufficient Dimension Folding with Categorical Variables **6**

2.1 Introduction 7

2.2 Review on sufficient dimension folding 8

2.3 Review on sufficient dimension reduction with categorical variables 10

2.4 Sufficient dimension folding with categorical variable 12

2.5 Estimation methods 15

2.6 Estimation of structural dimensions 26

2.7 Numerical studies 27

2.8 Application 37

2.9 Discussion 42

2.10 Appendix 44

3 Regularized Sufficient Dimension Folding and Variable Selection **54**

3.1 Introduction 55
3.2 Review on regularized sufficient dimension reduction and variable selection . 57
3.3 Regularized sufficient dimension folding and variable selection 60
3.4 Estimation of structural dimensions 89
3.5 Numerical Study 95
3.6 Application 101
3.7 Discussion 104
3.8 Appendix 104

4 Testing Variable Contributions in Sufficient Dimension Folding 115

4.1 Introduction 116
4.2 Review on testing predictor contribution in sufficient dimension reduction . . 117
4.3 Testing predictor contribution in sufficient dimension folding 122

List of Figures

2.1 Summary of conditional and partial folding subspaces in four examples . . . 36

2.2 Eigenvalue plots for estimating structural dimensions 39

2.3 Bootstrap confidence intervals for estimated directions 41

2.4 Summary of smoothing splines with reduced predictors 43

3.1 Description on different types of sparseness 63

3.2 Visualization of true central folding space in Example 3.3 and Example 3.4 . 96

3.3 Summary of smoothing splines with reduced predictors for three models . . . 103

List of Tables

2.1 Example 2.1, accuracy of estimates on partial folding subspace 30

2.2 Example 2.2, accuracy of estimates on partial folding subspace 32

2.3 Example 2.3, accuracy of estimates on partial folding subspace 34

2.4 Example 2.4, accuracy of estimates on partial folding subspace 35

2.5 Estimated directions from three methods 40

3.1 Average of TPR and FPR based on 100 replications for example 3.3 98

3.2 Average of TPR and FPR based on 100 replications for example 3.4 99

3.3 Percentages of estimated AIC, BIC and RIC structural dimensions 100

Chapter 1

Overview

Due to the explosion of large size and complicated structured data in recent years, how to reduce redundant information and retain only key information about the data to a smaller amount remains a challenging problem. A typical use-case is regression analysis, where one seeks to explore the relationship between a univariate response variable Y and a $p \times 1$ vector predictor X, by studying the conditional distribution $Y|X$, and in particular the mean function $E(Y|X)$ and variance function $Var(Y|X)$.

Sufficient dimension reduction (SDR) combines the model-free reduction idea with regression analyses by studying the conditional distribution of $X|Y$. The purpose of SDR (Li, 1991; Cook, 1994, 1996) leads to the estimation of a smallest subspace S_B, such that $Y \perp\!\!\!\perp X|B^T X$ and $span(B)$ is referred as *central subspace* $S_{Y|X}$ (Cook, 1996). The conditions under which such space exists are studied in Cook (1998) and Yin, Li and Cook (2008). Despite the wide applications of SDR on vector predictor, adaption was also proposed for dealing with complicated structure data, such as for categorical predictor (Chiaromonte, Cook and Li, 2002), longitudinal data (Li and Yin, 2009) and multivariate response variable (Li et. al, 2003) etc.

A recent trend of need to analyze matrix/array predictor variable such as large sized images and videos has brought SDR into a more computational intensive challenge. A naive way is to convert matrix/array predictor into vector predictor and utilize SDR methods to achieve reduction. Such simplification, however, suffers from estimating large numbers of parameters and potentially loses the inner structure of the matrix/array. Sufficient dimension folding (SDF; Li, Kim and Altman, 2010) defined an alternative reduction approach aimed at reducing dimensions for matrix/array predictor variable while preserving its inner structure. For matrix predictor $\mathbf{X} \in \mathbb{R}^{p_l \times p_r}$, SDF raised a similar concept to *central subspace* as *central folding subspace*, which defines the smallest subspace $S_{\beta \otimes \alpha}$, such that $Y \perp\!\!\!\perp X | \alpha^T \mathbf{X} \beta$. SDF naturally extended moment-based SDR methods such as sliced inverse regression (SIR; Li, 1993), sliced average variance estimation (SAVE; Cook and Weisberg, 1991) and directional regression (Li and Wang, 2007) to estimate *central folding subspace* with methods as folded-SIR, folded-SAVE and folded-DR. Other developments include folded-MAVE and folded-OPG (Xue and Yin, 2014) which combined SDF framework with forward model minimum average variance estimation (MAVE; Xia et al, 2002). Ding and Cook (2014) developed dimension folding principal fitted components (DF-PFC) which estimated *central folding subspace* by constructing its maximum likelihood estimation.

In this dissertation, we focus on the theory and methods of SDF. We first consider the reduction of matrix/array predictor with the existence of categorical variable. Based on SDR with categorical variable discussed in Chiaromonte, Cook and Li (2002), we proposed the concept of *partial folding subspace* which incorporate categorical variable information into the reduced space and provide extra insights on how predictors influences the response variable. Three estimation methods are raised to efficiently estimate *partial folding subspace*. Estimation of SDF essentially recovers the linear combinations of rows and columns of the original matrix predictor $\mathbf{X} \in \mathbb{R}^{p_l \times p_r}$ that are related to response variable, thus may lack interpretation when the predictor size $p_l \times p_r$ is sufficiently large. In the second study, we

further pair SDF methods with regularization methods including ridge regression (Hoerl and Kennard, 1970) and lasso regression (Tibshirani, 1996) to produce sparse estimation on the *central folding subspace*. In parallel with regularized SDF methodology, we also evaluate predictors' contribution to response variable through hypothesis testing. We adopt the idea in Cook (2004) and define three tests including marginal dimension test, and conditional and marginal coordinate tests. Constructing corresponding test statistics and studying their asymptotic properties are our future work. Last but not least, we provide both simulation studies and data applications to each study to promote the efficacy of our proposed methods.

Bibliography

[1] F. Chiaronmonte, R. D. Cook and B. Li. Sufficient dimension reduction in regressions with categorical predictors. *The Annals of Statistics* **30(2)**, 475-497, 2002.

[2] R. D. Cook. Using dimension-reduction subspaces to identify important inputs in models of physical systems. *Proceedings of the Section on Physical and Engineering Sciences*, 1825, 1994.

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

[4] R. D. Cook. Testing predictor contribution in sufficient dimension reduction. *The Annals of Statistics* **32**, 1062-1092, 2004.

[5] R. D. Cook and S. Weisberg. Discussion of "Sliced inverse regression for dimension reduction". *Journal of the American Statistical Association* **86**, 28-33, 1991.

[6] S. Ding and R. D. Cook. Dimension folding PCA and PFC for matrix-valued predictors. *Statistica Sinica* **24**, 463-492, 2014.

[7] A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55-67, 1970.

[8] B. Li, M. Kim and N. Altman. On dimension folding of matrix- or array-valued statistical objects. *The Annals of Statistics* **38**, 1094-1121, 2010.

[9] B. Li, and S. Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association* **102**, 2143-2172, 2007.

[10] L. Li and X. Yin. Longitudinal data analysis using sufficient dimension reduction. *Computational Statistics and Data Analysis* **53**, 4106-4115, 2009.

[11] K.-C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**, 316-342, 1991.

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

[13] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical Society, Series B* **58**, 267-288, 1996.

[14] Y. Xia, H. Tong, W.K. Li and L. Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society, Series B* **64**, 363-410, 2002.

[15] Y. Xue and X. Yin. Sufficient dimension folding for regression mean function. *Journal of Computational and Graphical Statistics* **39**, 1028-1043, 2014.

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

Chapter 2

Sufficient Dimension Folding with Categorical Variables

Abstract

In this chapter, we study dimension folding where predictor variables are matix/array structured with categorical variables. Li, Kim and Altman (2010) proposed the framework of sufficient dimension folding which successfully reduces the dimensionality of matrix/array predictor, while preserving the inner matrix/array structure. Chiaromonte, Cook and Li (2002), on the other hand discuss the extension of sufficient dimension reduction with both quantitative and categorical predictor variables in the regression context. In this chapter, based on their works, we adapt the methodology of sufficient dimension folding to incorporate categorical variable information into the reduction. We first introduce the concepts of marginal, conditional and partial folding subspaces, and shed light on their connections with central folding subspace proposed by Li, Kim and Altman (2010). We propose three estimation methods to estimate the desired partial folding subspace. An empirical eigenvalue proportion plot method is also developed to determine the structural dimensions of

the associated partial folding subspace. Effectiveness of our proposed methods is evaluated through simulations and an application to a longitudinal data with gender being the categorical predictor.

2.1 Introduction

As IBM Big Data & Analytics Hub (IBM, 2014) has pointed out, nowadays "Big Data" often come with 4Vs, "Volume", "Variety", "Velocity" and "Veracity", among which "Volume" and "Variety" usually refer to data with high dimensions and complicated structures. In particular, for complex data such as images or audios, each observation can take the form of a large sized 2-D matrix, or even higher dimensional array. Effectively reducing the number of dimensions while preserving the data structure remains a challenging task, since most dimension reduction methods are well suited for regular data where each observation is a vector instead. Li, Kim and Altman (2010) proposed the framework of sufficient dimension folding which extended the moment-based dimension reduction method such as sliced inverse regression (SIR; Li, 1993), sliced average variance estimation (SAVE; Cook and Weisberg, 1991) and directional regression (DR; Li and Wang, 2007), to methods that gain insights by studying conditional distribution of $\mathbf{X}|Y$, such as Folded-SIR, Folded-SAVE and Folded-DR.

Nevertheless, real data usually come with both matrix predictors as well as traditional categorical variables. Based on Chiaromonte, Cook and Li (2002) who developed sufficient dimension reduction with categorical predictors in regression context, we also extend sufficient dimension folding with consideration of categorical predictor variables. Similar to Chiaromonte, Cook and Li (2002), we define three types of reduced spaces including *marginal folding subspace* that does not involve categorical variable information, *conditional folding subspace* that corresponds to the *central folding subspace* within each category and *partial folding subspace* that combines different *conditional folding subspaces*. We bridge the gap

between these spaces by theoretical results that also shed lights on how to estimate them, in particular, *partial folding subspace*. Three numerical algorithms are established to estimate *partial folding subspace*, which summarizes the *central folding subspace* with consideration on categorical predictors. An empirical criterion is proposed to further determine the structural dimension for the *partial folding subspace*. Our simulation, as well as an application on a longitudinal dataset indicate that our proposed *partial folding subspace* provides better insights by including both matrix predictor and categorical predictor in the same model.

The rest of the chapter is organized as the follows: Section 2.2 and Section 2.3 mainly review on the concepts of sufficient dimension folding for matrix/array predictor and sufficient dimension reduction with categorical variables for vector predictor. Section 2.4 is devoted to the theoretical foundations of sufficient dimension folding with categorical variables, where we define three types of reduced spaces including marginal, conditional and partial folding subspaces, and we end this section with their theoretical connections among these spaces. In Section 2.5, we proceed with three estimation methods to estimate in particular *partial folding subspace*. Section 2.6 explores different simulation setting including forward and inverse models with different types of *partial central subspaces*. In Section 2.7, we apply our estimation methods to a longitudinal dataset where time points and bio-markers form the matrix predictor, with the presence of gender information as a categorical variable. Our results clearly indicate that the combined *partial folding subspace* greatly reduces the matrix dimensions while providing better insights on predicting associated response variable.

2.2 Review on sufficient dimension folding

Suppose B is a $p \times q$ matrix, then $Span(B)$ is the space spanned by the columns of B and $P_B = B(B^TB)^-B$ is the projection matrix onto $span(B)$. We consider a regression or a classification problem involving a univariate response variable Y with a random $p_l \times p_r$

predictor matrix $\mathbf{X}$.According to Li, Kim and Altman (2010), if there are two matrices $\alpha \in p_l \times d_l$ and $\beta \in p_r \times d_r$ ($d_l < p_l$ and $d_r < p_r$) satisfying

$$Y \perp\!\!\!\perp \mathbf{X} | \alpha^T \mathbf{X} \beta, \tag{2.1}$$

then we can conclude that Y depends on $\mathbf{X}$ only through the transformed matrix $\alpha^T \mathbf{X} \beta$. $Span(\alpha)$ and $Span(\beta)$ are defined as *left dimension folding subspace* and *right dimension folding subspace*.

Furthermore, Li, Kim and Altman (2010) pointed out that the intersection of *left dimension folding subspace* (*right dimension folding subspace*) is again, *left dimension folding space* (*right dimension folding subspace*), under some regularity conditions. Thus, if we denote $S_{Y|\circ\mathbf{X}}$ and $S_{Y|\mathbf{X}\circ}$ as the respective intersections of all *left dimension folding subspace* and *right dimension folding subspace*, we can define the following space $S_{Y|\circ\mathbf{X}\circ}$ as

$$S_{Y|\circ\mathbf{X}\circ} = S_{Y|\mathbf{X}\circ} \otimes S_{Y|\circ\mathbf{X}} = span(\beta \otimes \alpha), \tag{2.2}$$

where β and α are the basis matrices of $S_{Y|\mathbf{X}\circ}$ and $S_{Y|\circ\mathbf{X}}$, respectively. And $S_{Y|\circ\mathbf{X}\circ}$ is called *central folding subspace* (Li, Kim and Altman, 2010).

The relationship between the new estimators of *central folding subspace* and the traditional inverse method estimators of *central subspace* (for vectorized data) is captured through the concept of Kronecker envelope developed by Li, Kim and Altman (2010), an idea similar to Σ -envelope in Cook, Li and Chiaromonte (2010). A Kronecker envelope $\epsilon^{\otimes}(U) = S_{Y|\mathbf{U}\circ} \otimes S_{Y|\circ U}$ of a random matrix $U \in (r_R r_L) \times k$ satisfies the following conditions:

1. $span(U) \subseteq S_{Y|U\circ} \otimes S_{Y|\circ U}$ almost surely.
2. If there exist another pair of subspaces $S_R \in \mathbb{R}^{r_R}$ and $S_L \in \mathbb{R}^{r_L}$ satisfying 1, then $S_{Y|U\circ} \otimes S_{Y|\circ U} \subseteq S_R \otimes S_L$.

Li, Kim and Altman (2010) further illustrated that if $Span(U) \subseteq S_{Y|vec(\mathbf{X})}$, then we have the following relationship,

$$Span(U) \subseteq S_{Y|vec(\mathbf{X})} \subseteq \epsilon^{\otimes}(U) \subseteq S_{Y|\circ\mathbf{X}\circ}. \tag{2.3}$$

Thus, by specifying matrix U through traditional sufficient dimension reduction estimators such as sliced inverse regression (SIR; Li, 1991), sliced average variance estimation (SAVE; Cook and Weisberg, 1991) and directional regression (DR; Li and Wang, 2007) which target the *central subspace* for vectorized data, one can recover *central folding subspace* $S_{Y|\circ\mathbf{X}\circ}$ by estimating the Kronecker envelope of matrix U. The underlying algorithm for each method turns out to be an alternate ordinary least square method which recovers the basis matrices α and β, alternately.

Our theoretical foundation is mostly based on Li, Kim and Altman (2010), while other existing works with regards to dimension reduction for matrix/array data include dimension folding PCA (DF-PCA) and dimension folding principal fitted components (DF-PFC) in Ding and Cook (2014) and tensor sliced inverse regression in Ding and Cook (2015). Xue and Yin (2014, 2015) also extended and adapted the concept of sufficient dimension folding for various modeling purposes such as for modeling regression mean function and for a functional of conditional distribution of matrix- or array-valued objects.

2.3 Review on sufficient dimension reduction with categorical variables

We first focus on dimension reductions for vector predictor $\mathbf{X} \in \mathbb{R}^p$, with the presence of one or more additional categorical variables characterized by one random variable W with levels $w = 1, 2, ..., C$.

Define the intersection of all subspaces $S \subseteq \mathbb{R}^p$ satisfying

$$Y \perp\!\!\!\perp \mathbf{X} | P_S \mathbf{X}, \tag{2.4}$$

as *marginal central subspace*, denoted by $S_{Y|\mathbf{X}}$, which is essentially *central subspace* when only considering continuous predictor vector $\mathbf{X}$ (Cook, 1996).

Define that, for each level of categorical predictor W, that is, $W = w$, the intersection of all S satisfying

$$Y \perp\!\!\!\perp \mathbf{X}(P_S \mathbf{X}, W = w), \tag{2.5}$$

as *conditional central subspace*, denoted by $S_{Y_w|\mathbf{X}_w}$ (Chiaromonte, Cook and Li, 2002).

Define the intersection of all subspaces $S \subseteq \mathbb{R}^p$ satisfying

$$Y \perp\!\!\!\perp \mathbf{X} | (P_S \mathbf{X}, W), \tag{2.6}$$

as *partial central subspace*, denoted by $S^{(W)}_{Y|\mathbf{X}}$ (Chiaromonte, Cook and Li, 2002).

According to Chiaromonte, Cook and Li (2002), the three types of central subspaces are usually not identical, but the following proposition summarizes their connections.

Proposition 1 (*Chiaromonte, Cook and Li (2002)*)

(a) *If* $W \perp\!\!\!\perp \mathbf{X} | P_{S^{(W)}_{Y|\mathbf{X}}} \mathbf{X}$ *or* $W \perp\!\!\!\perp Y | P_{S^{(W)}_{Y|\mathbf{X}}} \mathbf{X}$, *then* $S^{(W)}_{Y|\mathbf{X}} \supseteq S_{Y|\mathbf{X}}$,

(b) *If* $W \perp\!\!\!\perp Y | \mathbf{X}$, *then* $S^{(W)}_{Y|\mathbf{X}} \subseteq S_{Y|\mathbf{X}}$,

(c) $S^{(W)}_{Y|\mathbf{X}} = \oplus_{w=1}^{C} S_{Y_w|\mathbf{X}_w}$,

(d) $S_{Y|\mathbf{X}} \subseteq S_{W|\mathbf{X}} \oplus (\oplus_{w=1}^{C} S_{Y_w|\mathbf{X}_w}) = S_{W|\mathbf{X}} \oplus S^{(W)}_{Y|\mathbf{X}}$.

Using part (c) of Proposition 1, Chiaromonte, Cook and Li (2002) constructed partial slice inverse regression to estimate the *partial central subspace* $S^{(W)}_{Y|\mathbf{X}}$. Part (c) of Proposition 1 suggests that $S^{(W)}_{Y|\mathbf{X}}$ can be estimated by combining sliced inverse regression estimators within subpopulations. If for each subpopulation $W = w$, $w = 1, ..., C$, we denote mean of

predictor as μ_w, covariance $\mathbf{\Sigma}_w$ for each subpopulation predictor $\mathbf{X}_w$. After standardization $\mathbf{Z}_w = \Sigma_w^{-\frac{1}{2}}(\mathbf{X}_w - \mu_w)$, we have

$$S_{Y|\mathbf{X}}^{(W)} = \oplus_{w=1}^{C} \Sigma_w^{-\frac{1}{2}} S_{Y_w|\mathbf{Z}_w}. \tag{2.7}$$

By assuming covariance structure is the same across subpopulations $\Sigma_w = \Sigma_{pool}$, and using weight average covariance matrix as its estimate $\hat{\Sigma}_{pool} = \sum_{w=1}^{C} \frac{n_w}{n} \hat{\Sigma}_w$ where n_w is the number of observations within each category and n is the total number of observations. One can construct the estimate for covariance matrix of $E(Z_w|Y_w)$ as $\hat{\Theta}^{(W)}$ by a weighted average $\hat{\Theta}^{(W)} = \sum_{w=1}^{C} \frac{n_w}{n} \hat{\Theta}_w$ where $\hat{\Theta}_w = \sum_{s=1} H_w \frac{n_{sw}}{n_w} \bar{\mathbf{Z}}_{sw} \bar{\mathbf{Z}}_{sw}'$, $\quad w = 1, 2, ...C$, $\bar{\mathbf{Z}}_{sw} = \frac{1}{n_{sw}} \sum_{i|s} \hat{\mathbf{Z}}_{iw}$ is the intra-slice mean vector and H_w is the number of slices.

By eigenvalue decomposition on $\hat{\Theta}^{(W)} = \sum_{j=1}^{p} \hat{\theta}_j \hat{\mathbf{t}}_j \hat{\mathbf{t}}_j'$, we obtain the basis directions of the projection matrix for *partial central space* as $\hat{\mathbf{v}}_j = \hat{\Sigma}_{pool}^{-\frac{1}{2}} \hat{\mathbf{t}}_j$, $\quad j = 1, ..., p$. A large sample testing statistic $T(m) = n \sum_{j=m+1}^{p} \hat{\theta}_j$ is applied to test the number of directions.

2.4 Sufficient dimension folding with categorical variable

Following the similar logic in sufficient dimension reduction with categorical variables, we derive concepts of *marginal, conditional and partial folding subspaces* for dimension reduction on matrix/array predictors in this section. Through out this section, we assume $\mathbf{X}$ is a $p_l \times p_r$ random matrix predictor with a categorical variable W characterized by $w = 1, 2, ..., C$.

2.4.1 Marginal, conditional and partial folding subspace

We denote $S_{Y|\circ\mathbf{X}\circ}$ as the *marginal folding subspace*, which only considers reduction on continuous matrix predictor $\mathbf{X}$.

Define that, for each $W = w$, the intersection of all spaces S_L and the intersection of all spaces S_R satisfying

$$Y \perp\!\!\!\perp \mathbf{X} | (P_{S_L} \mathbf{X} P_{S_R}, W = w), \tag{2.8}$$

conditional left folding subspace as $S_{Y_w|\circ \mathbf{X}_w}$, *conditional right folding subspace* as $S_{Y_w|\mathbf{X}_w \circ}$, and their Kronecker product $S_{Y_w|\mathbf{X}_w \circ} \otimes S_{Y_w|\circ \mathbf{X}_w}$ as *conditional folding subspace*, denoted by $S_{Y_w|\circ \mathbf{X}_w \circ}$. If we denote the basis matrices of $S_{Y_w|\circ \mathbf{X}_w}$ and $S_{Y_w|\mathbf{X}_w \circ}$ as $\alpha_w \in \mathbb{R}^{p_l \times d_l}$ and $\beta_w \in \mathbb{R}^{p_r \times d_r}$, respectively, then the *conditional folding subspace* $S_{Y_w|\circ \mathbf{X}_w \circ}$ can be equivalently defined as $S_{Y_w|\circ \mathbf{X}_w \circ} = span(\beta_w) \otimes span(\alpha_w) = span(\beta_w \otimes \alpha_w)$.

Define that, for discrete random variable W, the intersection of all spaces S_L and the intersection of all spaces S_R satisfying

$$Y \perp\!\!\!\perp \mathbf{X} | (P_{S_L} \mathbf{X} P_{S_R}, W) \tag{2.9}$$

partial left folding subspace as $S^{(W)}_{Y|\circ \mathbf{X}}$, *partial right folding subspace* as $S^{(W)}_{Y|\mathbf{X} \circ}$ and their Kronecker product $S^{(W)}_{Y|\mathbf{X} \circ} \otimes S^{(W)}_{Y|\circ \mathbf{X}}$ as *partial folding subspace*, denoted by $S^{(W)}_{Y|\circ \mathbf{X} \circ}$. If we denote the basis matrices of $S^{(W)}_{Y|\circ \mathbf{X}}$ and $S^{(W)}_{Y|\mathbf{X} \circ}$ as $\alpha^{(W)} \in \mathbb{R}^{p_l \times d_l}$ and $\beta^{(W)} \in \mathbb{R}^{p_r \times d_r}$, respectively, then the *partial folding subspace* $S^{(W)}_{Y|\circ \mathbf{X} \circ}$ can be equivalently defined as $S^{(W)}_{Y|\circ \mathbf{X} \circ} = span(\beta^{(W)}) \otimes span(\alpha^{(W)}) = span(\beta^{(W)} \otimes \alpha^{(W)})$.

We can easily infer from Li, Kim and Altman (2010) with the following equivalent relationships.

$$\begin{gathered}
Y \perp\!\!\!\perp \mathbf{X} | (P_{S_{Y|\circ \mathbf{X}}} \mathbf{X} P_{S_{Y|\mathbf{X} \circ}}) \Leftrightarrow Y \perp\!\!\!\perp \mathbf{X} | ([P_{S_{Y|\circ \mathbf{X} \circ}}]^T vec(\mathbf{X})), \\
Y \perp\!\!\!\perp \mathbf{X} | (P_{S_{Y_w|\circ \mathbf{X}_w}} \mathbf{X} P_{S_{Y_w|\mathbf{X}_w \circ}}, W = w) \Leftrightarrow Y \perp\!\!\!\perp \mathbf{X} | ([P_{S_{Y_w|\circ \mathbf{X}_w \circ}}]^T vec(\mathbf{X}), W = w), \\
Y \perp\!\!\!\perp \mathbf{X} | (P_{S^{(W)}_{Y|\circ \mathbf{X}}} \mathbf{X} P_{S^{(W)}_{Y|\mathbf{X} \circ}}, W) \Leftrightarrow Y \perp\!\!\!\perp \mathbf{X} | ([P_{S^{(W)}_{Y|\circ \mathbf{X} \circ}}]^T vec(\mathbf{X}), W).
\end{gathered} \tag{2.10}$$

It is worthwhile to point out that if we convert the matrix predictor $\mathbf{X}$ into a vector predictor, denoted as $vec(\mathbf{X})$ and look for the associated *partial central subspace* $S_{Y|\circ vec(\mathbf{X})\circ}$, then we have $S_{Y|\circ vec(\mathbf{X})\circ} \subseteq S_{Y|\circ \mathbf{X}\circ}$ as suggested in (2.3) by Li, Kim and Altman (2010). This is because the *conditional central subspace* derived from the vectorized predictor as $S_{Y_w|vec(\mathbf{X}_w)}$, is also a subspace in $S_{Y_w|\circ \mathbf{X}_w\circ}$, i.e., $S_{Y_w|vec(\mathbf{X}_w)} \subseteq S_{Y_w|\circ \mathbf{X}_w\circ}$. Therefore, vectorizing $\mathbf{X}$ and apply dimension deduction methods such as partial SIR developed by Chiaromonte, Cook and Li (2002) may lose information and recover a smaller space than the desired *partial folding subspace* $S^{(W)}_{Y|\circ \mathbf{X}\circ}$. Similar to the aforementioned Proposition 1 from Chiaromonte, Cook and Li (2002), we obtain the following result.

Proposition 2

(a) *If* $W \perp\!\!\!\perp \mathbf{X}|P_{S^{(W)}_{Y|\circ \mathbf{X}}}\mathbf{X}P_{S^{(W)}_{Y|\mathbf{X}\circ}}$ *or* $W \perp\!\!\!\perp Y|P_{S^{(W)}_{Y|\circ \mathbf{X}}}\mathbf{X}P_{S^{(W)}_{Y|\mathbf{X}\circ}}$, *then* $S^{(W)}_{Y|\circ \mathbf{X}} \supseteq S_{Y|\circ \mathbf{X}}$, $S^{(W)}_{Y|\mathbf{X}\circ} \supseteq S_{Y|\mathbf{X}\circ}$, *and thus* $S^{(W)}_{Y|\circ \mathbf{X}\circ} \supseteq S_{Y|\circ \mathbf{X}\circ}$.

(b) *If* $W \perp\!\!\!\perp Y|\mathbf{X}$, *then* $S^{(W)}_{Y|\circ \mathbf{X}\circ} \subseteq S_{Y|\circ \mathbf{X}\circ}$.

(c)

$$\begin{aligned} S^{(W)}_{Y|\circ X} &= \oplus^{C}_{w=1} S_{Y_w|\circ X_w}, \\ S^{(W)}_{Y|X\circ} &= \oplus^{C}_{w=1} S_{Y_w|X_w\circ} \text{ and therefore,} \\ S^{(W)}_{Y|\circ X\circ} &= (\oplus^{C}_{w=1} S_{Y_w|X_w\circ}) \otimes (\oplus^{C}_{w=1} S_{Y_w|\circ X_w}). \end{aligned} \tag{2.11}$$

(d)

$$S^{(W)}_{Y|\circ \mathbf{X}\circ} \supseteq \oplus^{C}_{w=1} S_{Y_w|\circ \mathbf{X}_w\circ} = \oplus^{C}_{w=1}(S_{Y_w|\mathbf{X}_w\circ} \otimes S_{Y_w|\circ \mathbf{X}_w}), \tag{2.12}$$

(e) *If* $span(U_w) \subseteq S_{Y_w|vec(\mathbf{X})_w}$ *almost surely. Then if we denote the basis matrix of space* $\oplus^{c}_{w=1} span(U_w)$ *as* U^*, *we have*

$$\epsilon^{\otimes}(U^*) \subseteq S^{(W)}_{Y|\circ \mathbf{X}\circ} \text{ almost surely}, \tag{2.13}$$

and consequently, $S_{\circ U^*} \subseteq \oplus_{w=1}^{c} S_{Y_w|\circ \mathbf{X}_w}$ *and* $S_{U^*\circ} \subseteq \oplus_{w=1}^{c} S_{Y_w|\mathbf{X}_w\circ}$.

(f)

$$
\begin{aligned}
S_{Y|\circ\mathbf{X}} &\subseteq S_{W|\circ\mathbf{X}} \oplus S^{(W)}_{Y|\circ\mathbf{X}}, \\
S_{Y|\mathbf{X}\circ} &\subseteq S_{W|\mathbf{X}\circ} \oplus S^{(W)}_{Y|\mathbf{X}\circ} \text{ and therefore,} \\
S_{Y|\circ\mathbf{X}\circ} &\subseteq (S_{W|\mathbf{X}\circ} \oplus S^{(W)}_{Y|\mathbf{X}\circ}) \otimes (S_{W|\circ\mathbf{X}} \oplus S^{(W)}_{Y|\circ\mathbf{X}}).
\end{aligned}
\tag{2.14}
$$

Part (a) and part (b) of Proposition 2 conclude with the general relationships between the *partial folding subspace* and the *marginal folding subspace*. That is, the two spaces do not always preserve one way containing relationship. Thus by estimating *marginal folding subspace* $S_{Y|\circ\mathbf{X}\circ}$ as *partial folding subspace* $S^{(W)}_{Y|\circ\mathbf{X}\circ}$, one could be estimating a either larger or a smaller space.

Based on part (c) of Proposition 2 that $S^{(W)}_{Y|\circ X\circ} = (\oplus_{w=1}^{C} S_{Y_w|X_w\circ}) \otimes (\oplus_{w=1}^{C} S_{Y_w|\circ X_w})$, we propose an individual direction ensemble method to estimate $\oplus_{w=1}^{C} S_{Y_w|X_w\circ}$, $\oplus_{w=1}^{C} S_{Y_w|\circ X_w}$, and thus $S^{(W)}_{Y|\circ X\circ}$. Part (d) of Proposition 2 concludes that $\oplus_{w=1}^{C}(S_{Y_w|\mathbf{X}_w\circ} \otimes S_{Y_w|\circ\mathbf{X}_w})$ is a subspace of $S^{(W)}_{Y|\circ\mathbf{X}\circ}$, i.e, estimation on $\oplus_{w=1}^{C}(S_{Y_w|\mathbf{X}_w\circ} \otimes S_{Y_w|\circ\mathbf{X}_w})$ does not recover $S^{(W)}_{Y|\circ\mathbf{X}\circ}$ exhaustively. We propose ordinary least square type ensemble method to estimate $\oplus_{w=1}^{C}(S_{Y_w|\mathbf{X}_w\circ} \otimes S_{Y_w|\circ\mathbf{X}_w})$ in the later section. Finally, based on part (e) of Proposition 2, we propose a third method called objective function optimization method, which estimate $S^{(W)}_{Y|\circ\mathbf{X}\circ}$ through a direct ordinary least square formulation and does not involve embellishing *partial left folding subspace* (*partial right folding subspace*).

2.5 Estimation methods

In this section, we propose three numerical algorithms based upon sliced inverse regression for estimating *partial folding subspace*. We call it partial folded sliced inverse regression, or simply, partial folded-SIR. Note that, each estimation method could be easily extended to

partial folded-SAVE partial folded-DR etc., and we will demonstrate such extensions at the end of each estimation method.

2.5.1 Individual direction ensemble method

Part (c) of Proposition 2 concludes that $S^{(W)}_{Y|\circ X\circ} = (\oplus^C_{w=1} S_{Y_w|X_w\circ}) \otimes (\oplus^C_{w=1} S_{Y_w|\circ X_w})$. Based on this equivalent relationship, we propose an individual direction ensemble method to estimate $\oplus^C_{w=1} S_{Y_w|X_w\circ}$ and $\oplus^C_{w=1} S_{Y_w|\circ X_w}$, respectively, and thus $S^{(W)}_{Y|\circ X\circ}$.

The idea of individual direction ensemble methods is similar to Outer Product of the Gradient method (OPG; Xia et al, 2002), where we first estimate the partial left-folding (right-folding) SIR directions α (β) by ensembling all the conditional left-folding direction α_w (β_w). We then use $\beta \otimes \alpha$ as the estimation of the basis directions of *partial folding subspace*. The algorithm can be developed as the following:

1. For each category $W = w$, estimate the basis matrix of *conditional left-folding subspace* $S_{Y_w|\circ X_w}$ as $\hat{\alpha}_w$, and basis matrix of *conditional right-folding subspace* $S_{Y_w|X_w\circ}$ as $\hat{\beta}_w$ using Li, Kim and Altman's (2010) folded-SIR algorithm.

2. Ensemble all the conditional left-folding SIR directions $\hat{\alpha}_w$ as partial left-folding SIR directions $\hat{\Theta}_\alpha = \sum^C_{w=1} \frac{n_w}{n} \hat{\alpha}_w \hat{\alpha}^T_w$, then apply eigen-value decomposition on $\hat{\Theta}_\alpha$ and extract the first d_l eigen-vectors as the partial left-folding SIR direction $\hat{\alpha}$.

3. Do the similar estimation for $\hat{\beta}$, i.e, ensemble all the conditional right-folding SIR directions $\hat{\beta}_w$ as partial left-folding SIR directions $\hat{\Theta}_\beta = \sum^C_{w=1} \frac{n_w}{n} \hat{\beta}_w \hat{\beta}^T_w$, then apply eigen-value decomposition on $\hat{\Theta}_\beta$ to extract the first d_r eigen-vectors as the partial right-folding S.I.R direction $\hat{\beta}$.

4. Use space spanned by the columns of $\hat{\beta} \otimes \hat{\alpha}$ as the estimation of basis matrix for *partial folding subspace*.

By changing the estimation $S_{Y_w|\circ X_w}$ in step 1 using Li, Kim and Altman's (2010) folded-SAVE and folded-DR algorithm, one can extend this method to estimate partial folded-SAVE, partial folded-DR etc.

2.5.2 Ordinary least square ensemble method

Part (d) of Proposition 2 concludes that $\oplus_{w=1}^{C}(S_{Y_w|\mathbf{X}_w\circ} \otimes S_{Y_w|\circ\mathbf{X}_w})$. Thus, by estimating $\oplus_{w=1}^{C}(S_{Y_w|\mathbf{X}_w\circ} \otimes S_{Y_w|\circ\mathbf{X}_w})$, we are targeting a subspace of $S^{(W)}_{Y|\circ\mathbf{X}\circ}$, i.e, estimation on $\oplus_{w=1}^{C}(S_{Y_w|\mathbf{X}_w\circ} \otimes S_{Y_w|\circ\mathbf{X}_w})$ may not recover $S^{(W)}_{Y|\circ\mathbf{X}\circ}$ exhaustively. We propose an ordinary leas square ensemble method to estimate $\oplus_{w=1}^{C}(S_{Y_w|\mathbf{X}_w\circ} \otimes S_{Y_w|\circ\mathbf{X}_w})$ in the this section. We follow Cook (2004) to formulate an OLS type solution to estimate α and β. First observe that

$$S^{(W)}_{Y|\circ\mathbf{X}\circ} = span\left\{\beta^{(W)} \otimes \alpha^{(W)}\right\} \supseteq \oplus_{w=1}^{C} S_{Y_w|\circ X_w\circ} = \oplus_{w=1}^{C} span\left\{\beta_w \otimes \alpha_w\right\}, \tag{2.15}$$

which means $\forall w = 1, ..., C$, we can find a matrix $f(w) \in p_l \times p_r$ such that $\beta_w \otimes \alpha_w = (\beta^{(W)} \otimes \alpha^{(W)})f(w)$. Thus in terms of population solution, we could minimize

$$E_W||(\beta \otimes \alpha)f(W) - \beta_W \otimes \alpha_W||, \tag{2.16}$$

over α, β and projection matrix $f(W)$ with constraint that $\alpha^T\alpha = I_{d_l}$ and $\beta^T\beta = I_{d_r}$ to obtain estimate for *partial folding subspace*. Before we propose the numerical algorithm for sample estimation, we introduce three lemmas that help explain how the the algorithm is developed. Lemma 1 and Lemma 3 can be found in Li, Kim and Altman (2010) and Lemma 2 can be found in Magnus and Neudecker (1999).

Lemma 1 *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be matrices in* $\mathbb{R}^{r_1 \times r_2}$ *and* $\mathbb{R}^{r_3 \times r_4}$*, where* $r_1, ..., r_4$ *are positive integers. Then,*

$$vec(\mathbf{A} \otimes \mathbf{B}) = \Pi[vec(\mathbf{A}) \otimes vec(\mathbf{B})], \tag{2.17}$$

where $\Pi = I_{r_2} \otimes [(I_{r_4} \otimes K_{r_1,r_3})K_{r_3 r_4, r_1}]$.

Lemma 2 *Let* $\mathbf{A}$*,* $\mathbf{X}$ *and* $\mathbf{B}$ *be matrices in* $\mathbb{R}^{r_1 \times r_2}$*,* $\mathbb{R}^{r_2 \times r_3}$ *and* $\mathbb{R}^{r_3 \times r_4}$*, where* $r_1, ..., r_4$ *are positive integers. Then,*

$$vec(\mathbf{AXB}) = (B^T \otimes \mathbf{A})vec(\mathbf{X}). \tag{2.18}$$

Lemma 3 *If* $\mathbf{V}_1$ *is an* r_1 *dimensional random vector,* V_2 *is an* $r_1 \times r_2$ *dimensional random matrix, each having finite second moments, then the minimizer of* $E||\mathbf{V}_1 - \mathbf{V}_2\mathbf{c}||^2$ *over all* $c \in \mathbb{R}^{r_2}$ *is* $c^* = [E({\mathbf{V}_2}^T\mathbf{V}_2)]^{-1}E({\mathbf{V}_2}^T\mathbf{V}_1)$.

Therefore, we can estimate β, α and $f(W)$ by the minimizer of

$$E_W||\beta_W \otimes \alpha_W - (\beta \otimes \alpha)f(W)||^2. \tag{2.19}$$

Rewrite the objective function as

$$E||vec(\beta_W \otimes \alpha_W) - vec[(\beta \otimes \alpha)f(W)]||^2, \tag{2.20}$$

and that,

$$\begin{aligned} vec[(\beta \otimes \alpha)f(W)] &= (f^T(W) \otimes I_{p_l p_r})vec(\beta \otimes \alpha) \\ &= (f^T(W) \otimes I_{p_l p_r})\Pi[vec(\beta) \otimes vec(\alpha)]. \end{aligned} \tag{2.21}$$

According to Lemma 1, $\Pi = I_{d_r} \otimes [(I_{d_l} \otimes \mathbf{K}_{p_r,p_l})\mathbf{K}_{p_l d_l, p_r}]$, such that $vec(\beta \otimes \alpha) = \Pi[vec(\beta) \otimes vec(\alpha)]$. The definitions of communication matrices $\mathbf{K}_{p_r,p_l}$ and $\mathbf{K}_{p_l d_l, p_r}$ can be found in

Magnus and Neudecker (1999). Based on Lemma 2,

$$\begin{aligned} vec(\beta) \otimes vec(\alpha) &= vec[vec(\alpha)vec^T(\beta)] \\ &= [I_{p_r d_r} \otimes vec(\alpha)]vec(\beta). \end{aligned} \tag{2.22}$$

Therefore, combining equations (2.21) and (2.22), we can rewrite the objective function (2.20) as

$$E||vec(\beta_W \otimes \alpha_W) - (f^T(W) \otimes I_{p_l p_r})\Pi[I_{p_r d_r} \otimes vec(\alpha)]vec(\beta)]||. \tag{2.23}$$

Thus if we denote

$$\mathbf{V}_1 = vec(\beta_W \otimes \alpha_W), \qquad \mathbf{V}_2 = (f^T(W) \otimes I_{p_l p_r})\Pi[I_{p_r d_r} \otimes vec(\alpha)], \tag{2.24}$$

then the minimizer of (2.19) over $\beta \in R^{p_r \times d_r}$ for fixed $f \in L_2^{d_l d_r, c}$ and $\alpha \in R^{p_l \times d_l}$is given by

$$vec(\beta) = [E(\mathbf{V}_2^T \mathbf{V}_2)]^{-1} E(\mathbf{V}_2^T \mathbf{V}_1). \tag{2.25}$$

Similarly, we can find the minimizer of (2.19) over $\alpha \in R^{p_l \times d_l}$ for fixed $f \in L_2^{d_l d_r, c}$ and $\beta \in R^{p_r \times d_r}$ is given by

$$vec(\alpha) = [E(\mathbf{V}_2^T \mathbf{V}_2)]^{-1} E(\mathbf{V}_2^T \mathbf{V}_1), \tag{2.26}$$

where $V_1(W) = vec(\beta_W \otimes \alpha_W)$, $V_2(W) = (f^T(W) \otimes I_{p_l p_r})\Pi[vec(\beta) \otimes I_{p_l d_l}]$. To summarize the above results, the algorithm can then be developed as follows:

1. For each category $W = w$, estimate the directions for conditional left-folding space $S_{Y_w|\circ X_w}$ as $\hat{\alpha}_w$, and directions for conditional right-folding space $S_{Y_w|X_w\circ}$ as $\hat{\beta}_w$ using Li, Kim and Altman's (2010) folded-SIR algorithm.

2. Generate initial value of $\hat{\alpha}^{(0)} \in \mathbb{R}^{p_l \times d_l}$ (or equivalently, $vec(\alpha_1)^{(0)}$) and $\left\{\hat{f_0(w)} : w = 1, ..., c\right\} \in \mathbb{R}^{d_l d_r \times d_l d_r}$ from a sample of $N(0, 1)$ variables.

   For $k \geq 0$,

3. For $w = 1, ..., c$, compute $\hat{p}_w = \frac{n_w}{n}$ and

$$\hat{\mathbf{V}}_1(w) = vec(\hat{\beta}_w \otimes \hat{\alpha}_w),$$
$$\hat{\mathbf{V}}_2(w) = (f^T(w) \otimes I_{p_l p_r})\Pi[\hat{vec(\alpha)}^{(k)} \otimes I_{p_r d_r}].$$

   Then compute $\hat{vec(\beta)}^{(k+1)}$ by

$$[\sum_{w=1}^{c} \hat{p}_w \hat{\mathbf{V}}_2^T(w)\hat{\mathbf{V}}_2(w)]^{-1}[\sum_{w=1}^{c} \hat{p}_w \hat{\mathbf{V}}_2^T(w)\hat{\mathbf{V}}_1(w)]. \tag{2.27}$$

4. For $w = 1, ..., c$, recompute $\hat{\mathbf{V}}_2(w)$ as

$$\hat{\mathbf{V}}_2(w) = (f^T(w) \otimes I_{p_l p_r})\Pi[I_{p_l d_l} \otimes \hat{vec(\beta)}^{(k+1)})].$$

   Then compute $\hat{vec(\alpha)}^{(k+1)}$ by the same equation as $\hat{vec(\beta)}^{(k+1)}$ but with an updated $\hat{\mathbf{V}}_2(w)$, that is, $\hat{vec(\alpha)}^{(k+1)}$ is

$$[\sum_{w=1}^{c} \hat{p}_w \hat{\mathbf{V}}_2^T(w)\hat{\mathbf{V}}_2(w)]^{-1}[\sum_{w=1}^{c} \hat{p}_w \hat{\mathbf{V}}_2^T(w)\hat{\mathbf{V}}_1(w)]. \tag{2.28}$$

5. For $w = 1, ..., c$, recompute $\hat{\mathbf{V}}_2(w)$ as

$$\hat{\mathbf{V}}_2(w) = I_{d_l d_r} \otimes [\hat{\beta}^{(k+1)} \otimes \hat{\alpha}^{(k+1)}],$$

then compute

$$\hat{f}(w) = (\hat{\mathbf{V}}_2(w)^T \hat{\mathbf{V}}_2(w))^{-1} [\hat{\mathbf{V}}_2(w)^T \hat{\mathbf{V}}_1(w)].$$

6. Iterate from step 2 to step 5, until

$$\sum_{w=1}^{c} \hat{p}_w || \hat{\beta}_w \otimes \hat{\alpha}_w - (\hat{\beta}^{(k)} \otimes \hat{\alpha}^{(k)}) \hat{f}(w) ||,$$

is smaller than some pre-specified threshold ϵ.

By changing the estimation $S_{Y_w|\circ X_w}$ in step 1 using Li, Kim and Altman's (2010) folded-SAVE and folded-DR algorithm, one can extend this method to estimate partial folded-SAVE, partial folded-DR etc.

2.5.3 Objective function optimization method

Instead of ensembling the directions of *conditional left (right) folding subspace*, one could also estimate $\beta \otimes \alpha$ directly through an OLS type formulation, in order to reduce the aggregated errors when estimating the basis matrices β_w and α_w for their corresponding *conditional folding subspaces* respectively. Part (e) of Proposition 2 concludes that we can estimate *partial folding subspace* $S^{(W)}_{Y|\circ \mathbf{X}_\circ}$ by targeting on the Kronecker envelope of U^* where U^* is the basis matrix of $\oplus_{w=1}^{c} span(U_w)$. We propose the following theorem which lays out the theoretical foundations for our third estimation method.

First we clarify the notations as follows: random matrix $U_W \in (p_l p_r) \times k_W$ follows the previous definition, and discrete variable W has finite possible values with distribution $0 < P(W = w) = p_w < 1$. Let $\alpha_0 \in p_l \times d_l$ and $\beta_0 \in p_r \times d_r$ be the basis matrices of $S_{\circ U^*}$ and $S_{U^* \circ}$ which span the Kronecker envelope of U^* with respect to integer pair (p_l, p_r). For positive integers k_1 and k_2, and a random vector Z defined on Ω_Z, let ${L_2}^{k_1 \times k_2}(\Omega)$ be the

class of functions $f : Z \to R^{k_1 \times k_2}$ such that $E||f(Z)||^2 < \infty$ and $||.||$ is the Frobenius norm of a matrix.

Theorem 1 *Suppose that the for each $W = w$, elements of U_w have finite variances and are measurable with respect to a random vector Z and that A is a $p_l p_r \times p_l p_r$ nonrandom and nonsingular matrix. Let $(\alpha^*, \beta^*, (f_1^*, ..., f_C^*))$ be the minimizer of*

$$E||AU_W - A(\beta \otimes \alpha) f_W(Z)||^2, \tag{2.29}$$

over all $\alpha \in p_l \times d_l$, $\beta \in p_r \times d_r$ and $f_w \in {L_2}^{d_l d_r \times k_w}(\Omega_Z)$ for each $W = w$. Then

$$span(\beta^* \otimes \alpha^*) = \epsilon^{\otimes}(U^*). \tag{2.30}$$

Proof:

Using double expectation formula, we can further write the objective function as

$$E_W(E_{U_w}||AU_w - A(\beta \otimes \alpha) f_w(Z)|W = w||^2), \tag{2.31}$$

where the inside expectation is with respect to random matrices $U_1, ..., U_C$ and the outside expectation is with respect to categorical variable W. This is equivalent to

$$\sum_{w=1}^{C} p_w (E||AU_w - A(\beta \otimes \alpha) f_w(Z)|W = w||^2). \tag{2.32}$$

Assume $\epsilon^{\otimes}(U^*) = span(\beta_0 \otimes \alpha_0)$. Since for each $W = w$, $span(\beta_0 \otimes \alpha_0) = \epsilon^{\otimes}(U^*) \supseteq \oplus_{w=1}^{c} span(U_w) \supseteq U_w$ and the elements of U_w are measurable with respect to Z, there exists a random projection matrix $\phi_w(Z) \in L^{d_l d_r \times k_w}$ such that $U_w = (\beta_0 \otimes \alpha_0)\phi_w(Z)$, which is equivalent to

$$AU_w = A(\beta_0 \otimes \alpha_0)\phi_w(Z). \tag{2.33}$$

Thus (2.29), or equivalently (2.32), reaches its minimum 0 within the range of $(\alpha, \beta, f_1, ..., f_C)$ given in the theorem. This implies that any minimizer $(\alpha^*, \beta^*, f_1^*, ..., f_C^*)$ of (2.29) must satisfy $A(\beta^* \otimes \alpha^*) f_w^*(Z) = AU_w$ almost surely for every $W = w$ and, consequently,

$$(\beta_0 \otimes \alpha_0)\phi_w(Z) = (\beta^* \otimes \alpha^*) f_w{}^*(Z), \tag{2.34}$$

almost surely. But this means that $span(\beta^* \otimes \alpha^*)$ contains each U_w almost surely, thus we also have $span(\beta^* \otimes \alpha^*) \supseteq \oplus_{w=1}^{c} span(U_w)$. Since $span(\beta^* \otimes \alpha^*)$ has the same dimensions as $\epsilon^{\otimes}(U^*)$, the theorem now follows from the uniqueness of the Kronecker envelope. $\square$.

Now we will connect the solution for minimizing the above expectation with results from Li, Kim and Altman (2010). First if we still use the equivalent formula (2.35) that we need to minimize the following function:

$$\sum_{w=1}^{C} p_w (E||AU_w - A(\beta \otimes \alpha) f_w(Z)|W = w||^2). \tag{2.35}$$

Then by using the following notation

$$U^* = [\sqrt{p_1}U_1, \sqrt{p_2}U_2, ..., \sqrt{p_C}U_C], f(Z)^* = [\sqrt{p_1}f_1(Z), ..., \sqrt{p_C}f_C(Z)],$$

we can rewrite (2.35) as $E||AU^* - A(\beta \otimes \alpha) f(Z)^*||^2$. Then, following Li, Kim and Altman (2010), we implement their algorithm to obtain estimation on β and α directly from the objective function. Note that Li, Kim and Altman (2010) proposed a simplified estimation method for folded-SIR when dimension of $f(Y)$ is $d_l d_r \times 1$. However, in our case, since $f(Y)^{(W)}$ has dimension $d_l d_r \times c$, such simplification could not be applied here. In summary, the third algorithm can then be developed as the following. We only demonstrate the detailed algorithm for partial folded SIR, extensions for SAVE and DR can be easily derived

by replacing with appropriate U_w.

First, we estimate pooled covariance matrix Σ_{pool} through vectorized data

$$\hat{\Sigma}_{pool} = n^{-1} \sum_{i=1}^{n} vec(\mathbf{X}_i - \bar{\mathbf{X}}) vec^T(\mathbf{X}_i - \bar{\mathbf{X}}).$$

Similarly, we estimate covariance matrix Σ_w for each category $W = w$ as

$$\hat{\Sigma}_w = n_w^{-1} \sum_{i=1}^{n} vec(\mathbf{X}_i - \bar{\mathbf{X}}) vec^T(\mathbf{X}_i - \bar{\mathbf{X}}) | W = w, \qquad w = 1, 2, ..., c.$$

We discretize response variable Y. Let $J_1, ..., J_s$ be the partition of Ω_Y and let $D = \delta(Y)$ be the discrete random variable defined by

$$\delta(Y) = \ell \quad if \quad Y \in J_\ell, \ell = 1, ..., s.$$

For a function h of $(\mathbf{X}, Y)$, let $E_n h(\mathbf{X}, Y)$ denote the sample average $n^{-1} \times \sum_{i=1}^{n} h(\mathbf{X}_i, Y_i)$. The formal objective function optimization method is described as follows:

1. Generate initial value of $\hat{\alpha}^{(0)} \in \mathbb{R}^{p_l \times d_l}$ (or equivalently, $\hat{v}ec(\alpha)^{(0)}$) from the individual direction ensemble method stated previously. Then generate $d_l d_r \times c$ random matrix $\left\{ \hat{f_0}(\ell) : \ell = 1, ..., s \right\}$ from a sample of $N(0, 1)$ variables.

   For $k \geq 0$,

2. For $\ell = 1, ..., s$, compute $\hat{p}_\ell = E_n[I(D = \ell)]$ and

$$\begin{aligned}
\hat{\mathbf{V}}_1(\ell) &= \hat{p}_\ell^{-1} vec\{\Sigma_{pool}^{\frac{1}{2}} [\sqrt{P(W=1)} \Sigma_1^{-1} E_n[vec(X_{W=1}) I(D = \ell)], \\
&\qquad ..., \sqrt{P(W=c)} \Sigma_c^{-1} E_n[vec(X_{W=c}) I(D = \ell)]]\}, \\
\hat{\mathbf{V}}_2(\ell) &= (f^T(\hat{\ell})^{(k)} \otimes \Sigma_{pool}^{\frac{1}{2}}) \Pi [I_{p_r d_r} \otimes vec(\hat{\alpha})^{(k)}].
\end{aligned}$$

Then compute $\hat{vec}(\beta)^{(k+1)}$ by

$$[\sum_{\ell=1}^{s} \hat{p}_\ell \hat{\mathbf{V}}_2^T(\ell)\hat{\mathbf{V}}_2(\ell)]^{-1}[\sum_{\ell=1}^{s} \hat{p}_\ell \hat{\mathbf{V}}_2^T(\ell)\hat{\mathbf{V}}_1(\ell)]. \tag{2.36}$$

3. For $\ell = 1, ..., s$, recompute $\hat{\mathbf{V}}_2(\ell)$ as

$$\hat{\mathbf{V}}_2(\ell) = (f^T\hat{(\ell)}^{(k)} \otimes \Sigma_{pool}^{\frac{1}{2}})\Pi[\hat{vec}(\beta)^{(k+1)}) \otimes I_{p_l d_l}].$$

Then compute $\hat{vec}(\alpha)^{(k+1)}$ by the same equation as $\hat{vec}(\beta)^{(k+1)}$, but with an updated $\hat{\mathbf{V}}_2(\ell)$, that is, $\hat{vec}(\alpha)^{(k+1)}$ is

$$[\sum_{\ell=1}^{s} \hat{p}_\ell \hat{\mathbf{V}}_2^T(\ell)\hat{\mathbf{V}}_2(\ell)]^{-1}[\sum_{\ell=1}^{s} \hat{p}_\ell \hat{\mathbf{V}}_2^T(\ell)\hat{\mathbf{V}}_1(\ell)]. \tag{2.37}$$

4. For $\ell = 1, ..., s$, recompute $\hat{\mathbf{V}}_2$ as

$$\hat{\mathbf{V}}_2 = I_c \otimes [\Sigma_{pool}^{\frac{1}{2}} \quad \hat{\beta}^{(k+1)} \otimes \hat{\alpha}^{(k+1)}],$$

then compute

$$\hat{f(l)}^{(k+1)} = (\hat{\mathbf{V}}_2^T\hat{\mathbf{V}}_2)^{-1}[\hat{\mathbf{V}}_2^T\hat{\mathbf{V}}_1(\ell)].$$

5. Iterate from step 2 to step 5, until

$$||P_{\hat{\beta}^{(k+1)}\otimes\hat{\alpha}^{(k+1)}} - P_{\hat{\beta}^{(k)}\otimes\hat{\alpha}^{(k)}}||,$$

is smaller than some pre-specified threshold ϵ.

By plugging in different U_w in $U^* = \left[\sqrt{p_1}U_1, \sqrt{p_2}U_2, ..., \sqrt{p_c}U_c\right]$, and $f_w(Z)$ in $f(Z)^* = \left[\sqrt{p_1}f_1(Z), ..., \sqrt{p_c}f_c(Z)\right]$ in the objective function using Li, Kim and Altman's (2010) folded-

SAVE and folded-DR formulation, one can extend this method to estimate partial folded-SAVE, partial folded-DR etc. For partial folded-SIR, partial folded-SAVE and partial folded-DR, the respective U_w's are:

$$\begin{aligned} SIR : U_w &= \Sigma_w^{-1} E[vec(\mathbf{X})|Y, W = w], \\ SAVE : U_w &= \Sigma_w^{-1}[\Sigma_w - var[vec(\mathbf{X})|Y, W = w]\Sigma_w^{-\frac{1}{2}}, \\ DR : U_w &= \Sigma_w^{-1}\left\{2\Sigma_w - E[(vec(\mathbf{X}) - var(\tilde{\mathbf{X}}))(vec(\mathbf{X}) - var(\tilde{\mathbf{X}}))^T|Y, \tilde{Y}, W = w]\right\}\Sigma_w^{-\frac{1}{2}}. \end{aligned} \tag{2.38}$$

And for partial folded-SIR, partial folded-SAVE and partial folded-DR, the $f(Z_w)$'s are:

$$\begin{aligned} SIR : f_w(Z) &= f_w(Y) \text{ is a random } d_l d_r \text{ - dimenstional vector,} \\ SAVE : f_w(Z) &= f_w(Y) \text{ is a random } d_l d_r \times p_l p_r \text{ matrix,} \\ DR : f_w(Z) &= f_w(Y, \tilde{Y}) \text{ is a random } d_l d_r \times p_l p_r \text{ matrix.} \end{aligned} \tag{2.39}$$

2.6 Estimation of structural dimensions

Previously, we regard the structural dimensions ${d_l}^{(W)}$ and ${d_r}^{(W)}$ for the associated *partial folding subspace* as known quantities. In reality, these dimensions need to be inferred from the data. One can utilize the first estimation: individual direction ensemble method, to construct an empirical method for estimating structural dimensions ${d_l}^{(W)}$ and ${d_r}^{(W)}$.

First of all, denote that for each category $W = w$, the estimated basis matrix for *conditional left-folding subspace* $S_{Y_w|\circ X_w}$ as $\hat{\alpha}_w$, and basis matrix of *conditional right-folding subspace* $S_{Y_w|X_w\circ}$ as $\hat{\beta}_w$ using Li, Kim and Altman's (2010) folded-SIR algorithm. The ensembled directions as defined as:

$$\hat{\Theta}_\alpha = \sum_{w=1}^{C} \frac{n_w}{n} \hat{\alpha}_w \hat{\alpha}_w^T, \ \hat{\Theta}_\beta = \sum_{w=1}^{xuC} \frac{n_w}{n} \hat{\beta}_w \hat{\beta}_w^T, \tag{2.40}$$

and the descending ordered eigenvalues for $\hat{\Theta}_\alpha$ as $(\hat{\lambda}_1, ..., \hat{\lambda}_{p_l})$ and eigenvalues for $\hat{\Theta}_\beta$ as $(\hat{\phi}_1, ..., \hat{\phi}_{p_r})$. From there, one can compute ratio of two consecutive eigenvalues as $\hat{r}_i = \hat{\lambda}_i/\hat{\lambda}_{i+1}$ for $i = 1, ..., p_l - 1$ and $\hat{q}_j = \hat{\phi}_j/\hat{\phi}_{j+1}$ for $j = 1, ..., p_r - 1$. Then we adopt the maximal eigenvalue ration criterion implemented in contour projected dimension reduction (MERC; Luo, Wang and Tsai, 2009) and (Li and Yin, 2009) to estimate structural dimensions as:

$$\begin{aligned} \hat{d_l}^{(W)} &= argmax_{1 \le i \le d_{l\,max}} \{\hat{r}_i\} \\ \hat{d_r}^{(W)} &= argmax_{1 \le j \le d_{r\,max}} \{\hat{q}_j\} \end{aligned} \tag{2.41}$$

As suggested by Luo, Wang and Tsai, one can usually set the maximum possible structural dimensions as $d_{lmax} = d_{rmax} = 5$. In practice, one can also rely on graphical tools to help determine structural dimensions. By plotting the eigen-values ratios $\hat{\lambda}_i/\hat{\lambda}_{i+1}$ (or $\hat{\phi}_i/\hat{\phi}_{i+1}$) as Y-axis against the index of the pair as X-axis, one can look for the elbow of the plot and estimate the structural dimensions according to the position of the elbow.

2.7 Numerical studies

Our numerical studies consist of two parts. The first part includes Example 2.1 and Example 2.2, where response variable Y is continuous and generated from forward models. Response variable in the second part from Example 2.3 to Example 2.4 is discrete with two levels and generated from inverse models. Throughout the four examples, we assume the categorical variable $W = 0, 1$, is independent of $\mathbf{X}$ and Y, and follows a binomial distribution with success probability 0.5. The proposed three estimation methods perform differently in different simulation settings and details on their performances, as well as possible reasonings are also provided in this section. For abbreviation, we call three estimations individual direction ensemble method as individual ensemble method, ordinary least square ensemble method as OLS method, and objective function optimization method as objective function method.

2.7.1 Part I, continuous Y, forward model

In this part, examples are modified based on the examples from Xue and Yin (2014), where continuous response variable Y is generated from forward models. In addition, we have a categorical variable W with two levels 0 and 1. In Example 2.1 and Example 2.2, we set *conditional central subspaces* as proper subspaces of *conditional folding subspaces*, i.e, for $w = 0, 1$,

$$S_{Y_w|vec(\mathbf{X})_w} \subsetneq S_{Y_w|\circ\mathbf{X}_w\circ}. \tag{2.42}$$

Therefore, the *partial central subspace* $S^{(W)}_{Y|vec(\mathbf{X})}$ is also a proper subspace of *partial folding subspace* $S^{(W)}_{Y|\circ X\circ}$ because,

$$S^{(W)}_{Y|vec(\mathbf{X})} = \oplus^1_{w=0} S_{Y_w|vec(\mathbf{X})_w} \subsetneq \oplus^1_{w=0} S_{Y_w|\circ\mathbf{X}_w\circ} \subseteq S^{(W)}_{Y|\circ X\circ}. \tag{2.43}$$

For such simulation settings, methods such as partial sliced inverse regression (partial SIR; Chiaromonte, Cook and Li, 2002) which targets at the *conditional central subspace* $S^{(W)}_{Y|vec(\mathbf{X})}$ of the vectorized predictor $vec(\mathbf{X})$ (Chiaromonte, Cook and Li, 2002), can not recover the *partial folding subspace* $S^{(W)}_{Y|\circ\mathbf{X}\circ}$ exhaustively. Comparison between the proposed three estimation methods and partial SIR are not recommended given that they are targeting different subspaces. We compare numerical results only based on the proposed three estimation methods all estimating *partial folding subspace*. Example 2.1 and Example 2.2 differ by how the two *conditional folding subspaces* overlap with each other, with Example 2.1 having two subspaces being exact the same, and Example 2.2 with two complete orthogonal subspaces.

Let predictor $\mathbf{X}$ be a $p_l = p_r = 5$ matrix, and $\mathbf{X}_{ij}$ stands for the element on the ith row and jth column. Its vectorized version follows a standard normal distribution: $vec(\mathbf{X}) \sim MVN_{p_l p_r}(\mathbf{0}, I_{p_l p_r})$.The error term ϵ is independent of $\mathbf{X}$ and follows a standard

normal distribution as well. To compute the accuracy of the partial folded SIR with categorical variable, we follow the same criterion as Li, Zha and Chiaromonte (2005) by using the Frobenius form between estimated project matrix and underlying true projection matrix,

$$||P_{\hat{\beta}\otimes\hat{\alpha}} - P_{\beta\otimes\alpha}||_F,$$

where $||\cdot||_F$ standards for the Frobenius norm with respect to usual inner product. By specifying different sample size $n = 200, 400, 600, 800, 1000$ and 1600, Table 2.1 and Table 2.2 summarize the results of Example 2.1 and Example 2.2 with average estimation errors and their standard deviation using three types of estimation methods based on 100 iterations of simulations. They are also compared with benchmark performance where the matrices α_0 and β_0 are obtained by randomly and independently generating $vec(\alpha_0)$ and $vec(\beta_0)$ from standard normal distribution, then calculate the Frobenius norm as $||P_{\beta_0\otimes\alpha_0} - P_{\beta\otimes\alpha}||_F$. We describe the details of the two examples as follows:

Example 2.1

The conditional distribution of Y is as follows:

$$Y = \mathbf{X}_{11} \times (\mathbf{X}_{12} + \mathbf{X}_{21} + 1) + 0.2 \times \epsilon \; for \; W = 0,$$
$$Y = \mathbf{X}_{22} \times (\mathbf{X}_{12} + \mathbf{X}_{21} + 1) + 0.2 \times \epsilon \; for \; W = 1.$$

In this example, *partial folding subspace* is exactly the same as two corresponding *conditional folding subspace*.

$$\begin{aligned} S^{(W)}_{Y|\circ X\circ} &= S_{Y_{w=0}|\circ\mathbf{X}_{w=0}\circ} = S_{Y_{w=1}|\circ\mathbf{X}_{w=1}\circ} = S_{Y_{w=0}|\circ\mathbf{X}_{w=0}\circ} \oplus S_{Y_{w=1}|\circ\mathbf{X}_{w=1}\circ} \\ &= span(e_1\otimes e_1, e_1\otimes e_2, e_2\otimes e_1, e_2\otimes e_2). \end{aligned} \tag{2.44}$$

Therefore, all three estimation methods can recover $S^{(W)}_{Y|\circ X\circ}$ exhaustively. However, for vectorized data, its corresponding *partial central subspace* is a proper subset of *partial folding subspace*.

$$\begin{aligned} S_{Y_{w=0}|vec(\mathbf{X})_{w=0}} &= span(e_1 \otimes e_1, e_1 \otimes e_2 + e_2 \otimes e_1) \subsetneq S_{Y_{w=0}|\circ \mathbf{X}_{w=0}\circ}, \\ S_{Y_{w=1}|vec(\mathbf{X})_{w=1}} &= span(e_2 \otimes e_2, e_1 \otimes e_2 + e_2 \otimes e_1) \subsetneq S_{Y_{w=1}|\circ \mathbf{X}_{w=1}\circ}, \end{aligned} \tag{2.45}$$

thus

$$S^{(W)}_{Y|vec(\mathbf{X})} = span(e_1 \otimes e_1, e_1 \otimes e_2 + e_2 \otimes e_1, e_2 \otimes e_2) \subsetneq S^{(W)}_{Y|\circ X\circ}. \tag{2.46}$$

The results in Table 2.1 indicate that among the proposed three methods, individual ensemble method and OLS method both perform much better than benchmark measure, as well as compared with the objective function method. As sample size increases, all three methods steadily improve their performance. Objective function method with or without pooled variance provide similar accuracy compared with the other two methods when sample size is small, but improve slower when sample size increases.

Table 2.1: Example 2.1, accuracy of estimates on partial folding subspace

n	Bench Mark	Individual ensemble	OLS	Objective	Objective (pooled)
200		1.9928 (0.3540)	2.0090 (0.3845)	2.0267 (0.3735)	1.9949 (0.3112)
400		1.4318 (0.4069)	1.4370 (0.4235)	1.5484 (0.3922)	1.5503 (0.4013)
600	2.5894	1.0036 (0.2966)	0.9965 (0.2981)	1.1935 (0.3407)	1.2176 (0.3119)
1000		0.7105 (0.2046)	0.7103 (0.2057)	0.9018 (0.2675)	0.9605 (0.2626)
1600		0.4584 (0.1110)	0.4580 (0.1110)	0.7025 (0.1779)	0.7186 (0.1881)

Example 2.2

In the second example, we modify the previous example so that the two *conditional folding subspaces* are completely orthogonal with each other. The conditional distribution of Y given

$\mathbf{X}$ and W is:

$$Y = \mathbf{X}_{11} \times (\mathbf{X}_{12} + \mathbf{X}_{21} + 1) + 0.2 \times \epsilon \; for \; W = 0,$$
$$Y = \mathbf{X}_{33} \times (\mathbf{X}_{34} + \mathbf{X}_{43} + 1) + 0.2 \times \epsilon \; for \; W = 1.$$

In this case,

$$S_{Y_{w=0}|\circ \mathbf{X}_{w=0}\circ} = span(e_1, e_2) \otimes span(e_1, e_2) = span(e_1 \otimes e_1, e_1 \otimes e_2, e_2 \otimes e_1, e_2 \otimes e_2)$$
$$S_{Y_{w=1}|\circ \mathbf{X}_{w=1}\circ} = span(e_3, e_4) \otimes span(e_3, e_4) = span(e_3 \otimes e_3, e_3 \otimes e_4, e_4 \otimes e_3, e_4 \otimes e_4),$$

Based on part (c) of Proposition 2, we have:

$$\begin{aligned} S^{(W)}_{Y|\circ X\circ} &= (span(e_1, e_2) \oplus span(e_3, e_4)) \otimes (span(e_1, e_2) \oplus span(e_3, e_4)) \\ &= span(e_i \otimes e_j) \; i, j = 1, ..., 4 \end{aligned}$$

On the other hand, based on part (d) of Proposition 2, we have:

$$\begin{aligned} S_{Y|\circ X_{w=0}\circ} \oplus S_{Y|\circ X_{w=1}\circ} &= (span(e_1, e_2) \otimes span(e_1, e_2)) \oplus (span(e_1, e_2) \oplus span(e_3, e_4)) \\ &= span(e_1 \otimes e_1, e_1 \otimes e_2, e_2 \otimes e_1, e_2 \otimes e_2, e_3 \otimes e_3, e_3 \otimes e_4, e_4 \otimes e_3, e_4 \otimes e_4) \\ &\subsetneq S^{(W)}_{Y|\circ X\circ}. \end{aligned}$$

The second OLS method only targets at $S_{Y|\circ X_{w=0}\circ} \oplus S_{Y|\circ X_{w=1}\circ}$, which is a proper subspace of our desired space $S^{(W)}_{Y|\circ X\circ}$ according to part (d) of Proposition 4 and the experiment setting. For vectorized data, we could also derive their *conditional central subspaces* and *partial central subspaces* for the vectorized predictor as proper subspaces of the corresponding *conditional folding subspaces* and *partial folding subspaces*.

$$\begin{aligned} S_{Y_{w=0}|vec(\mathbf{X})_{w=0}} &= span(e_1 \otimes e_1, e_1 \otimes e_2 + e_2 \otimes e_1) \subsetneq S_{Y_{w=0}|\circ \mathbf{X}_{w=0}\circ} \\ S_{Y_{w=1}|vec(\mathbf{X})_{w=1}} &= span(e_3 \otimes e_3, e_3 \otimes e_4 + e_4 \otimes e_3) \subsetneq S_{Y_{w=1}|\circ \mathbf{X}_{w=1}\circ}, \end{aligned} \tag{2.47}$$

Table 2.2: Example 2.2, accuracy of estimates on partial folding subspace

n	Bench Mark	Individual ensemble	OLS	Objective	Objective (pooled)
200		2.8420 (0.6133)	2.8857 (0.5835)	3.1280 (0.4645)	3.1580 (0.4780)
400		2.2741 (0.7187)	2.3071 (0.7305)	3.0424 (0.5441)	3.1002 (0.5327)
600	3.3844	1.7783 (0.7106)	1.7465 (0.6960)	2.9843 (0.5206)	3.0103 (0.5324)
1000		1.0698 (0.3741)	1.0919 (0.4277)	2.8106 (0.6115)	2.7481 (0.6649)
1600		0.7958 (0.3014)	0.7961 (0.3023)	2.6598 (0.7274)	2.6990 (0.7179)

therefore,

$$\begin{aligned} S^{(W)}_{Y|vec(\mathbf{X})} &= span(e_1 \otimes e_1, e_1 \otimes e_2 + e_2 \otimes e_1, e_3 \otimes e_3, e_3 \otimes e_4 + e_4 \otimes e_3) \\ &\subsetneq S_{Y|\circ X_{w=0}\circ} \oplus S_{Y|\circ X_{w=1}\circ} \subsetneq S^{(W)}_{Y|\circ X\circ}. \end{aligned} \tag{2.48}$$

Though in theory OLS method does not estimate *partial folding subspace* exhaustively, it still provides similar results compared with individual ensemble method. This is probably due to the fact that it uses initial values from the results of individual ensemble method.

2.7.2 Part II, discrete Y, inverse model

In the second part, the response variable Y is set to be discrete, and is generated from inverse model. The two new examples are modified from Li, Kim and Altman (2010), where we add one more categorical variable W. They share the same two *partial folding subspaces* as the previous Example 2.1 and Example 2.2, correspondingly.

Example 2.3

The example is constructed from an inverse model: let $d_l = d_r = 2$ and $p_l = p_r = p = 5$. The response Y follows Bernoulli distribution with success probability $\pi = 0.5$. The conditional distribution of vectorized data $vec(\mathbf{X})$ given Y follows multivariate normal distribution with different means and variances with different categories $W = w,\ w = 0, 1$.

When $W = 0$,

$$E(\mathbf{X}|Y = 0, W = 0) = \mathbf{0}_{p\times p}, \quad E(\mathbf{X}|Y = 1, W = 0) = \begin{pmatrix} \mu\mathbf{I}_2 & \mathbf{0}_{2\times(p-2)} \\ \mathbf{0}_{(p-2)\times 2} & \mathbf{0}_{(p-2)\times(p-2)} \end{pmatrix}.$$

Here $\mu = 1.5$ and $\mathbf{0}_{r\times s}$ is an $r \times s$ matrix with all elements equal to 0. The conditional variances of each element of $\mathbf{X}$ is set to be

$$var(X_{ij}|Y = 0, W = 0) = \begin{cases} \sigma^2 & (i,j) \in A \\ 1 & (i,j) \notin A \end{cases},$$

$$var(X_{ij}|Y = 1, W = 0) = \begin{cases} \tau^2 & (i,j) \in A \\ 1 & (i,j) \notin A \end{cases}.$$

Here $\sigma = 0.1$, $\tau = 1.5$ and A is the index set $\{(1,2),(2,1)\}$. We also assume $vec(\mathbf{X})$ has covariance $cov(X_{ij}, X_{i'j'}) = 0$, when $(i,j) \neq (i',j')$. When $W = 1$, it follows the exact same setting as $W = 0$. Then from Bayes theorem, we can derive the conditional probability $P(Y = 0|\mathbf{X}, W = 0)$ (or equivalently, $P(Y = 1|\mathbf{X}, W = 0)$) is a function of $X_{11} + X_{22}$, X_{12}^2 and X_{21}^2, and so does conditional posterior probability $P(Y = 1|X, W = 1)$ (or equivalently, $P(Y = 1|\mathbf{X}, W = 1)$. The smallest sub-matrix containing the elements is

$$\begin{pmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} \\ \mathbf{X}_{21} & \mathbf{X}_{22} \end{pmatrix}.$$

Therefore, the two *conditional folding subspaces*, as well as the desired *partial folding subspace* are all the same,

$$\begin{aligned} S^{(W)}_{Y|\circ X\circ} &= S_{Y_{w=0}|\circ\mathbf{X}_{w=0}\circ} = S_{Y_{w=1}|\circ\mathbf{X}_{w=1}\circ} = S_{Y_{w=0}|\circ\mathbf{X}_{w=0}\circ} \oplus S_{Y_{w=1}|\circ\mathbf{X}_{w=1}\circ} \\ &= span(e_1 \otimes e_1, e_1 \otimes e_2, e_2 \otimes e_1, e_2 \otimes e_2). \end{aligned} \tag{2.49}$$

Table 2.3: Example 2.3, accuracy of estimates on partial folding subspace

n	Bench Mark	Individual ensemble	OLS	Objective	Objective (pooled)
200		0.9905 (0.2339)	0.9913 (0.2353)	1.6270 (0.4409)	1.4212 (0.4283)
400		0.6349 (0.1404)	0.6349 (0.1402)	0.8305 (0.1969)	0.8082 (0.2120)
600	2.5894	0.5012 (0.1015)	0.5012 (0.1015)	0.5858 (0.1245)	0.6031 (0.1391)
1000		0.3763 (0.0749)	0.3763 (0.0749)	0.4175 (0.0897)	0.4277 (0.0865)
1600		0.2906 (0.0542)	0.2906 (0.0542)	0.3039 (0.0624)	0.3134 (0.0660)

In this case, all three estimation methods can recover $S^{(W)}_{Y|\circ X\circ}$ exhaustively. However, for vectorized data,

$$\begin{aligned} S_{Y_{w=0}|vec(\mathbf{X})_{w=0}} &= span(e_1\otimes e_1+e_2\otimes e_2, e_1\otimes e_2, e_2\otimes e_1) \subsetneq S_{Y_{w=0}|\circ\mathbf{X}_{w=0}\circ}, \\ S_{Y_{w=1}|vec(\mathbf{X})_{w=1}} &= span(e_1\otimes e_1+e_2\otimes e_2, e_1\otimes e_2, e_2\otimes e_1) \subsetneq S_{Y_{w=1}|\circ\mathbf{X}_{w=1}\circ}, \end{aligned} \tag{2.50}$$

thus

$$S^{(W)}_{Y|vec(\mathbf{X})} = span(e_1\otimes e_1+e_2\otimes e_2, e_1\otimes e_2, e_2\otimes e_1) \subsetneq S^{(W)}_{Y|\circ X\circ}. \tag{2.51}$$

The results are listed in Table 2.3, where the proposed individual ensemble method and OLS method outperform the third estimation method objective function method. But their performance gap is closing in when sample size n increases.

Example 2.4

For example 2.4, its corresponding two *conditional folding subspaces* are completely orthogonal, leading to the largest *partial folding subspace* overall.

For $W = 0$, it follows exact same setting as in Example 2.3. For $W = 1$, however, the

Table 2.4: Example 2.4, accuracy of estimates on partial folding subspace

n	Bench Mark	Individual ensemble	OLS	Objective	Objective (pooled)
200		1.5467 (0.3854)	1.5464 (0.3853)	1.8733 (0.5051)	1.1086 (0.2908)
400		0.9871 (0.2605)	0.9873 (0.2605)	1.1736 (0.2996)	0.7344 (0.1787)
600	3.3844	0.7996 (0.2024)	0.7996 (0.2024)	0.9048 (0.2343)	0.5818 (0.1352)
1000		0.6191 (0.1555)	0.6191 (0.1555)	0.6961 (0.1759)	0.4617 (0.1066)
1600		0.5013 (0.1407)	0.5013 (0.1407)	0.5607 (0.1427)	0.3725 (0.1071)

condition mean of $\mathbf{X}$ given Y is changed to:

$$E(\mathbf{X}|Y=1,W=1)=\begin{pmatrix}\mathbf{0}_{2\times 2} & \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times(p-4)}\\ \mathbf{0}_{2\times 2} & \mu\mathbf{I}_2 & \mathbf{0}_{2\times(p-4)}\\ \mathbf{0}_{(p-4)\times 2} & \mathbf{0}_{(p-4)\times 2} & \mathbf{0}_{(p-4)\times(p-4)}\end{pmatrix}.$$

Correspondingly, the conditional covariance structure stay the same as Example 3 except the index set $A=\{(3,4),(4,3)\}$. We can easily verify that the desired *partial folding subspace* $S^{(W)}_{Y|\circ X\circ}$ is the same as in Example 2.2. But for vectorized data $vec(\mathbf{X})$,

$$\begin{aligned}S_{Y_{w=0}|vec(\mathbf{X})_{w=0}} &= span(e_1\otimes e_1+e_2\otimes e_2, e_1\otimes e_2, e_2\otimes e_1)\subsetneq S_{Y_{w=0}|\circ\mathbf{X}_{w=0}\circ},\\ S_{Y_{w=1}|vec(\mathbf{X})_{w=1}} &= span(e_3\otimes e_3+e_4\otimes e_4, e_3\otimes e_4, e_4\otimes e_3)\subsetneq S_{Y_{w=1}|\circ\mathbf{X}_{w=1}\circ},\end{aligned} \tag{2.52}$$

thus

$$S^{(W)}_{Y|vec(\mathbf{X})}\subsetneq S^{(W)}_{Y|\circ X\circ}. \tag{2.53}$$

The results are listed in Table 2.4, similarly, the proposed individual ensemble method and OLS method outperform the third estimation method objective function optimization method, while the best model still remains to be objective function method with pooled covariance. Figure 2.1 summarizes the four examples. The three estimation methods then can be "loosely" interpreted as this: Since *partial folding subspace* $S^{(W)}_{Y|\circ X\circ}$ with its basis matrix

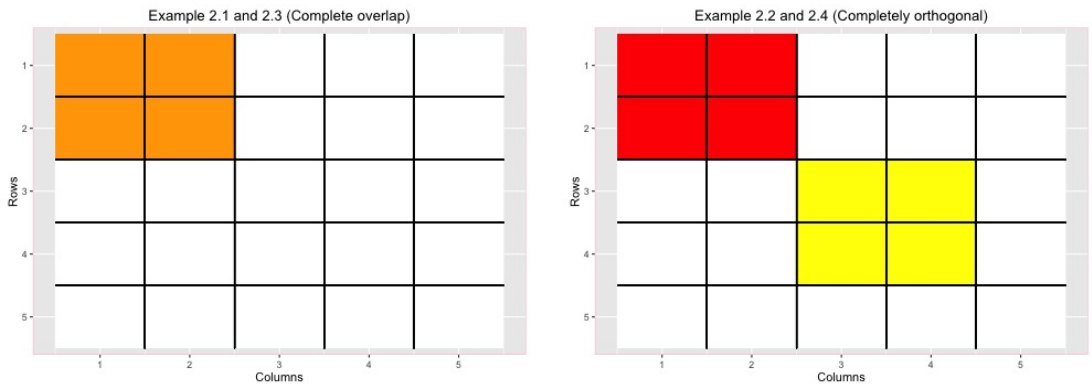

Figure 2.1: Summary of conditional and partial folding subspaces in four examples

must be presented as a Kronecker product, it can only be covered by a "rectangle space". Therefore, exhaustive methods including individual ensemble method and objective function method attempt to find one minimal "rectangle space" that covers both of the *conditional folding subspaces*; On the other hand, OLS method estimates $\oplus S_{Y|\circ X_w \circ}$, which look for two minimal "rectangles" that cover all the *conditional folding subspaces*, thus can be smaller than *partial folding subspace*. Traditional *partial central subspace* $S^{(W)}_{Y|vec(\mathbf{X})}$, which stack the columns together, and its estimation method partial SIR looks for "blocks" which cover all the *conditional central subspaces*.

Our take-aways from Example 2.1 - Example 2.4 are that:

- Individual ensemble method and OLS method outperform objective function method in forward and continuous models, but objective function method with pooled variance performs much better than the other two methods in inverse model and discrete response variable setting. This is probably because that folded-SIR (Li, Kim and Altman, 2010) itself is developed as an inverse model, and the fact that discrete response also ease the difficulty of choosing appropriate number of slices because each slice would

naturally contain observations belonging to the same category.

- All three methods are influenced by how overlap the two *conditional folding subspaces* are, the more they overlap, the fewer parameters need to be estimated, thus higher accuracy. The smaller the overlap the two spaces are, the better the objective function method with pooled variance is, compared with individual ensemble method and OLS method in the inverse models. We infer that because of when two *conditional folding subspaces* are more or less different, the errors of estimating their corresponding basis matrices will aggregate when applying individual ensemble or OLS methods. However, for the objective function method with pooled variance, we do not need to estimate *conditional folding subspaces* at all, thus no errors are aggregated at this stage. Instead, we form the optimization function directly which directly aims at estimating *partial folding subspace*.

- As sample size increases, all three methods increase their accuracy as well as decrease their standard errors.

2.8 Application

In this section, we apply our proposed three numerical methods to estimate *partial folding subspace* with a longitudinal Primary Biliary Cirrhosis (PBC) dataset provided by Mayo Clinic. The dataset was collected in a follow up experiment conducted between 1974 and 1984 with 312 patientients participated in the experiment. Its detailed description can be found in Fleming and Harrington (1991) and Murtaugh et al. (1994). The dataset can be found in the "survival" package named as "pbcseq" in computing software R. PBC is a progressive cholesteric liver disease on adults which may result in severe liver failure, need of transplantation and even death. (Talwalkar and K.D. Lindor, 2003); (Xue and Yin, 2014).

This disease is related to various bio-markers including biliribin, albumin and prothrombin time. In this PBC dataset for each subject, repeated measurements of the these bio-markers and other demographic information are collected at different visit time points. In this article we are particular interested how these bio-markers, correlated with time and other demographic information, may influence on patients' transplantation time (or death time). We denote the variables in this analysis as:

Response variable: Time between registration and the earlier of transplanting or death (In Years).

Predictor variable: 3×4 matrix predictor, where columns of the predictor represents the discrete visit time from subjects. Similar to Xue and Yin (2012), we regard visits between 90 days and 270 days as 6-month, 270 - 550 days as 1 year, 550 - 910 days as 2 years and 910 -1275 as 3 years. The rows of the predictor matrix consist of three types of bio-markers including bilirubin, albumin and prothrombin time.

Categorical variable: gender including males and females.

Note that several sufficient dimension reduction and sufficient dimension folding methods have been devoted to analyze longitudinal data. Li and Yin (2009) developed sufficient dimension reduction framework where time is regarded as discrete random variable. As a extension to the partial ordinary least square method from Li, Cook and Chiaromonte (2002), a partial ordinary least square method was proposed to estimate the *partial central mean subspaces*, and all the subspaces are at last ensembled together as a final estimate. Despite its straightforward formulation, the proposed method neglects the inner correlation of the measure bio-markers among different time points.. A similar method called longitudinal sliced

inverse regression (LSIR; Pfeiffer, Forzani and Bura, 2011) was developed to extend sliced inverse regression in the longitudinal data context. Xue and Yin (2014) applied folded-MAVE and folded-OPG under the framework of sufficient dimension folding, which treat time as one fold dimension of the variable, and the bio-markers quantities as the other fold. In this way, correlation across different time points are taken into consideration. However, this method cannot involve categorical predictor information such as gender. Our method simultaneously reduce the dimension of matrix predictors while considering categorical variable information. We begin our analysis by estimating the structural dimensions ${d_l}^{(W)}$ and ${d_r}^{(W)}$ for the associated *partial folding subspace*. Since Xue and Yin (2014) have demonstrated

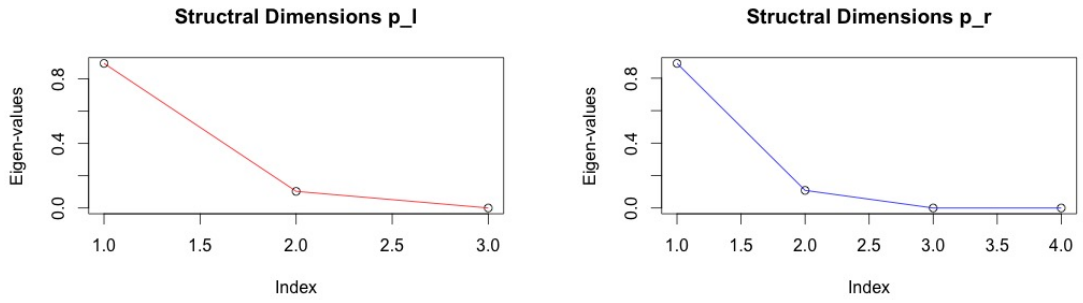

Figure 2.2: Eigenvalue plots for estimating structural dimensions

that the structural dimensions for *marginal folding subspace* are d_l and $d_r = 1$, we remain to use the same dimensions $d_{lw} = 1$ and $d_{rw} = 1$ $\forall w = 0, 1$, for the associated *conditional folding subspace* within each gender sub-population. We then ensemble the directions from each sub-population and implement eigen-value decomposition based on the ensembled left folding matrix and right folding matrix. The eigen-value plot suggest that the structural dimensions for the *partial folding subspace* are also ${d_l}^{(W)} = 1$ and ${d_r}^{(W)} = 1$.

Table 2.5 summarized the estimated directions from three algorithms. We observe that:

- The first two estimation methods: individual direction ensemble method and ordinary least square method provide very similar results on the produced *partial folding*

subspace, while the last method objective function optimization method yield slightly different direction estimates.

- The estimated *partial folding subspace* is essentially very similar to the *marginal folding subspace* that without considering gender information, or simply the *conditional folding subspace* for female subgroups. This is due to the fact the female subjects dominate the majority experiment subjects (128 females v.s. 16 males). We notice that for the other male subgroup, its *conditional folding subspace* is very different from the previous ones. Our results also are consistent with the results from Xue and Yin (2016) where folded-OPG and folded-MAVE are utilized to estimate *marginal folding subspace*.

Table 2.5: Estimated directions from three methods

Direction	Individual Ensemble	OLS	Objective function	Male	Female	Marginal
$\hat{\alpha}_1$	0.0253	0.0498	0.0203	0.9319	0.0530	0.0530
$\hat{\alpha}_2$	-0.9943	-0.9955	-0.9934	-0.3577	-0.9955	-0.9948
$\hat{\alpha}_3$	-0.1029	-0.0801	-0.1124	-0.0589	-0.0797	-0.0868
$\hat{\beta}_1$	-0.1127	-0.1067	-0.0220	0.7553	-0.10643	-0.0598
$\hat{\beta}_2$	-0.1301	-0.1127	-0.0406	- 0.4679	-0.1116	-0.0631
$\hat{\beta}_3$	-0.3294	-0.3319	-0.3479	0.2123	-0.3321	-0.2657
$\hat{\beta}_4$	-0.9283	-0.9304	-0.9363	-0.4067	-0.9305	-0.9601

For the three estimation methods, we also compute the confidence intervals for the estimated directions using Bootstrap methods with 200 replications. Among the computed intervals, β_4, α_2 are significantly smaller than 0, α_1 is significantly larger than 0, across all three estimation methods. Three methods do not agree on the estimation of β_3 as it appears not significant in individual ensemble method and OLS method, but significantly smaller than 0 in objective function method. Note that the variability of the estimates is similar between individual ensemble method and OLS method. Objective function methods provide similar variability when estimating β_2, but it also produces narrower interval for β_1, β_3, β_4, α_2 and α_3. The largest inconsistency between objective function method and other methods appears when estimating α_1, as objective function method produces a much smaller coefficients as well as narrower interval.

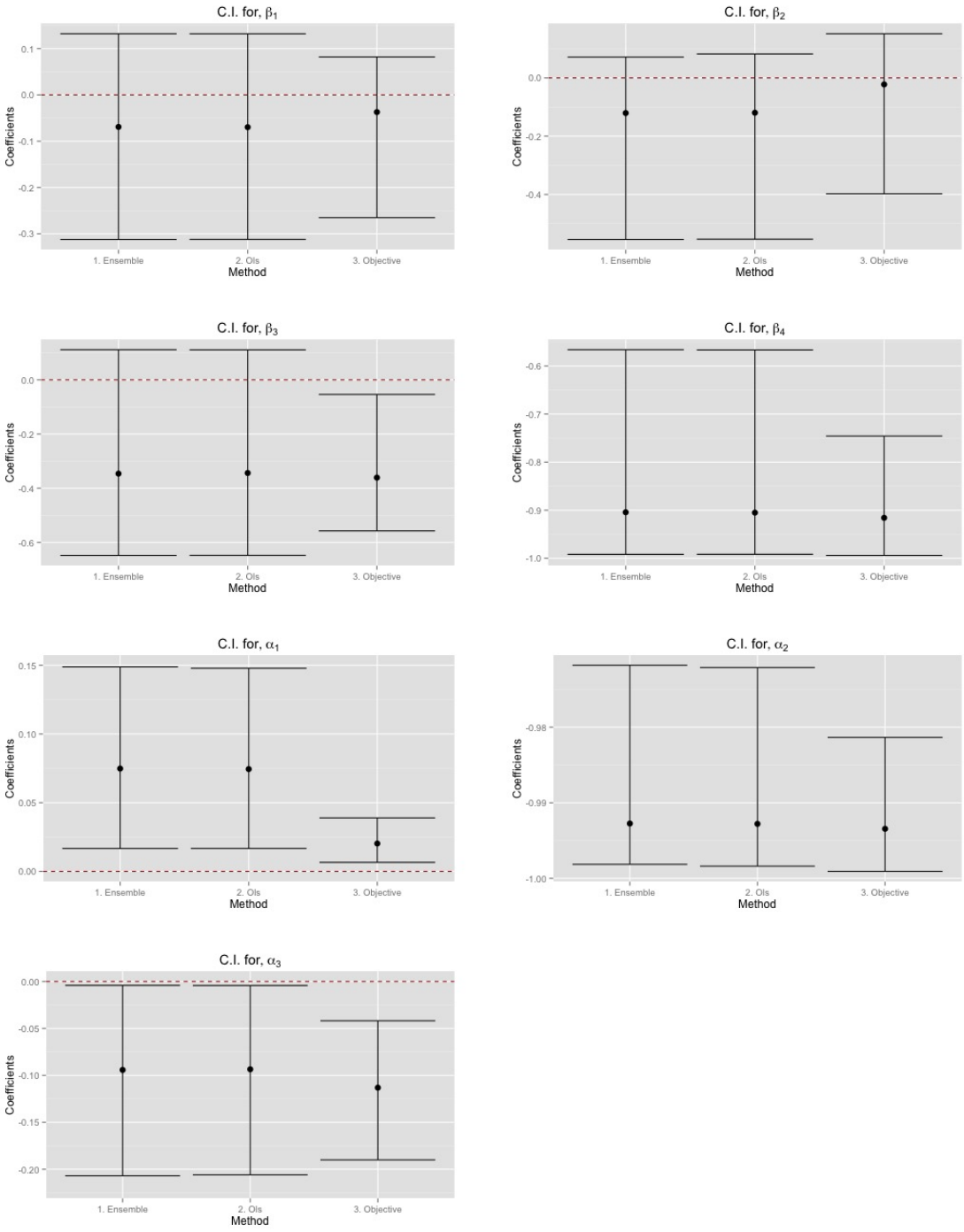

Figure 2.3: Bootstrap confidence intervals for estimated directions

Finally, based on the estimated *partial folding subspace* and two *conditional folding subspaces*, we compute the reduced predictor as ${\alpha^{T}}^{(W)}\mathbf{X}\beta^{(W)} \in \mathbb{R}^{1\times 1}$ for all data points, ${\alpha_{w=0}}^{T}\mathbf{X}\beta_{w=0} \in \mathbb{R}^{1\times 1}$ for 16 males, and ${\alpha_{w=1}}^{T}\mathbf{X}\beta_{w=1} \in \mathbb{R}^{1\times 1}$ for all 128 females, which are all univariate variables. We estimate smoothing splines with these variable against response variable. We adopt a leave one out cross validation to selecting tuning parameter λ in the smoothing spline models. The following three plots provide the details about the fitted models, including the cross-validation procedures to select smoothing penalty $\lambda = 1, 1, 0.25$ for the three model in the first row, the fitted lines in the second row and the residual patterns in the last row. Notice that the fitted lines for both females and for all data points is very similar to results obtained in Xue and Yin (2014), but with smaller magnitude of residuals provided in our model. On the other hand, male group also exhibit an increasing but not necessarily same pattern as females. Due to the limited number of points, the fitted smoothing spline might still over-fit the male observations, lacking a more suitable representation of the correct pattern. We believe the results indicate better performance when using the reduced predictors from the *partial folding subspace* which simultaneously reduce dimensions for predictor matrix, as well as taking into consideration on the categorical variable gender. Most importantly, further investigation are more male data points collections are needed in order to study the pattern of the reduced predictor against the response variable.

2.9 Discussion

In this chapter, we mainly discuss how to implement dimension folding for complex matrix/array predictor with the existence of categorical variables. Aligning with sufficient dimension reduction with categorical variables, we define similar concept *partial folding subspace* for matrix/array predictor against *partial central subspace*. Three types of estimators are proposed to efficiently estimate *partial folding subspace*, which enable us to analyze ma-

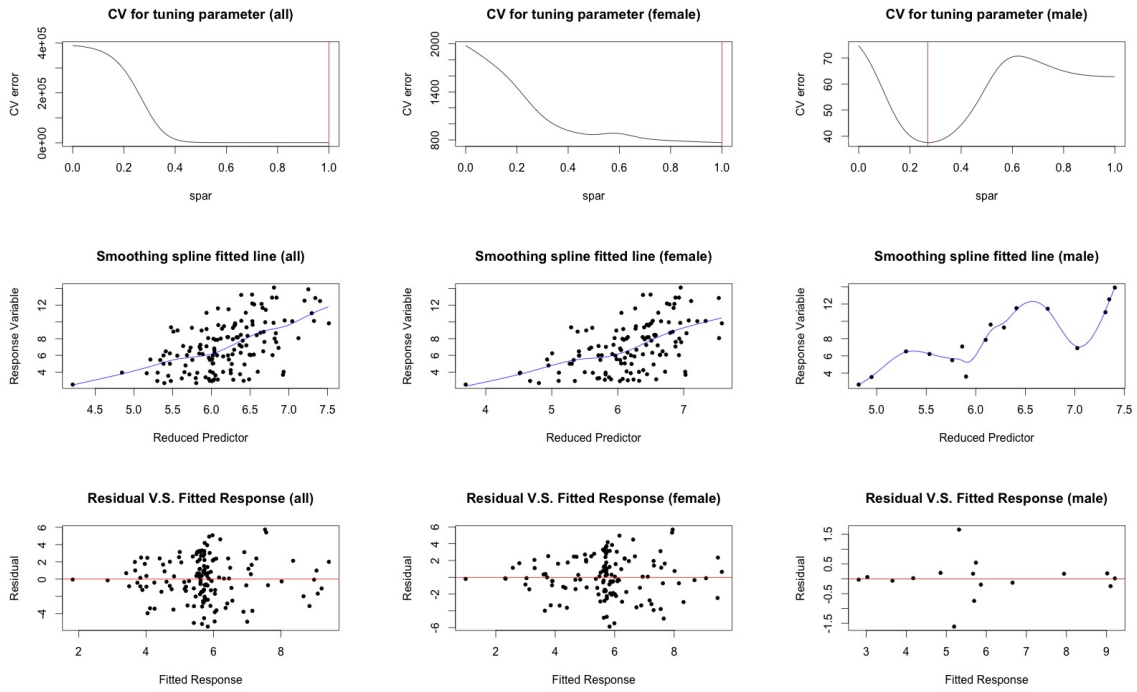

Figure 2.4: Summary of smoothing splines with reduced predictors

trix predictor with categorical variable. Both simulation and data application indicates the proposed *partial folding subspace* provide better insights and summary on how matrix predictors are associated with response variable, combining the additional information from categorical variable.

2.10 Appendix

Proof of Proposition 2 part (a):

In Lemma 4, let

$$\begin{cases} V_1 = vec(\mathbf{X}) \\ V_2 = W \\ V_3 = P_{S_R \otimes S_L} vec(\mathbf{X}) \\ V_4 = Y, \end{cases}$$

and apply the first part of Lemma 4, and the equivalent relationship that $Y \perp\!\!\!\perp \mathbf{X} | \alpha^T \mathbf{X} \beta \Leftrightarrow Y \perp\!\!\!\perp vec(\mathbf{X}) | (\beta \otimes \alpha)^T vec(\mathbf{X})$, we have

$$\begin{aligned} & \mathbf{X} \perp\!\!\!\perp W | (P_{S_L} \mathbf{X} P_{S_R}, Y) \text{ and } \mathbf{X} \perp\!\!\!\perp Y | P_{S_L} \mathbf{X} P_{S_R} \\ \Leftrightarrow\ & vec(\mathbf{X}) \perp\!\!\!\perp W | (P_{S_R \otimes S_L} vec(\mathbf{X}), Y) \text{ and } vec(\mathbf{X}) \perp\!\!\!\perp Y | P_{S_R \otimes S_L} vec(\mathbf{X}) \\ \Leftrightarrow\ & vec(\mathbf{X}) \perp\!\!\!\perp Y | (W, P_{S_R \otimes S_L} vex(\mathbf{X})) \text{ and } vec(\mathbf{X}) \perp\!\!\!\perp W | P_{S_R \otimes S_L} vec(\mathbf{X}) \\ \Leftrightarrow\ & \mathbf{X} \perp\!\!\!\perp Y | (W, P_{S_L} \mathbf{X} P_{S_R}) \text{ and } \mathbf{X} \perp\!\!\!\perp W | P_{S_L} \mathbf{X} P_{S_R}. \end{aligned} \tag{2.54}$$

Therefore, under the assumption that

$$W \perp\!\!\!\perp \mathbf{X} | P_{S^{(W)}_{Y|\circ\mathbf{X}}} \mathbf{X} P_{S^{(W)}_{Y|\mathbf{X}\circ}},$$

$$Y \perp\!\!\!\perp \mathbf{X} | (W, P_{S^{(W)}_{Y|\circ\mathbf{X}}} \mathbf{X} P_{S^{(W)}_{Y|\mathbf{X}\circ}}) \Rightarrow Y \perp\!\!\!\perp P_{S^{(W)}_{Y|\circ\mathbf{X}}} \mathbf{X} P_{S^{(W)}_{Y|\mathbf{X}\circ}}. \tag{2.55}$$

We have $S^{(W)}_{Y|\circ X} \supseteq S_{Y|\circ X}$, $S^{(W)}_{Y|X\circ} \supseteq S_{Y|X\circ}$ and $S^{(W)}_{Y|\circ X\circ} \supseteq S_{Y|\circ X\circ}$.

Now in Lemma 4, let

$$\begin{cases} V_1 = Y \\ V_2 = vec(\mathbf{X}) \\ V_3 = P_{S_R \otimes S_L} vec(\mathbf{X}) \\ V_4 = W, \end{cases}$$

and again apply the first part of Lemma 4, and the equivalent relationship that $Y \perp\!\!\!\perp \mathbf{X} | \alpha^T \mathbf{X} \beta \Leftrightarrow Y \perp\!\!\!\perp vec(\mathbf{X}) | (\beta \otimes \alpha)^T vec(\mathbf{X})$, we have

$$\begin{aligned} & Y \perp\!\!\!\perp W | (P_{S_L} \mathbf{X} P_{S_R}, Y) \text{ and } Y \perp\!\!\!\perp \mathbf{X} | P_{S_L} \mathbf{X} P_{S_R} \\ & \Leftrightarrow \mathbf{Y} \perp\!\!\!\perp W | (P_{S_R \otimes S_L} vec(\mathbf{X}), Y) \text{ and } Y \perp\!\!\!\perp vec(\mathbf{X}) | P_{S_R \otimes S_L} \mathbf{X} \\ & \Leftrightarrow Y \perp\!\!\!\perp vec(\mathbf{X}) | (P_{S_R \otimes S_L} vec(\mathbf{X}), W) \text{ and } Y \perp\!\!\!\perp W | P_{S_R \otimes S_L} vec(\mathbf{X}) \\ & \Leftrightarrow Y \perp\!\!\!\perp \mathbf{X} | (P_{S_L} \mathbf{X} P_{S_R}, W) \text{ and } Y \perp\!\!\!\perp W | P_{S_L} \mathbf{X} P_{S_R}. \end{aligned} \tag{2.56}$$

Therefore, under the assumption that

$$Y \perp\!\!\!\perp W | P_{S^{(W)}_{Y|\circ\mathbf{X}}} \mathbf{X} P_{S^{(W)}_{Y|\mathbf{X}\circ}},$$

$$Y \perp\!\!\!\perp \mathbf{X} | (P_{S^{(W)}_{Y|\circ\mathbf{X}}} \mathbf{X} P_{S^{(W)}_{Y|\mathbf{X}\circ}}, W) \Rightarrow Y \perp\!\!\!\perp \mathbf{X} | P_{S^{(W)}_{Y|\circ\mathbf{X}}} \mathbf{X} P_{S^{(W)}_{Y|\mathbf{X}\circ}}. \tag{2.57}$$

We have $S^{(W)}_{Y|\circ X} \supseteq S_{Y|\circ X}$, $S^{(W)}_{Y|X\circ} \supseteq S_{Y|X\circ}$ and $S^{(W)}_{Y|\circ X\circ} \supseteq S_{Y|\circ X\circ}$.

Proof of Proposition 2 part (b):

In Lemma 4, let

$$\begin{cases} V_1 = Y \\ V_2 = W \\ V_3 = P_{S_R \otimes S_L} vec(\mathbf{X}) \\ V_4 = vec(\mathbf{X}), \end{cases}$$

and apply the first part of Lemma 4, and the equivalent relationship that $Y \perp\!\!\!\perp \mathbf{X}|\alpha^T\mathbf{X}\beta \Leftrightarrow Y \perp\!\!\!\perp vec(\mathbf{X})|(\beta \otimes \alpha)^T vec(\mathbf{X})$, we have

$$\begin{aligned} & Y \perp\!\!\!\perp W|(P_{S_L}\mathbf{X}P_{S_R}, \mathbf{X}) \text{ and } Y \perp\!\!\!\perp \mathbf{X}|P_{S_L}\mathbf{X}P_{S_R} \\ & \Leftrightarrow Y \perp\!\!\!\perp W|(P_{S_R\otimes S_L}vec(\mathbf{X}), vec(\mathbf{X})) \text{ and } Y \perp\!\!\!\perp vec(\mathbf{X})|P_{S_R\otimes S_L}vec(\mathbf{X}) \\ & \Leftrightarrow Y \perp\!\!\!\perp vec(\mathbf{X})|(W, P_{S_R\otimes S_L}vex(\mathbf{X})) \text{ and } Y \perp\!\!\!\perp W|P_{S_R\otimes S_L}vec(\mathbf{X}) \\ & \Leftrightarrow Y \perp\!\!\!\perp \mathbf{X}|(W, P_{S_L}\mathbf{X}P_{S_R}) \text{ and } Y \perp\!\!\!\perp W|P_{S_L}\mathbf{X}P_{S_R}. \end{aligned} \tag{2.58}$$

Therefore, under the assumption that $W \perp\!\!\!\perp Y|\mathbf{X}$, we also have $Y \perp\!\!\!\perp W|(P_{S_{Y|\circ X}}\mathbf{X}P_{S_{Y|X\circ}}, \mathbf{X})$, thus,

$$Y \perp\!\!\!\perp \mathbf{X}|P_{S_{Y|\circ X}}\mathbf{X}P_{S_{Y|X\circ}} \Rightarrow Y \perp\!\!\!\perp \mathbf{X}|(W, P_{S_{Y|\circ X}}\mathbf{X}P_{S_{Y|X\circ}}). \tag{2.59}$$

Hence, $S_{Y|\circ X} \supseteq S^{(W)}_{Y|\circ X}$, $S_{Y|X\circ} \supseteq S^{(W)}_{Y|X\circ}$ and $S_{Y|\circ X\circ} \supseteq S^{(W)}_{Y|\circ X\circ}$.

Proof of Proposition 2 part (c):

For generic subspace S_L and S_R, we have

$$\begin{aligned} Y \perp\!\!\!\perp \mathbf{X}|(P_{S_L}\mathbf{X}P_{S_R}, W) &\Leftrightarrow Y \perp\!\!\!\perp \mathbf{X}|(P_{S_L}\mathbf{X}P_{S_R}, W = w) \quad \forall w = 1, ..., C. \\ &\Leftrightarrow Y \perp\!\!\!\perp vec(\mathbf{X})|P_{(S_R\otimes S_L)}vec(\mathbf{X}), W = w) \quad \forall w = 1, ..., C. \end{aligned} \tag{2.60}$$

Since $S^{(W)}_{Y|\circ\mathbf{X}\circ} = S^{(W)}_{Y|\mathbf{X}\circ} \otimes S^{(W)}_{Y|\circ\mathbf{X}}$, $S^{(W)}_{Y|\circ\mathbf{X}}$ and $S^{(W)}_{Y|\mathbf{X}\circ}$ satisfy the left-handside of equation (2.61) by their definitions, they also satisfy

$$Y \perp\!\!\!\perp \mathbf{X}|(P_{S^{(W)}_{Y|\circ\mathbf{X}}}\mathbf{X}P_{S^{(W)}_{Y|\mathbf{X}\circ}}, W = w) \quad \forall w = 1, ..., C.$$

This implies that $\forall w = 1, ..., C$, $S^{(W)}_{Y|\circ\mathbf{X}} \supseteq S_{Y_w|\circ\mathbf{X}_w}$, $S^{(W)}_{Y|\mathbf{X}\circ} \supseteq S_{Y_w|\mathbf{X}_w\circ}$ and thus $S^{(W)}_{Y|\circ\mathbf{X}} \supseteq \oplus^C_{w=1} S_{Y_w|\circ\mathbf{X}_w}$, $S^{(W)}_{Y|\mathbf{X}\circ} \supseteq \oplus^C_{w=1} S_{Y_w|\mathbf{X}_w\circ}$. Therefore,

$$S^{(W)}_{Y|\circ\mathbf{X}\circ} = S^{(W)}_{Y|\mathbf{X}\circ} \otimes S^{(W)}_{Y|\circ\mathbf{X}} \supseteq (\oplus^C_{w=1} S_{Y_w|\mathbf{X}_w\circ}) \otimes (\oplus^C_{w=1} S_{Y_w|\circ\mathbf{X}_w}).$$

Because $(\oplus^C_{w=1} S_{Y_w|\circ\mathbf{X}_w}) \supseteq S_{Y_w|\circ\mathbf{X}_w}$ and $(\oplus^C_{w=1} S_{Y_w|\mathbf{X}_w\circ}) \supseteq S_{Y_w|\mathbf{X}_w\circ}$ $\forall w = 1, ..., C$, the two direct sum spaces also satisfy the right-hand side of equation (2.61). Therefore, we have

$$Y \perp\!\!\!\perp \mathbf{X} | (P_{\oplus^C_{w=1} S_{Y_w|\circ\mathbf{X}_w}} \mathbf{X} P_{\oplus^C_{w=1} S_{Y_w|\mathbf{X}_w\circ}}, W).$$

This implies the other containing relationship

$$(\oplus^C_{w=1} S_{Y_w|\mathbf{X}_w\circ}) \otimes (\oplus^C_{w=1} S_{Y_w|\circ\mathbf{X}_w}) \supseteq S^{(W)}_{Y|\circ\mathbf{X}\circ}.$$

We then conclude that $S^{(W)}_{Y|\circ\mathbf{X}\circ} = (\oplus^C_{w=1} S_{Y_w|\mathbf{X}_w\circ}) \otimes (\oplus^C_{w=1} S_{Y_w|\circ\mathbf{X}_w})$.

Proof of Proposition 2 part (d):

For generic subspace S_L and S_R, we have

$$\begin{aligned} Y \perp\!\!\!\perp \mathbf{X} | (P_{S_L} \mathbf{X} P_{S_R}, W) &\Leftrightarrow Y \perp\!\!\!\perp \mathbf{X} | (P_{S_L} \mathbf{X} P_{S_R}, W = w) \quad \forall w = 1, ..., C. \\ &\Leftrightarrow Y \perp\!\!\!\perp vec(\mathbf{X}) | P_{(S_R \otimes S_L)} vec(\mathbf{X}), W = w) \quad \forall w = 1, ..., C. \end{aligned} \tag{2.61}$$

Since $S^{(W)}_{Y|\circ\mathbf{X}\circ} = S^{(W)}_{Y|\mathbf{X}\circ} \otimes S^{(W)}_{Y|\circ\mathbf{X}}$, $S^{(W)}_{Y|\circ\mathbf{X}}$ and $S^{(W)}_{Y|\mathbf{X}\circ}$ satisfy the left-handside of equation (2.61) by their definitions, they also satisfy

$$Y \perp\!\!\!\perp \mathbf{X} | (P_{S^{(W)}_{Y|\circ\mathbf{X}}} \mathbf{X} P_{S^{(W)}_{Y|\mathbf{X}\circ}}, W = w) \quad \forall w = 1, ..., C.$$

This implies that $\forall w = 1, ..., C$, $S^{(W)}_{Y|\circ\mathbf{X}} \supseteq S_{Y_w|\circ\mathbf{X}_w}$, $S^{(W)}_{Y|\mathbf{X}\circ} \supseteq S_{Y_w|\mathbf{X}_{w}\circ}$ and thus $S^{(W)}_{Y|\circ\mathbf{X}\circ} = S^{(W)}_{Y|\mathbf{X}\circ} \otimes S^{(W)}_{Y|\circ\mathbf{X}} \supseteq S_{Y_w|\mathbf{X}_{w}\circ} \otimes S_{Y_w|\circ\mathbf{X}_w}$. Therefore,

$$S^{(W)}_{Y|\circ\mathbf{X}\circ} \supseteq \oplus^{C}_{w=1}(S_{Y_w|\mathbf{X}_{w}\circ} \otimes S_{Y_w|\circ\mathbf{X}_w}).$$

Because $\oplus^{C}_{w=1}(S_{Y_w|\mathbf{X}_{w}\circ} \otimes S_{Y_w|\circ\mathbf{X}_w}) \supseteq S_{Y_w|\mathbf{X}_{w}\circ} \otimes S_{Y_w|\circ\mathbf{X}_w}$ and $S_{Y_w|\mathbf{X}_{w}\circ} \otimes S_{Y_w|\circ\mathbf{X}_w}$ satisfy the second right-hand equation (2.61) $\forall w = 1, ..., C$,

$$\oplus^{C}_{w=1}(S_{Y_w|\mathbf{X}_{w}\circ} \otimes S_{Y_w|\circ\mathbf{X}_w}) = span(\mathbf{U}^*) \subseteq S^{(W)}_{Y|\circ\mathbf{X}\circ} = S^{(W)}_{Y|\mathbf{X}\circ} \otimes S^{(W)}_{Y|\circ\mathbf{X}},$$

where $\mathbf{U}^*$ is a random basis matrix of space $\oplus^{C}_{w=1} span(\beta_w \otimes \alpha_w)$ in $\mathbb{R}^{p_l p_r \times k}$. Therefore, by Lemma 2, we have that the Kronecker envelope of $\mathbf{U}^*$with respect to integer p_l and p_r, that is $\epsilon^{\otimes}_{p_l,p_r}(\mathbf{U}^*) = S_{\mathbf{U}^*\circ} \otimes S_{\circ\mathbf{U}^*}$ satisfies the following:

1. $span(\mathbf{U}^*) = \oplus^{C}_{w=1}(S_{Y_w|\mathbf{X}_{w}\circ} \otimes S_{Y_w|\circ\mathbf{X}_w}) \subseteq S_{\mathbf{U}^*\circ} \otimes S_{\circ\mathbf{U}^*}$ almost surely.
2. If there is another pair of subspaces $S_R \in \mathbb{R}^{p_r}$ and $S_L \in \mathbb{R}^{p_l}$ that satisfies condition 1, then $S_{\mathbf{U}^*\circ} \otimes S_{\circ\mathbf{U}^*} \subseteq S_R \otimes S_L$.

However, from the previous proof

$$S^{(W)}_{Y|\circ\mathbf{X}\circ} = S^{(W)}_{Y|\mathbf{X}\circ} \otimes S^{(W)}_{Y|\circ\mathbf{X}} \supseteq \oplus^{C}_{w=1}(S_{Y_w|\mathbf{X}_{w}\circ} \otimes S_{Y_w|\circ\mathbf{X}_w}) = span(\mathbf{U}^*),$$

and by definition, $S^{(W)}_{Y|\mathbf{X}\circ} \in \mathbb{R}^{p_r}$ and $S^{(W)}_{Y|\circ\mathbf{X}} \in \mathbb{R}^{p_l}$. Therefore,

$$\epsilon^{\otimes}_{p_l,p_r}(\mathbf{U}^*) = S_{\mathbf{U}^*\circ} \otimes S_{\circ\mathbf{U}^*} \subseteq S^{(W)}_{Y|\mathbf{X}\circ} \otimes S^{(W)}_{Y|\circ\mathbf{X}} = S^{(W)}_{Y|\circ\mathbf{X}\circ}.$$

On the other hand, for $\forall w = 1, ..., C$,

$$\epsilon^{\otimes}_{p_l,p_r}(\mathbf{U}^*) = S_{\mathbf{U}^*\circ} \otimes S_{\circ\mathbf{U}^*} \supseteq span(\mathbf{U}^*) = \oplus_{w=1}^{C}(S_{Y_w|\mathbf{X}_w\circ} \otimes S_{Y_w|\circ\mathbf{X}_w}) \supseteq S_{Y_w|\mathbf{X}_w\circ} \otimes S_{Y_w|\circ\mathbf{X}_w}.$$

Therefore, $S_{\circ\mathbf{U}^*}$ and $S_{\mathbf{U}^*\circ}$ satisfy the second right-handside of equation (2.61), hence the left-handside of equation (2.61) that

$$Y \perp\!\!\!\perp \mathbf{X} | (P_{S_{\circ\mathbf{U}^*}}\mathbf{X}P_{S_{\mathbf{U}^*\circ}}, W).$$

Thus $S_{\circ\mathbf{U}^*} \supseteq S^{(W)}_{Y|\circ\mathbf{X}}$ and $S_{\mathbf{U}^*\circ} \supseteq S^{(W)}_{Y|\mathbf{X}\circ}$.
This implies the relationship

$$\epsilon^{\otimes}_{p_l,p_r}(\mathbf{U}^*) = S_{\mathbf{U}^*\circ} \otimes S_{\circ\mathbf{U}^*} \supseteq S^{(W)}_{Y|\circ\mathbf{X}\circ}.$$

Therefore

$$\epsilon^{\otimes}_{p_l,p_r}(\mathbf{U}^*) = S_{\mathbf{U}^*\circ} \otimes S_{\circ\mathbf{U}^*} = S^{(W)}_{Y|\circ\mathbf{X}\circ}.$$

This theorem concludes that $S^{(W)}_{Y|\circ\mathbf{X}\circ}$ equals to the Kronecker envelope of $\oplus_{w=1}^{C}(S_{Y_w|\mathbf{X}_w\circ} \otimes S_{Y_w|\circ\mathbf{X}_w})$, thus by estimating $\oplus_{w=1}^{C}(S_{Y_w|\mathbf{X}_w\circ} \otimes S_{Y_w|\circ\mathbf{X}_w})$, we are targeting a proper subspace of $S^{(W)}_{Y|\circ\mathbf{X}\circ}$, i.e, estimation on $\oplus_{w=1}^{C}(S_{Y_w|\mathbf{X}_w\circ} \otimes S_{Y_w|\circ\mathbf{X}_w})$ does not recover $S^{(W)}_{Y|\circ\mathbf{X}\circ}$ exhaustively.

Proof of Proposition 2 part (e):

First note that if for each $W = w$, $span(U_w) \subseteq S_{Y_w|vec(\mathbf{X})_w}$ almost surely, then from part (d)

of Proposition 2, we have

$$\begin{aligned}\oplus_{w=1}^{c} span(U_w) &\subseteq \oplus_{w=1}^{c} S_{Y_w|vec(\mathbf{X})_w} = S^{(W)}_{Y|vec(\mathbf{X})}, \\ &\subseteq \oplus_{w=1}^{c} S_{Y_w|\mathbf{X}_{w\circ}} \otimes S_{Y_w|\mathbf{X}_{w\circ}} \\ &\subseteq S^{(W)}_{Y|\circ\mathbf{X}\circ} = (\oplus_{w=1}^{c} S_{Y_w|\mathbf{X}_{w\circ}}) \otimes (\oplus_{w=1}^{c} S_{Y_w|\circ\mathbf{X}_w}),\end{aligned} \tag{2.62}$$

almost surely. Therefore, by the definition of Kronecker product, we have

$$\begin{aligned} S_{\circ U_{new}} &\subseteq \oplus_{w=1}^{c} S_{Y_w|\circ\mathbf{X}_w}, \\ S_{U_{new}\circ} &\subseteq \oplus_{w=1}^{c} S_{Y_w|\mathbf{X}_{w}\circ}, \ and \\ \epsilon^{\otimes}(U_{new}) = S_{U_{new}\circ} \otimes S_{\circ U_{new}} &\subseteq S^{(W)}_{Y|\circ\mathbf{X}\circ} = (\oplus_{w=1}^{c} S_{Y_w|\mathbf{X}_{w}\circ}) \otimes (\oplus_{w=1}^{c} S_{Y_w|\circ\mathbf{X}_w}). \end{aligned} \tag{2.63}$$

□

Bibliography

[1] F. Chiaronmonte, R. D. Cook and B. Li. Sufficient dimension reduction in regressions with categorical predictors. *The Annals of Statistics* **30(2)**, 475-497, 2002.

[2] R. D. Cook. Graphics for Regressions With a Binary Response. *Journal of the American Statistical Association* **91**, 983-992, 1996.

[3] R. D. Cook. Regression Graphics: Ideas for Studying Regressions Through Graphics. *Wiley*, **New York**, 1998.

[4] R. D. Cook and S. Weisberg. Discussion of "Sliced inverse regression for dimension reduction". *Journal of the American Statistical Association* **86**, 28-33, 1991.

[5] S. Ding and R. D. Cook. Dimension folding PCA and PFC for matrix-valued predictors. *Statistica Sinica* **24**, 463-492, 2014.

[6] S. Ding and R. D. Cook. Tensor sliced inverse regression. *Journal of Multivariate Analysis* **133**, 216-231, 2015.

[7] T. R. Fleming and D. P. Harrington. Counting process and survival analysis. *Wiley, New York*, 1991.

[8] IBM Big Data and Analytics Hub. The Four V's of Big Data. *http://www.ibmbigdatahub.com/infographic/four-vs-big-data*, 2014.

[9] J. R. Magnus and H. Neudecker. Matrix differential calculus with applications in statistics and econometrics, 2nd Edition. *Wiley*, **New York**, 1999.

[10] B. Li, R. D. Cook and F. Chiaromonte. Dimension reduction for the conditional mean in regression with categorical predictors. *The Annals of Statistics* **31**, 1636-1668, 2003.

[11] B. Li, M. Kim and N. Altman. On dimension folding of matrix- or array-valued statistical objects. *The Annals of Statistics* **38**, 1094-1121, 2010.

[12] B. Li and S. Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association* **102**, 2143-2172, 2007.

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

[14] K.-C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**, 316-342, 1991.

[15] L. Li and X. Yin. Longitudinal data analysis using sufficient dimension reduction. *Computational Statistics and Data Analysis* **53**, 4106-4115, 2009.

[16] R. Luo, H. Wang and C. L. Tsai. Contour projected dimension reduction. *The Annals of Statistics* **37**, 3743-3778, 2009.

[17] P. A. Murtaugh, E. R. Dickson, G. M. Van Dam, M. Malinchoc, P. M. Grambsch, A. L. Langworthy and C. H. Gips. Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits. *Hepatology* **20**, 126-136, 1994.

[18] R. M. Pfeiffer, L. Forzani and E. Bura. Sufficient dimension reduction for longitudinally measured predictors. *Statistics in Medicine*, 2011.

[19] J. A Talwalkar and K. D. Lindor. Primary biliary cirrhosis. *The Lancet 362* July 5, 2003.

[20] Y. Xia, H. Tong, W. Li and L. Zhu. An adaptive estimation of dimension reduction. *Journal of Royal Statistical Society, Series B* **64**, 363-410, 2002.

[21] Y. Xue and X. Yin. Sufficient dimension folding for regression mean function. *Journal of Computational and Graphical Statistics* **39**, 1028-1043, 2014.

[22] Y. Xue and X. Yin. Sufficient dimension folding for a functional of conditional distribution of matrix- or array-valued objects. *Journal of Nonparametric Statistics* **27-2**, 253-269, 2015.

Chapter 3

Regularized Sufficient Dimension Folding and Variable Selection

Abstract

Dimension folding methods effectively perform dimension reduction while preserving data's inner structure for complex matrix data such as images and videos. The folding methods, however, would become unstable when sample size n is smaller than the matrix dimensions $p_l \times p_r$, or when collinearity exists among the predictors. In addition, the reduced space consists of linear combinations of rows and columns of the original matrix, thus can be difficult to interpret. In this chapter, based on the least square formulation of folding methods, we develop sufficient dimension folding with regularization to achieve sparse estimations on folding spaces and produce better interpretations. We briefly introduce two types of sparseness existed in matrix predictor data. We develop folding with L_1 regularization which imposes the first type of sparseness, sparseness on the vectorized basis matrices, as well as folded-SIR with L_2 regularization and folded-SIR with both L_1 and L_2 regularizations, which produce the second type of sparseness, coordinate-wise sparse basis matrices that further re-

duce the number of rows and columns in the original predictor matrix. We demonstrate the advantages of our methods over traditional variable selection method such as Lasso on both simulation data and a longitudinal PBC dataset.

3.1 Introduction

Matrix and array data have aroused great interests from researchers recently since more and more data are presented in such a format. In image analysis and facial recognition, based on the marginal distribution of predictor matrix $\mathbf{X}$, a number of dimension reduction methods such as 2DPCA (Yang at al, 2004), $(2D)^2$ PCA (Zhang and Zhou, 2005), Unified PCA (Shan et al, 2008) have been proposed to effectively reduce the dimensionality of the original predictor matrix while preserving its inner matrix structures. These methods, however, were not able to include information from response variable Y, as the core task of image analysis and facial recognition is often more than just reducing the size of the image, but also including tasks such as classification with reduced data.

Li, Kim and Altman (2010) proposed supervised sufficient dimension folding which extends traditional moment-based dimension reduction approaches such as sliced inverse regression (SIR; Li, 1993), sliced average variance estimator (SAVE; Cook and Weisberg, 1991) and directional regression (DR; Li and Wang, 2007) to matrix and array formated predictors. The folding methods simultaneously reduce the dimensionality of matrix or array predictors while preserving the inner matrix or array structure. The estimation of folding methods, such as folded-SIR, requires the estimation of inverse of the covariance matrix Σ_x of the vectorized predictor, i.e., $\Sigma_x = Cov(vec(\mathbf{X}))$. In may application of matrix data analysis, such as facial expression classification, the number of predictors of a matrix $\mathbf{X} \in \mathbb{R}^{p_l \times p_r}$ as $p_l \times p_r$ could easily exceed the sample size n. Moreover, the predictors, in particular, the pixels in the near neighborhood, may be highly correlated, adding more unstableness

when estimating the inverse of covariance matrix of the predictor matrix. In addition, the estimated basis matrices of central folding subspaces are still linear combinations of rows and columns, thus no individual variable selection, or in the case of matrix predictors, row selections and column selections are achieved. In facial expression classification, one would be desired to answer questions such as which specific area or pixels determine humans' facial expression.

In this chapter, motivated by the least square formulation of folding estimators (Li, Kim and Altman, 2010), as well as sparse sufficient dimension reduction (Li, 2007) and sliced inverse regression with regularizations (Li and Yin, 2008), we propose sufficient dimension folding with regularizations. The reminder of this chapter will be organized as the following. In Section 3.2, we begin by reviewing on sparse sufficient dimension reduction (Li, 2007) and sliced inverse regression with regularizations (Li and Yin, 2008). In Section 3.3, we demonstrate the methodology part of sufficient dimension folding with regularizations, including: an introduction on two types of sparseness, namely, sparseness on the vectorized basis matrices, and sparseness on the coordinates of basis matrices. The first type of sparseness corresponds to cases where only a smaller dimensional (in terms of row numbers and column numbers) sub-matrix of the original predictors are associated with response variable. The second type of sparseness corresponds to scenarios where only a few elements in the predictor matrix are associated with the response Y, but the sub-matrix predictors containing these elements are not necessarily a much smaller dimensional matrix compared to the original predictor matrix. We propose sufficient dimension folding with L_1 regularization, and especially folded-SIR with L_1 regularization, to address both type of sparseness. In order to solve the collinearity among predictor elements, and produce better estimation on the first type of sparseness, we further proposed sufficient dimension folding with both L_2 and L_1 regularizations to achieve simultaneous reduction estimation and rows (columns) selections. In Section 3.4, we develop two types of criteria to determine the reduced structural dimensions

d_l and d_r. We demonstrate the effectiveness of the proposed sufficient dimension folding with regularizations methods on both simulate data in Section 3.5 and a longitudinal PBC data in Section 3.6. We conclude this chapter with a short discussion in Section 3.7.

3.2 Review on regularized sufficient dimension reduction and variable selection

3.2.1 Sparse sufficient dimension reduction

Li (2007) proposed sparse sufficient dimension reduction (Sparse SDR; Li, 2007) by converting the estimation of *central subspace* (Cook, 1996) using moment-based inverse estimator such as sliced inverse regression (SIR; Li, 1991) and sliced average variance estimation (SAVE; Cook and Weisberg, 1991), under the framework of generalized eigenvalue decomposition. A generalized eigenvalue decomposition for sufficient dimension reduction has the following form:

$$M\beta_i = \rho_i G\beta_i, i = 1, ..., p, \tag{3.1}$$

where M is a method specific symmetric nonnegative definite kernel matrix (for example $M = cov[E\{X - E(X)|Y\}]$) for sliced inverse regression; G is a symmetric positive definite matrix (in most dimension reduction methods, $G = \Sigma_x$), $\beta_i^T G\beta_j = 1$ for $i \neq j$ and 0 for $i = j$. Then the eigenvectors $\{\beta_1, ..., \beta_d\}$ that corresponds to non-zero eigenvalues $\{\phi_1 \geq ... \geq \phi_d \geq 0\}$ span the central subspace, i.e., $span(\beta_1, ..., \beta_d) \subseteq S_{Y|X}$.

In order to impose sparsity structure on the estimated space, the basis matrix $[\beta_1, ..., \beta_d]$ is estimated through a regression type formulation with a square loss function as its optimization function, so that penalty terms can be added on the loss function later on. If m_i

stands for the ith column of $M^{\frac{1}{2}}$, and β to be a $p \times d$ matrix, then $\hat{\beta}$ can be estimated by

$$\hat{\beta} = argmin_{\beta} \sum_{i=1}^{p} ||G^{-1}m_i - \beta\beta^T m_i||_G^2, \tag{3.2}$$

with constraint that $\beta^T G\beta = I_d$. Then $\hat{\beta}_j = v_j$, $j = 1, ..., d$ (Li, 2007). A regularized eigenvalue decomposition with respect to minimizing the above function, plus a L_1 penalty on each column of the basis matrix β turns out to be complicated in terms of optimization, as shown in Jolliffe, Trendafilov and Uddin. (2003).

Instead, Li (2007) followed the idea the formulation of sparse principal component analysis from Zou, Hastie and Tibshirani (2006) and came up with a more efficient algorithm, with estimator $\hat{\alpha}$ and $\hat{\beta}$ estimated through

$$(\hat{\alpha}, \hat{\beta}) = min_{\alpha\beta}\{\sum_{i=1}^{p} ||G^{-1}m_i - \alpha\beta^T m_i||_G^2 + \lambda_2 trace(\beta^T G\beta) + \sum_{j=1}^{d} \lambda_{1j|\beta_j|_1}\}. \tag{3.3}$$

An alternating minimization algorithm was proposed to estimate α and β and an information type criteria very similar to Akaike information criterion (AIC; Akaike, 1979) was developed to estimate tuning parameter λ_1 and λ_2 as following:

$$\sum_{i=1}^{p} ||G^{-1}m_i - \hat{\beta}_\lambda \hat{\beta}_\lambda^T m_i||_G^2 + 2p_\lambda/n, \tag{3.4}$$

where p_λ is the number of non-zeros components of $\hat{\beta}_\lambda$.

3.2.2 Sliced inverse regression with regularization

Li and Yin (2008) focused their attention on sliced inverse regression (SIR; Li, 1993) by incorporating L_2 regularization into the estimation of SIR called ridge SIR estimator, and adding both L_2 and L_1 regularizations into SIR as sparse ridge estimator. They utilized the

least square formulation of SIR from Cook (2004) that the SIR estimate can be reformulated by minimizing

$$G(B,C) = \sum_{y=1}^{h} \hat{p}_y(\bar{X}_y - \bar{X}) - \hat{\Sigma}_x BC_y)^T \hat{\Sigma}_x^{-1}(\bar{X}_y - \bar{X}) - \hat{\Sigma}_x BC_y), \tag{3.5}$$

over $B \in \mathbb{R}^{p\times d}$ and $C = (C_1, ..., C_h) \in \mathbb{R}^{d\times h}$ under constraint that $B^T B = I_p$. The solution $\hat{B}$ forms the basis of *central subspace* $S_{Y|X}$. To promote regularized estimation on basis matrix B and avoid calculation of inverse of $\hat{\Sigma}_x$, Li and Yin (2008) defined that for a given tuning parameter τ, (B, C) that minimizes

$$G_\tau(B,C) = \sum_{y=1}^{h} \hat{p}_y||(\bar{X}_y - \bar{X}) - \hat{\Sigma}_x BC_y||^2 + \tau vec(B)^T vec(B) \tag{3.6}$$

leads to an estimation of ridge SIR estimator of *central subspace* $S_{Y|X}$ as $span(\hat{B})$. The solution $(\hat{B}, \hat{C})$ can be found through an alternate weighted least square algorithms stated in Li and Yin (2008). In practice, τ is selected through a generalized cross validation criterion (GCV) following Golub, Heath and Wahba (1979). In order to provide variable selection results, Li and Yin (2008) further denoted that for a given tuning parameter λ, and ridge SIR estimator $(\hat{B}, \hat{C})$, the minimizer of

$$G_\lambda(\alpha) = \sum_{y=1}^{h} \hat{p}_y||(\bar{X}_y - \bar{X}) - \hat{\Sigma}_x diag(\alpha)\hat{B}\hat{C})y||, \tag{3.7}$$

over $\alpha \in \mathbb{R}^p$ with constraint that $\sum_{j=1}^{p} |\alpha_i| \leq \lambda$ leads to a sparse ridge SIR estimator as $span(diag(\hat{\alpha})\hat{B})$ on *central subspace* $S_{Y|X}$. Such $\hat{\alpha}$ can be found by reformulating the problem into a typical Lasso problem, and apply Lasso algorithms such as Tibshirani (1996) and Efron et. al (2004). Tuning parameter λ which controls the sparseness of the sparse ridge SIR estimator, is selected through criteria such as AIC, BIC and RIC (Li and Yin,

2008). A BIC type criterion similar to Zhu, Miao and Peng (2006) was also developed to determine structural dimension d.

3.3 Regularized sufficient dimension folding and variable selection

In this section, we focus our attention on matrix data $\mathbf{X} \in \mathbb{R}^{p_l \times p_r}$ where one seeks to find sparse solutions for basis of *left folding space* $S_{Y|\circ\mathbf{X}}$ as $\alpha \in \mathbb{R}^{p_l \times d_l}$ and for basis matrix of *right folding space* $S_{Y|\mathbf{X}\circ}$ as $\beta \in \mathbb{R}^{p_r \times d_r}$. We begin this section with discussions on two types of sparseness for the associated *central folding subspace* (Li, Kim and Altman, 2010) and develop methods correspondingly.

3.3.1 Different types of sparseness

Sparseness on the coordinates of basis matrices

Sparseness on coordinates corresponds to cases where certain coordinates across all columns of basis matrices α for *left folding subspace* $S_{Y|\circ\mathbf{X}}$ (β for *right folding subspace* $S_{Y|\mathbf{X}\circ}$) are 0. This type of sparseness usually result in a *folding central subspace* that contains a much smaller dimensional sub-matrix compared to the original predictor matrix. Thus one can achieve row (column) selection from the high dimensional original predictor matrix and further build predictive models on the sub-matrix. Among the following proposed methods, both dimension folding with L_1 regularization and dimension folding with both L_2 and L_1 regularizations are capable to achieve such coordinate sparseness and produce smaller dimensional sub-matrix. Dimension folding with L_2 regularization further utilizes ridge regression type adjustment to reduce the instability of estimating inverse of covariance matrix, when dimension $p_l \times p_r$ exceeds sample size n, or when predictors are highly correlated.

Sparseness on the vectorized basis matrices

Instead of producing a sub-matrix with smaller number of rows (columns), sparseness on the vectorized basis matrices, on the other hand, states that if we vectorize the basis matrix α (or β) as $vec(\alpha)$ (or $vec(\beta)$) for *left folding subspace* $S_{Y|\circ\mathbf{X}}$ (or *right folding subspace*), then $vec(\alpha)$ (or $vec(\beta)$) is a sparse vector where most of its elements are 0. This type of sparseness usually results in occasions where only a few elements among the the predictor matrix are associated with response variable Y, but the smallest matrix containing all these elements is not necessarily much smaller (i.e, in terms of number of rows and columns) compared to the original size of the predictor matrix. For this type of sparseness, as shown in the next section, Dimension folding with L_1 regularization was developed to promote such sparseness. The other method, dimension folding with both L_2 and L_1 regularizations may fail to capture such sparseness due to its intentions to only produce coordinate sparseness.

As a motivating example to understand better on two types of sparseness, consider a predictor matrix $\mathbf{X} \in \mathbb{R}^{4\times 4}$ with the following two different *central folding spaces* and their associated basis matrices:

Example 3.1

Assume that *central folding subspace* $S_{Y|\circ\mathbf{X}\circ}$ is spanned by

$$\beta \otimes \alpha = [e_1, e_2, e_5, e_6] \in \mathbb{R}^{16\times 4},$$

where $e_i \in \mathbb{R}^{16}$ is a vector with ith element 1, and other elements 0. In the predictor matrix $\mathbf{X}$, only elements $\mathbf{X}_{1,1}, \mathbf{X}_{1,2}, \mathbf{X}_{2,1}, \mathbf{X}_{2,2}$ are associated with response variable Y, and the smallest sub-matrix forming the *central folding subspace* is a 2×2 sub-matrix containing elements $\mathbf{X}_{1,1}, \mathbf{X}_{1,2}, \mathbf{X}_{2,1}, \mathbf{X}_{2,2}$. The *central folding subspace* $S_{Y|\circ\mathbf{X}\circ}$ can be further decomposed

into *left folding subspace* $S_{Y|\circ\mathbf{X}}$ and *right folding subspace* $S_{Y|\mathbf{X}\circ}$ with the same basis matrix

$$\alpha = \beta = [(1,0,0,0)^T, (0,1,0,0)^T] \in \mathbb{R}^{4\times 2}.$$

In this case, α and β are sparse both in terms of their vectorized version $vec(\alpha)$ and $vec(\beta)$, and in terms of coordinates, as the 3^{th} and 4^{th} elements for all columns of α and β are 0.

Example 3.2

Assume that *central folding subspace* $S_{Y|\circ\mathbf{X}\circ}$ is spanned by

$$\beta \otimes \alpha = diag(1, ..., 1) \in \mathbb{R}^{16\times 16}.$$

This is simply saying that all diagonal elements in the predictor matrix are associated with the response variable Y, and the smallest sub-matrix forming the *central folding subspace* is the same as predictor matrix $\mathbf{X}$. The *central folding subspace* $S_{Y|\circ\mathbf{X}\circ}$ can be decomposed into *left folding subspace* $S_{Y|\circ\mathbf{X}}$ and *right folding subspace* $S_{Y|\mathbf{X}\circ}$ with the same basis matrix

$$\alpha = \beta = diag(1, ..., 1) \in \mathbb{R}^{4\times 4}.$$

Therefore, both α and β are sparse in terms of their vectorized version $vec(\alpha)$ and $vec(\beta)$, but not in the sense of coordinates since there is no coordinates of the basis matrices producing exact 0 loadings among all columns.

As Figure 3.1 indicates, for both examples, there are 4 elements in the original predictor matrix $\mathbf{X}$ associated with *central folding subspace*, but the second example produces only sparseness on vectorized basis matrices while the first example exhibits sparseness on both vectorized basis matrices and coordinates. We develop method for both types of sparseness. In the following sections, sufficient dimension folding with L_1 regularization accounts for both

types of sparseness, and sufficient dimension folding with both L_1 and L_2 regularizations method only guarantees sparseness on the coordinates of basis matrices.

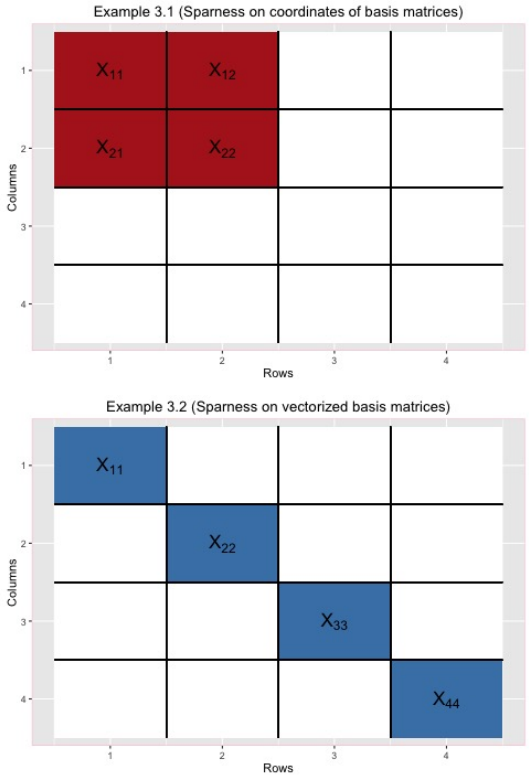

Figure 3.1: Description on different types of sparseness

Connections with sufficient variable selection subspace

The second type of sparseness, sparseness on the vectorized basis matrices, is closely related to the concept of *central variable selection subspace* in Yin and Hilafu (2015). They proposed a novel concept called sufficient variable selection, an idea similar to sufficient dimension reduction, but with focus on selecting relevant predictor variables. They define that:

Definition 1 (Yin and Hilafu, 2015) *For vector predictor $X \in \mathbb{R}^p$, if there is a $p \times m$ matrix β with $m \leq p$, and the the columns of matrix β only consist of unit vectors of e_i's where e_i is a vector with element 1 on its ith coordinate and 0 elsewhere, such that*

$Y \perp\!\!\!\perp X|\beta^T X$, *then column space of* β *is called the* variable selection space. *The intersection of all* variable selection spaces*, if still satisfying the independence condition called the* central variable selection space *as* $S^V_{Y|X}$.

Yin and Hilafu (2015) closed the cap between *central subspace* and *central variable selection space* by pointing out that $S_{Y|X} \subseteq S^V_{Y|X}$, and that if $S_{Y|X}$ exists, then $S^V_{Y|X}$ exists and is unique. To estimate *central variable selection space*, one strategy is to combine sufficient dimension reduction with a penalization approach such as Ni, Cook and Tsai (2005), Li and Nachtsheim (2006), Li (2007), Zhou and He (2008), Wang and Yin (2008) and Chen, Zou and Cook (2010). Yin and Hilafu (2015) took a different approach and proposed two paths to sequentially estimate $S^V_{Y|X}$ for large dimension p and sample sample size n.

For matrix predictor $\mathbf{X} \in \mathbb{R}^{p_l \times p_r}$, we demonstrate that an estimate of *central variable selection space* $S^V_{Y|vec(\mathbf{X})}$ can lead to an estimate of $\alpha \in \mathbb{R}^{p_l \times d_l}$ $(d_l \leq p_l)$ and $\beta \in \mathbb{R}^{p_r \times d_r}$ $(d_r \leq p_r)$ whose columns consist of unit vectors so that $Y \perp\!\!\!\perp \mathbf{X}|\alpha^T \mathbf{X}\beta$. Therefore it is not necessary to define *central folding variable selection space* for matrix predictor specifically.

Notice that $Y \perp\!\!\!\perp \mathbf{X}|\alpha^T X\beta$ is equivalent to $Y \perp\!\!\!\perp vec(\mathbf{X})|(\beta \otimes \alpha)^T vec(\mathbf{X})$. Suppose we have found the *central variable selection space* $S^V_{Y|vec(\mathbf{X})}$ as $span(B)$, where B consists of unit vectors $e_i \in \mathbb{R}^{p_l p_r}$, then B can always be decomposed into a Kronecker product of two matrices whose columns are also unite vectors. If we denote $p = mod(i, p_l)$, where $mod(a, b)$ defines the remainder of a/b, and that $q = \lceil i/pl \rceil$ where $\lceil a/b \rceil$ defines the ceiling integer of a/b, then we can decompose $e_i \in \mathbb{R}^{p_l \times p_r}$ as $e_i = \tilde{e}_p \otimes \tilde{e}_q$, where $\tilde{e}_p \in \mathbb{R}^{p_l}$ and $\tilde{e}_q \in \mathbb{R}^{p_r}$ are both unit vectors. Thus by decomposing each column of matrix B, one can obtain $span(B) = span(\beta \otimes \alpha)$ where α and β are matrices with unit vectors as columns. On the other hand, if we have found α and β both consisting of unit vectors such that $Y \perp\!\!\!\perp \mathbf{X}|\alpha^T X\beta$, then $span(\beta \otimes \alpha)$ is also a matrix whose columns are unit vectors.

For matrix predictor $\mathbf{X} \in \mathbb{R}^{p_l \times p_r}$, to connect the concept of *central variable selection space* $S^V_{Y|vec(\mathbf{X})}$ with sparseness on the vectorized basis matrices, note that for the estimated α in

central variable selection space, $vec(\alpha)$ is indeed sparse because it is composed by stacking unit vectors into a longer vector, and it is a special case of sparseness on the vectorized basis matrix, since all of its non-zero elements are restricted to be 1. Same statement holds for the estimation of β. On the other hand, α and β are not necessarily sparseness in their coordinates. The framework of sufficient variable selection (Yin and Hilafu, 2015) essentially sheds light on another approach to achieve sparse estimate on the vectorized basis matrices α and β: one can simply convert the matrix predictor into vector, apply two paths proposed in Yin and Hilafu (2015) to obtain estimate on the basis matrix B such that $span(B) = S^V_{Y|vec(\mathbf{X})}$, and then decompose B so that $B = \beta \otimes \alpha$, where columns of B, α and β are all unit vectors in their corresponding dimensions.

In order to explain how regularizations are incorporated into our proposed methods, we begin next section by reviewing on the least square formulation of sufficient dimension folding methods.

3.3.2 Sufficient dimension folding with L-1 regularization

Folding methods with least square formulation

Li, Altaman and Kim (2010) closed the gap between the traditional *central subspace* $S_{Y|vec(\mathbf{X})}$ and *central folding subspace* $S_{Y|\circ\mathbf{X}\circ}$ through the concept of *Kronecker envelope*. They stated that if we have a random matrix $U \in \mathbb{R}^{p_l p_r \times k}$ which serves as an estimator of *central subspace* $S_{Y|vec(\mathbf{X})}$ for the vectorized matrix predictor, i.e. $span(U) \subseteq S_{Y|vec(\mathbf{X})}$, then one can estimate its associated *central folding subspace* $S_{Y|\circ\mathbf{X}\circ}$ through the *Kronecker envelope* $\epsilon^{\otimes}(U)$ of random matrix U such that $\epsilon^{\otimes}(U) \subseteq S_{Y|\circ\mathbf{X}\circ}$. The definition and properties of *Kronecker envelope* can be found in Li, Kim and Altman (2010). The *Kronecker envelope* $\epsilon^{\otimes}(U) = \beta \otimes \alpha$

can be further estimated by minimizing:

$$E||AU - A(\beta \otimes \alpha)C||^2, \tag{3.8}$$

over $\alpha \in \mathbb{R}^{p_l \times d_l}$,$\beta \in \mathbb{R}^{p_r \times d_r}$ and $C \in \mathbb{R}^{d_l d_r \times k}$, where $A \in \mathbb{R}^{p_l p_r \times p_l p_r}$ is a non-singular matrix. One can specify random matrix U as different estimators of $S_{Y|vec(\mathbf{X})}$ and extend them to folding estimators through a least square formulation. For instance, by substituting $U = \Sigma_x^{-1} E(vec(\mathbf{X}|Y) - E(vec(\mathbf{X})))$ as a sliced inverse estimator (SIR; Li, 1993) of *central subspace* and A as $\Sigma^{\frac{1}{2}}$, one can estimate an Folded-SIR estimator of the *folding central subspace*.

For sample version estimation of (3.8), we first clarify notations as the following: suppose we have n independent and identically distributed sample points $(\mathbf{X}_1, Y_1), ..., (\mathbf{X}_n, Y_n)$. Let $vec(\bar{\mathbf{X}})$ denotes the vectorized grand average of predictor $\mathbf{X}$, and $\hat{\Sigma}_x$ is the estimated sample covariance matrix on $vec(\mathbf{X})$, that is, $\hat{\Sigma}_x = \frac{1}{n}\sum_{i=1}^{n} vec(\mathbf{X}_i - \bar{\mathbf{X}})vec^T(\mathbf{X}_i - \bar{X})$. Let $J_1, ..., J_h$ be the partition of subscripts $1, ..., n$ so that each J_y contains the indexes of points in slice y. Moreover, we partition response variable Y into h non-overlapped slices with n_y points in the yth slice, and denote the slice proportion as $\hat{p}_y = \frac{n_y}{n}$. We also denote the vectorized average of $\mathbf{X}$ in the yth slice as $vec(\bar{\mathbf{X}}_y)$. For folded-SIR, to estimate its corresponding *folding central subspace*, one minimize:

$$G(\alpha, \beta, C) = \sum_{y=1}^{h} \hat{p}_y ||\hat{\Sigma}_x^{-\frac{1}{2}} vec(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - \hat{\Sigma}_x^{\frac{1}{2}}(\beta \otimes \alpha)C_y||^2 \tag{3.9}$$

over $\alpha \in \mathbb{R}^{p_l \times d_l}$, $\beta \in \mathbb{R}^{p_r \times d_r}$ and $C = (C_1, ..., C_h) \in \mathbb{R}^{d_l d_r \times h}$ under constraint that $\beta^T\beta = I_{p_r}$ and $\alpha^T\alpha = I_{p_l}$, and that $||\cdot||$ stands for Frobenius norm with respect to its usual inner product. Equivalently, we can write the above objective function (3.9) in the original scale

of predictor $\mathbf{X}$ as,

$$\begin{aligned} G(\alpha, \beta, C) = \sum_{y=1}^{h} \hat{p}_y \{vec(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - \hat{\Sigma}_x(\beta \otimes \alpha) C_y\}^T \\ \times \hat{\Sigma}_x^{-1} \{vec(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - \hat{\Sigma}_x(\beta \otimes \alpha) C_y\}. \end{aligned} \tag{3.10}$$

Li, Kim and Altman (2010) develop an alternate least square method to estimate $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\mathbf{C}}$ in turn. Their simulation shows that such algorithm usually converge in a small number of steps.

Folded-SIR with L_1 regularization

To facilitate sparsity on the vectorized basis matrices $vec(\alpha)$ and $vec(\beta)$, we adopt folded-SIR and add L_1 regularization, and we delay the extension of other Folding methods such as folded-SAVE and folded-DR with L_1 regularization in the next part. First of all, similar to that of Li and Yin (2008), we propose the idea of sparse folded-SIR estimator by the following definition.

Definition 2 *A sparse folded-SIR estimator of the* central foldings subspace $S_{Y|\circ\mathbf{X}\circ}$ *is defined as span*$(\hat{\beta} \otimes \hat{\alpha})$*, where* $\hat{\alpha}$ *and* $\hat{\beta}$ *are estimated by minimizing*

$$\begin{aligned} G_{\lambda_\alpha, \lambda_\beta}(\alpha, \beta, C) = \sum_{y=1}^{h} \hat{p}_y \{vec(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - \hat{\Sigma}_x(\beta \otimes \alpha) C_y\}^T \\ \times \hat{\Sigma}_x^{-1} \{vec(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - \hat{\Sigma}_x(\beta \otimes \alpha) C_y\}, \end{aligned} \tag{3.11}$$

over α*,* β *and* C*, subject to* $|vec(\alpha)|_1 \leq \lambda_\alpha$ *and* $|vec(\beta)|_1 \leq \lambda_\beta$*.*

Here $|\cdot|_1$ stands for the L_1 norm of a vector with respect to it usual inner product, which is the sum of all the absolute values of its elements. To compute the optimization of (3.11) with the L_1 constraint, one can adopt a standard Lasso algorithm and estimate $\hat{\alpha}$, $\hat{\beta}$, and

C alternately, while fixing the other twos. We state the formal algorithm as the following 5 steps:

1. Generate initial values of $vec(\alpha) \in \mathbb{R}^{p_l d_l}$ and $vec(\beta) \in \mathbb{R}^{p_r d_r}$ from, say, standard normal distribution $MVN_{p_l d_l}(0, I_{p_l d_l})$ and $MVN_{p_r d_r}(0, I_{p_r d_r})$.

2. **For fixed α and β**:

   The optimization of (3.11) over C is equivalent to the usual least square estimation, since there is no penalty term associated with C. We can directly apply algorithm from Li, Kim and Altman (2010) as:

$$\begin{aligned} &\hat{C} = (\hat{C}_1, ..., \hat{C}_h), \quad where \\ &\hat{C}_y = ((\beta^T \otimes \alpha^T)\hat{\Sigma}_x(\beta \otimes \alpha))^{-1}(\beta^T \otimes \alpha^T)vec(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}),\ \forall y = 1, ..., h. \end{aligned} \tag{3.12}$$

3. **For fixed β and C**:

   The optimization of (3.11) over α is equivalent to minimize:

$$\begin{aligned} G_{\lambda_\alpha, \lambda_\beta}(\alpha, \beta, C) = \sum_{y=1}^{h} \hat{p}_y \{vec(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - ({C_y}^T \otimes \hat{\Sigma}_x)\Pi[vec(\beta) \otimes I_{p_l d_l}]vec(\alpha)\}^T \\ \times \hat{\Sigma}_x^{-1}\{vec(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - ({C_y}^T \otimes \hat{\Sigma}_x)\Pi[vec(\beta) \otimes I_{p_l d_l}]vec(\alpha)\}, \end{aligned} \tag{3.13}$$

   with constraint that $|vec(\alpha)| \leq \lambda_\alpha$. Here $\Pi \in \mathbb{R}^{p_l p_r d_l d_r \times p_l p_r d_l d_r}$ is a transformation matrix whose elements are 0s and 1s such that $vec(\beta \otimes \alpha) = \Pi[vec(\beta) \otimes vec(\alpha)]$ and its details can be found in Li, Kim and Altman (2010). If we define new variables as:

$$\begin{aligned} \tilde{Y} &= vec(vec(\bar{\mathbf{X}}_1) - vec(\bar{\mathbf{X}}), ..., vec(\bar{\mathbf{X}}_h) - vec(\bar{\mathbf{X}})) \in \mathbb{R}^{p_l p_r h}, \\ \tilde{X} &= (C^T \otimes \hat{\Sigma}_x)\Pi(vec(\beta) \otimes I_{p_l d_l}) \in \mathbb{R}^{p_l p_r h \times p_l d_l}, \\ \tilde{W}^{\frac{1}{2}} &= D_h^{1/2} \otimes \hat{\Sigma}_x^{-\frac{1}{2}} \in \mathbb{R}^{p_l p_r h \times p_l p_r h}, \end{aligned} \tag{3.14}$$

where $D_h = diag(\hat{p}_1, ..., \hat{p}_h)$. Then, one can estimate $\hat{\alpha}$ (or equivalently $\hat{v}ec(\alpha)$) by applying standard Lasso algorithm such as Tibshirani (1996) and Efron et al. (2004), with response variable $\tilde{W}^{\frac{1}{2}}\tilde{Y}$ and predictor variable $\tilde{W}^{\frac{1}{2}}\tilde{X}$, and constraint that $|vec(\alpha)|_1 \leq \lambda_\alpha$.

4. **For fixed α and C**

   Similarly, the optimization of (3.11) over β is equivalent to minimize:

$$\begin{aligned} G_{\lambda_\alpha,\lambda_\beta}(\alpha,\beta,C) = \sum_{y=1}^{h} \hat{p}_y \{vec(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - (C_y{}^T \otimes \hat{\Sigma}_x)\Pi[I_{p_r d_r} \otimes vec(\alpha)]vec(\beta)\}^T \\ \times \hat{\Sigma}_x^{-1}\{vec(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - (C_y{}^T \otimes \hat{\Sigma}_x)\Pi[I_{p_r d_r} \otimes vec(\alpha)]vec(\beta)\}, \end{aligned} \tag{3.15}$$

   with constraint that $|vec(\beta)| \leq \lambda_\beta$. If we keep the same definition of $\tilde{Y}$ and $\tilde{W}^{\frac{1}{2}}$, and replace $\tilde{X}$ with,

$$\tilde{X} = (C^T \otimes \hat{\Sigma}_x)\Pi(I_{p_r d_r} \otimes vec(\alpha)) \in \mathbb{R}^{p_l p_r h \times p_r d_r}. \tag{3.16}$$

   Then again, one can estimate $\hat{\beta}$ (or equivalently $\hat{v}ec(\beta)$) by applying Lasso algorithm with response variable $\tilde{W}^{\frac{1}{2}}\tilde{Y}$ and predictor variable $\tilde{W}^{\frac{1}{2}}\tilde{X}$, and constraint that $|vec(\beta)|_1 \leq \lambda_\beta$.

5. Check convergence of the algorithm.

   Calculate the objective function value at the kth iteration step as $G_{\lambda_\alpha,\lambda_\beta}(\hat{\alpha}^{(k)}, \hat{\beta}^{(k)}, \hat{C}^{(k)})$. Then if the absolute difference of the objective function between two consecutive steps is less than some pre-specified threshold ϵ (for example 10^{-4}), that is,

$$|G_{\lambda_\alpha,\lambda_\beta}(\hat{\alpha}^{(k)}, \hat{\beta}^{(k)}, \hat{C}^{(k)}) - G_{\lambda_\alpha,\lambda_\beta}(\hat{\alpha}^{(k-1)}, \hat{\beta}^{(k-1)}, \hat{C}^{(k)})| \leq \epsilon, \tag{3.17}$$

then stop the algorithm and use $(\hat{\alpha}^{(k)}, \hat{\beta}^{(k)}, \hat{C}^{(k)})$ as the estimation of (α, β, C). Otherwise iterate through step 2 to step 4 until it satisfies the convergence criterion.

Singularity of Σ_x

Note that the the algorithm involves calculating inverse of the covariance matrix for the vectorized predictor $\hat{\Sigma}_x$. $\hat{\Sigma}_x$ may not be invertible when dimension $p_l \times p_r$ exceeds the sample size n, or the inversion may become unstable when predictors are highly correlated. To address this problem, Li, Kim and Altman (2010) suggested that, in practice, one can either use Moore Penrose inverse $\hat{\Sigma}_x^{\dagger}$ or use a ridge-regression type inverse $(\hat{\Sigma}_x + \epsilon I_{p_l p_r})^{-1}$ where $\epsilon > 0$ is a small positive number. In our simulation studies and data application, we use ridge-regression type inverse.

Tuning Parameter

In our proposed methodology for sufficient dimension folding with L-1 regularization, there are two tuning parameters λ_α and λ_β awaited to be selected. Similar to that of Li and Yin (2008), we propose Akaike information criterion (AIC; Akaike, 1973), Bayesian information criterion (BIC; Schwarz, 1978) and residual information criterion (RIC; Shi and Tsai, 2002) as the following:

$$\begin{aligned}
AIC &= p_l p_r h \times log(\frac{G_{\lambda_\alpha,\lambda_\beta}(\hat{\alpha}, \hat{\beta}, \hat{C})}{p_l p_r h}) + 2 \times p_{\lambda_\alpha,\lambda_\beta}, \\
BIC &= p_l p_r h \times log(\frac{G_{\lambda_\alpha,\lambda_\beta}(\hat{\alpha}, \hat{\beta}, \hat{C})}{p_l p_r h}) + log(p_l p_r h) \times p_{\lambda_\alpha,\lambda_\beta}, \\
RIC &= (p_l p_r h - p_{\lambda_1,\lambda_2}) \times log(\frac{G_{\lambda_\alpha,\lambda_\beta}(\hat{\alpha}, \hat{\beta}, \hat{C})}{p_l p_r h - p_{\lambda_\alpha,\lambda_\beta}}) \\
&\quad + p_{\lambda_\alpha,\lambda_\beta}(log(p_l p_r h) - 1) + \frac{4}{p_l p_r h - p_{\lambda_\alpha,\lambda_\beta} - 2},
\end{aligned} \tag{3.18}$$

where $G_{\lambda_\alpha,\lambda_\beta}(\hat{\alpha}, \hat{\beta}, \hat{C})$ is the objective function evaluated with fixed tuning parameter λ_α and λ_β, and $p_{\lambda_\alpha,\lambda_\beta}$ denotes the total number of non-zero elements for matrices $\hat{\alpha}$ and $\hat{\beta}$. In

practice, we adopt a grid search strategy where a grid of pair $(\lambda_\alpha, \lambda_\beta)$ are provided, we apply the folded-SIR with L_1 regularization and calculate the associated AIC (or BIC, or RIC), then select the model with tuning parameters λ_α and λ_β that produced lowest AIC (or BIC, or RIC).

Extension to other folding methods with L_1 regularization

Extension of the previous idea can be easily adapted to other folding methods with L_1 regularizations, such as folded-SAVE with L_1 regularization, folded-DR with L_1 regularization. We briefly discuss the extensions as the following.

Folded-SAVE with L_1 regularization

To define a folded-SAVE with L_1 regularization estimator, one only needs to replace the term $vec(\bar{\mathbf{X}}_y - \mathbf{X})$ in the sample version of optimization function (3.11) by

$$\{\hat{\Sigma}_x - \frac{1}{n_y}\sum_{i\in J_y} vec(\mathbf{X}_i - \bar{\mathbf{X}}_y)vec(\mathbf{X}_i - \bar{\mathbf{X}}_y)^T\} \times \hat{\Sigma}_x^{-\frac{1}{2}}, \tag{3.19}$$

denoted as $\hat{M}_y \in \mathbb{R}^{p_l p_r \times p_l p_r}$ $\forall y = 1, ..., h$, and change dimension of $C_y \in \mathbb{R}^{d_l d_r \times 1}$ to $C_y \in \mathbb{R}^{d_l d_r \times p_l p_r}$ for $\forall y \in \{1, ..., h\}$.

Then, when computing solution for $\hat{C} = [\hat{C}_1, ..., \hat{C}_h] \in \mathbb{R}^{d_l d_r \times p_l p_r h}$, one can apply the same least square estimation in (3.12), but replacing the term $vec(\bar{\mathbf{X}}_y - \mathbf{X})$ with $\hat{M}_y$ $\forall y = 1, ..., h$, that is:

$$\begin{aligned} &\hat{C} = (\hat{C}_1, ..., \hat{C}_h), \quad where \\ &\hat{C}_y = ((\beta^T \otimes \alpha^T)\hat{\Sigma}_x(\beta \otimes \alpha))^{-1}(\beta^T \otimes \alpha^T)\hat{M}_y \quad y = 1, ..., h. \end{aligned} \tag{3.20}$$

When computing solutions for $\hat{\alpha}$ (or equivalently $\hat{v}ec(\alpha)$)), with fixed β and C, one can call standard Lasso algorithm with response variable $\tilde{W}^{\frac{1}{2}}\tilde{Y}$ and predictor matrix $\tilde{W}^{\frac{1}{2}}\tilde{X}$, where

$\tilde{X}$, $\tilde{Y}$ and $\tilde{W}$ are defined as:

$$\begin{aligned}\tilde{Y} &= vec\{\hat{M}_1, ..., \hat{M}_h\} \in \mathbb{R}^{{p_l}^2{p_r}^2 h}, \\ \tilde{X} &= (C^T \otimes \hat{\Sigma}_x)\Pi(vec(\beta) \otimes I_{p_l d_l}) \in \mathbb{R}^{{p_l}^2{p_r}^2 h \times p_l d_l}, \\ \tilde{W}^{\frac{1}{2}} &= {D_f}^{\frac{1}{2}} \otimes I_{p_l p_r} \otimes \hat{\Sigma}_x^{-1} \in \mathbb{R}^{{p_l}^2{p_r}^2 h \times {p_l}^2{p_r}^2 h},\end{aligned} \tag{3.21}$$

where Π is a permutation transformation matrix defined previously in (3.13). When computing solutions for $\hat{\beta}$ (or equivalently $\hat{v}ec(\beta)$) with fixed β and C, one can again call standard Lasso algorithm with the response variable $\tilde{W}^{\frac{1}{2}}\tilde{Y}$, but replacing term $\tilde{X}$ in predictor matrix $\tilde{W}^{\frac{1}{2}}\tilde{X}$ as:

$$\tilde{X} = (C^T \otimes \hat{\Sigma}_x)\Pi(I_{p_r d_r} \otimes vec(\alpha)) \in \mathbb{R}^{{p_l}^2{p_r}^2 h \times p_r d_r}. \tag{3.22}$$

Folded-DR with L_1 regularization

For folded-DR with L_1 regularization estimator, we define $(\mathring{\mathbf{X}}_1, \mathring{Y}_1), ..., (\mathring{\mathbf{X}}_n, \mathring{Y}_n)$ to be an independent copy of the original data, and similar notation follows for partition $\mathring{J}_1, ..., \mathring{J}_h$ on the new data. For simplification, if we denote $\bigtriangledown_i = vec(\mathbf{X}_i)$, $\mathring{\bigtriangledown}_j = vec(\mathring{\mathbf{X}}_j)$, $\triangle_{i,j} = \mathring{\bigtriangledown}_i - \bigtriangledown_j$ and $\hat{p}_{kl} = n_k n_l / n$, and for abbreviation, we define

$$\hat{M}_{kl} = [2\hat{\Sigma}_x - \frac{1}{n_k n_l}\sum_{i \in \mathring{J}_k}\sum_{j \in J_l} \triangle_{i,j}\triangle_{i,j}{}^T]\Sigma_x^{-\frac{1}{2}} \in \mathbb{R}^{p_l p_r \times p_l p_r} \ \forall i, j = 1, ..., h, \tag{3.23}$$

Again, one only needs to change the sample version of optimization function (3.11) as

$$G_{\lambda_\alpha, \lambda_\beta}(\beta, \alpha, C) = \sum_{k=1}^{h}\sum_{l=1}^{h} \hat{p}_{kl}\{\hat{M}_{kl} - \hat{\Sigma}_x(\beta \otimes \alpha)C_{kl}\}^T \hat{\Sigma}_x^{-1}\{\hat{M}_{kl} - \hat{\Sigma}_x(\beta \otimes \alpha)C_{kl}\}, \tag{3.24}$$

where $C_{kl} \in \mathbb{R}^{d_l d_r \times p_l p_r}$, $\forall k, l \in \{1, ..., h\}$, so that $C = [C_{11}, ..., C_{1h}, ..., C_{h1}, ..., C_{hh}] \in \mathbb{R}^{d_l d_r \times p_l p_r h^2}$. The optimization for (3.24) over C with fixed α and β can still be computed from a least

square estimation, that is,

$$\begin{aligned} \hat{C} &= [\hat{C}_{11}, ..., \hat{C}_{1h}, ..., \hat{C}_{h1}, ..., \hat{C}_{hh}], \quad where \\ \hat{C}_{kj} &= ((\beta^T \otimes \alpha^T)\hat{\Sigma}_x(\beta \otimes \alpha))^{-1}(\beta^T \otimes \alpha^T)\hat{M}_{kj} \quad k, j = 1, ..., h. \end{aligned} \tag{3.25}$$

When computing solutions for $\hat{\alpha}$ (or equivalently $\hat{v}ec(\alpha)$)), with fixed β and C, one can call standard Lasso algorithm with response variable $\tilde{W}^{\frac{1}{2}}\tilde{Y}$ and predictor matrix $\tilde{W}^{\frac{1}{2}}\tilde{X}$, where $\tilde{X}$, $\tilde{Y}$ and $\tilde{W}$ are defined as:

$$\begin{aligned} \tilde{Y} &= vec\{M_{11}, ..., M_{1h}, ..., M_{h1}, ..., M_{hh}\} \in \mathbb{R}^{{p_l}^2{p_r}^2h^2}, \\ \tilde{X} &= (C^T \otimes \hat{\Sigma}_x)\Pi(vec(\beta) \otimes I_{p_l d_l}) \in \mathbb{R}^{{p_l}^2{p_r}^2h^2 \times p_l d_l}, \\ \tilde{W}^{\frac{1}{2}} &= {D_f}^{\frac{1}{2}} \otimes {D_f}^{\frac{1}{2}} \otimes I_{p_l p_r} \otimes \hat{\Sigma}_x^{-1} \in \mathbb{R}^{{p_l}^2{p_r}^2h^2 \times {p_l}^2{p_r}^2h^2}. \end{aligned} \tag{3.26}$$

When computing solutions for $\hat{\beta}$ (or equivalently $\hat{v}ec(\beta)$) with fixed β and C, one can again call standard Lasso algorithm with the same response variable $\tilde{W}^{\frac{1}{2}}\tilde{Y}$, but replacing term $\tilde{X}$ in predictor matrix $\tilde{W}^{\frac{1}{2}}\tilde{X}$ as:

$$\tilde{X} = (C^T \otimes \hat{\Sigma}_x)\Pi(I_{p_r d_r} \otimes vec(\alpha)) \in \mathbb{R}^{{p_l}^2{p_r}^2h^2 \times p_r d_r}. \tag{3.27}$$

As for choosing tuning parameters λ_α and λ_β, the proposed folded-SAVE with L_1 regularization and folded-DR with L_1 regularization can take the same form of AIC, BIC and RIC as in (3.18) but with their own objective function evaluated at $(\hat{\alpha}, \hat{\beta}, \hat{C})$.

3.3.3 Sufficient dimension folding with L_2 regularization

Since the least square formulation of folding method involves computing the inverse of sample covariance matrix $\hat{\Sigma}_x$, the solution for sufficient dimension folding models may become extremely unstable when predictors are highly correlated, not to say $\hat{\Sigma}_x$ might be singular

when dimension $p_l \times p_r$ exceeds sample size n. In this section, we propose sufficient dimension folding with L_2 regularization which converts the original least square estimation into a ridge regression estimation to address the problem. We begin our description of the method by using folded-SIR with L_2 regularization as an example.

Folded-SIR with L_2 regularization

Recall that the folded-SIR estimator can be estimated through the minimization of objective function (3.10) over α, β and $C = (C_1, ..., C_h)$. In order to impose the L_2 regularization, one first needs to deal with the inverse covariance matrix in the objective function (3.10). First notice that, unlike Li and Yin (2008) who demonstrate that they could simply omit the inverse of covariance matrix in their similar objective function from their equation (3) to (4), without affecting the solution of the original objective function. In our case, we cannot simply ignore the inverse matrix. If we replace $\hat{\Sigma}_x$ in (3.10) with some other arbitrary non-negative definite B, and denote:

$$G(\beta, \alpha, C) = \sum_{y=1}^{h} \hat{p}_y \{vec(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - \hat{\Sigma}_x(\beta \otimes \alpha)C_y\}^T B \{vec(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - \hat{\Sigma}_x(\beta \otimes \alpha)C_y\}, \quad (3.28)$$

as the alternative objective function for (3.10), we can discuss the following different candidate B, including:

1. If $B = \hat{\Sigma}_x^{-1}$, then $\hat{\Sigma}_x^{-\frac{1}{2}} vec(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) \in S_{Y|vec(\mathbf{Z})}$, where $vec(\mathbf{Z}) = \hat{\Sigma}_x^{-\frac{1}{2}} vec(\mathbf{X})$, and that $Span(\hat{\Sigma}_x^{\frac{1}{2}}(\beta \otimes \alpha)) \in S_{Y|\circ\mathbf{Z}\circ}$ because $Span(\beta \otimes \alpha) \in S_{Y|\circ\mathbf{X}\circ}$. Thus the minimization of $G(\beta, \alpha, C)$ is thus with respect to space $S_{Y|\circ\mathbf{Z}\circ}$;

2. If $B = \hat{\Sigma}_x^{-2}$, then $\hat{\Sigma}_x^{-1} vec(\bar{\mathbf{X}}_y) \in S_{Y|vec(\mathbf{X})}$, and that $Span(\beta \otimes \alpha) \in S_{Y|\circ\mathbf{X}\circ}$. Thus the minimization of $G(\beta, \alpha, C)$, if replacing $vec(\bar{\mathbf{X}}_y - \bar{\mathbf{X}})$ by $vec(\bar{\mathbf{X}}_y)$, is with respect to space $S_{Y|\circ\mathbf{X}\circ}$;

3. If we are able to find the asymptotic distribution that $\frac{1}{n}\{vec(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - \hat{\Sigma}_x(\beta \otimes \alpha)C_y\} \sim MVN_{p_l p_r}(0, V)$ as sample size $n \to \infty$, where $V \in \mathbb{R}^{p_l p_r \times p_l p_r}$ is a non-negative definite symmetric matrix, then the ideal choice of B in (3.28) is simply V^{-1}.

Unfortunately, current literatures have not included thorough discussion on the asymptotic distribution of folding estimator obtained through the minimization of (3.9). Alternatively, Ding and Cook (2014) develop dimension folding with principal fitted components (DF-PMC), which claims that if predictor matrix $\mathbf{X} \in \mathbb{R}^{p_l \times p_r}$ follows a matrix normal distribution, and can be decomposed into:

$$\mathbf{X} = \mu + \mathcal{T}_1^T v_1^T f(Y) v_2 \mathcal{T}_2 + \epsilon, \tag{3.29}$$

where error term ϵ follows a matrix normal distribution, that is, $\epsilon \sim MN(0, \Omega, M)$, and other of details $\mathcal{T}_1$, $\mathcal{T}_2$, v_1, v_2 and μ can be found in Ding and Cook (2014), then $Span(\Omega^{-1}\mathcal{T}_1 \otimes M^{-1}\mathcal{T}_2) = S_{Y|\circ\mathbf{X}\circ}$. Note that the model specifies that predictor $\mathbf{X}$ follows a matrix normal distribution, thus one can estimate the likelihood of $\mathbf{X}$ and obtain maximum likelihood estimation on $\mathcal{T}_1$ and $\mathcal{T}_2$. From the formed likelihood, one can further derive the asymptotic distribution of the estimator $\hat{\mathcal{T}}_1$ and $\hat{\mathcal{T}}_2$. However, the authors develop an alternate eigen-value decomposition algorithm to obtain the the estimation of $\hat{\mathcal{T}}_1$ and $\hat{\mathcal{T}}_2$. Due to the fact the algorithm is an eigen-value decomposition problem, rather than a least square formulation, the generalization of DF-PMC with L_2 regularization is still challenging.

With the remaining of the section, we simply take $B = I_{p_l p_r}$ to avoid calculation of inverse of covariance matrix. We also compare the results of using $B = \hat{\Sigma}_x^{-1}$ and $B = \hat{\Sigma}_x^{-2}$ in our simulation to demonstrate such simplification is indeed feasible. Note that, although we cannot omit inverse of covariance matrix in the objective function, we indicate in the appendix that, even if we keep using $B = \hat{\Sigma}_x^{-1}$ in the objective function (3.28), the actual computation of finding L_2 regularized solution of $\hat{\alpha}$, $\hat{\beta}$ and $\hat{C}$ does not involve the calculation

of inverse of covariance matrix. This simplification is, however, only true for the development of folded-SIR with L_2 regularization, as for other methods such as folded-SAVE with L_2 regularization, such simplification is not correct.

Definition 3 *For non-negative constants τ_1 and τ_2, define*

$$\begin{aligned} G_{\tau_1,\tau_2}(\alpha,\beta,C) = \sum_{y=1}^{h} \hat{p}_y ||vec(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - \hat{\Sigma}_x(\beta \otimes \alpha)C_y||^2 \\ + \tau_1 vec(\alpha)^T vec(\alpha) + \tau_2 vec(\beta)^T vec(\beta). \end{aligned} \tag{3.30}$$

with constraint that $\alpha^T\alpha = I_{p_l}$, $\beta^T\beta = I_{p_r}$. Let $(\hat{\alpha},\hat{\beta},\hat{C}) = argmin_{\alpha,\beta,C} G_{\tau_1,\tau_2}(\alpha,\beta,C)$, then $span(\hat{\beta}\otimes\hat{\alpha})$ is called a ridge folded-SIR estimator of the central folding subspace *with respect to (τ_1,τ_2).*

To estimate $(\hat{\alpha},\hat{\beta},\hat{C})$, we propose an alternating least-square algorithm which is similar to the development of Li and Yin (2008) and Li, Altman, Kim (2010). We stated as follows:

1. Generate initial values for $vec(\alpha) \in \mathbb{R}^{p_l d_l}$ and $vec(\beta) \in \mathbb{R}^{p_r d_r}$ from, say, standard normal distribution $MVN_{p_l d_l}(0, I_{p_l d_l})$ and $MVN_{p_r d_r}(0, I_{p_r d_r})$.

2. **For fixed α and β**

   We can adopt the original least square algorithm to compute solution for $\hat{C}$ as Li, Altman, Kim (2010), which is:

$$\begin{aligned} &\hat{C} = (\hat{C}_1, ..., \hat{C}_h) \in \mathbb{R}^{d_l d_r \times h} \quad where \\ &\hat{C}_y = ((\beta^T \otimes \alpha^T)\hat{\Sigma}_x^2(\beta \otimes \alpha))^{-1}(\beta^T \otimes \alpha^T)\hat{\Sigma}_x vec(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) \in \mathbb{R}^{d_l d_r} \ y = 1, ..., h. \end{aligned} \tag{3.31}$$

3. **For fixed β and C:**

   If we define

$$\begin{aligned} \tilde{Y} &= vec(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}, ..., \bar{\mathbf{X}}_h - \bar{\mathbf{X}}) \in \mathbb{R}^{p_l p_r h}, \\ \tilde{X} &= (I_h \otimes \Sigma_x) U \in \mathbb{R}^{p_l p_r h \times p_l d_l}, \\ \tilde{W} &= D_f \otimes I_{p_l p_r} \in \mathbb{R}^{p_l p_r h \times p_l p_r h}, \end{aligned} \tag{3.32}$$

   where $U = [mat_{d_l}(C_1)\beta^T \otimes I_{p_l}, ..., mat_{d_l}(C_h)\beta^T \otimes I_{p_l}]^T \in \mathbb{R}^{p_l p_r h \times p_l d_l}$ and $D_f = diag(\hat{p}_1, ..., \hat{p}_h) \in \mathbb{R}^{h \times h}$. Here $mat_{d_l}(C_i)$ $i = 1, ..., h$, as defined in Li, Kim and Altman (2010), converts each vector $C_i \in R^{d_l d_r}$ into a $d_l \times d_r$ matrix. Then we can rewrite the objective function $G_{\tau_1,\tau_2}(\alpha, \beta, C)$ as

$$\begin{aligned} G_{\tau_1,\tau_2}(\alpha, \beta, C) = (\tilde{Y} - \tilde{X} vec(\alpha))^T \tilde{W} (\tilde{Y} - \tilde{X} vec(\alpha)) \\ + \tau_1 vec(\alpha)^T vec(\alpha) + \tau_2 vec(\beta)^T vec(\beta), \end{aligned} \tag{3.33}$$

   Therefore, the solution of $\hat{\alpha}$ (or equivalently $\hat{vec}(\alpha)$) can be computed from a weight least square problem, that is:

$$\begin{aligned} vec(\hat{\alpha}) &= (\tilde{X}^T \tilde{W} \tilde{X} + \tau_1 I_{p_l d_l})^{-1} \tilde{X}^T \tilde{W} \tilde{Y} \\ &= (U^T (D_f \otimes \hat{\Sigma}_x^2) U + \tau_1 I_{p_l d_l})^{-1} U^T (D_f \otimes \hat{\Sigma}_x) \tilde{Y}. \end{aligned} \tag{3.34}$$

   Standardize $\hat{\alpha}$ so that $\hat{\alpha}^T \hat{\alpha} = I_{p_l}$.

4. **For fixed α and C:**

   Similarly, using the same notation of $\tilde{Y}$ and $\tilde{W}$ as above, and define,

$$\tilde{\tilde{X}} = (I_h \otimes \Sigma_x) V \in \mathbb{R}^{p_l p_r h \times p_r d_r}, \tag{3.35}$$

   where $V = [K^T_{p_r,d_r}(I_{p_r} \otimes mat^L_{d_l}(C_1)\alpha^T), ..., K^T_{p_r,d_r}(I_{p_r} \otimes mat^L_{d_l}(C_h)\alpha^T)]^T \in \mathbb{R}^{p_l p_r h \times p_r d_r}$ and

$K_{p_r,d_r} \in \mathbb{R}^{p_r d_r \times p_r d_r}$ is a communication matrix composed of 0 and 1, and is intended to serve as "permutation" functionality to transform $vec(\hat{\beta}^T)$ to $vec(\hat{\beta})$, that is, $vec(\hat{\beta}^T) = K_{p_r,d_r} vec(\hat{\beta})$. Its details and explicit form can be found in Magnus and Neudecker (1979). Thus, we can also rewrite the objective function $G_{\tau_1,\tau_2}(\alpha, \beta, C)$ as:

$$\begin{aligned} G_{\tau_1,\tau_2}(\alpha, \beta, C) = (\tilde{Y} - \tilde{\tilde{X}} vec(\beta))^T \tilde{W} (\tilde{Y} - \tilde{\tilde{X}} vec(\beta)) \\ + \tau_1 vec(\alpha)^T vec(\alpha) + \tau_2 vec(\beta)^T vec(\beta). \end{aligned} \tag{3.36}$$

It is easy for us to obtain a ridge type solution for $\hat{\beta}$ (or equivalently $\hat{v}ec(\beta)$)as

$$\begin{aligned} vec(\hat{\beta}) &= (\tilde{\tilde{X}}^T \tilde{W} \tilde{\tilde{X}} + \tau_2 I_{p_r d_r})^{-1} \tilde{\tilde{X}}^T \tilde{W} \tilde{Y} \\ &= (V^T (D_f \otimes \hat{\Sigma}_x^2) V + \tau_2 I_{p_r d_r})^{-1} V^T (D_f \otimes \hat{\Sigma}_x) \tilde{Y}. \end{aligned} \tag{3.37}$$

Standardize $\hat{\beta}$ so that $\hat{\beta}^T \hat{\beta} = I_{p_r}$.

5. Check convergence of the algorithm.

   Calculate the objective function value at the kth iteration step as $G_{\tau_1,\tau_2}(\hat{\alpha}^{(k)}, \hat{\beta}^{(k)}, \hat{C}^{(k)})$. Then if the absolute difference of the objective function between two consecutive steps is less than some pre-specified threshold ϵ (for example 10^{-4}), that is,

$$|G_{\tau_1,\tau_2}(\hat{\alpha}^{(k)}, \hat{\beta}^{(k)}, \hat{C}^{(k)}) - G_{\tau_1,\tau_2}(\hat{\alpha}^{(k-1)}, \hat{\beta}^{(k-1)}, \hat{C}^{(k)})| \leq \epsilon, \tag{3.38}$$

   then stop the algorithm and use $(\hat{\alpha}^{(k)}, \hat{\beta}^{(k)}, \hat{C}^{(k)})$ as the estimation of (α, β, C). Otherwise iterate through step 2 to step 4 until it satisfies the convergence criterion.

To select tuning parameters τ_1 and τ_2, similar to those of Li and Yin (2008) and Golub, Heath, and Wahba (1979), if we further denote $\tilde{W}^{\frac{1}{2}} = D_f^{\frac{1}{2}} \otimes I_{p_l p_r}$, then we define a generalized

cross-validation criterion (GCV) as the following

$$GCV = \frac{||(I_{p_l p_r h} - S_{\tau_1})\tilde{W}^{1/2}\tilde{Y}||^2 + ||(I_{p_l p_r h} - S_{\tau_2})\tilde{W}^{1/2}\tilde{Y}||^2}{2p_l p_r h\{1 - \frac{trace(S_{\tau_1}) + trace(S_{\tau_2})}{2p_l p_r h}\}^2}, \tag{3.39}$$

where

$$\begin{aligned} S_{\tau_1} &= (D_f \otimes \hat{\Sigma}_x^{\frac{1}{2}})U(U^T(D_f \otimes \hat{\Sigma}_x^2)U + 2\tau_1 I_{p_l d_l})^{-1}U^T(D_f \otimes \hat{\Sigma}_x^{\frac{1}{2}}), \\ S_{\tau_2} &= (D_f \otimes \hat{\Sigma}_x^{\frac{1}{2}})V(V^T(D_f \otimes \hat{\Sigma}_x^2)V + 2\tau_2 I_{p_r d_r})^{-1}V^T(D_f \otimes \hat{\Sigma}_x^{\frac{1}{2}}). \end{aligned} \tag{3.40}$$

The details of the derivation of GCV can be found in the appendix. Through grid search strategy, we will select values of τ_1 and τ_2 which minimizes (3.39).

Extension to other folding methods with L_2 regularization

Similar to folded-SIR with L_2 regularization, we briefly describe how to extend the methodology to folded-SAVE with L_2 regularization and folded-DR with L_2 regularization.

Folded-SAVE with L_2 regularization

To define a folded-SAVE with L_2 regularization estimator, one only needs to replace the term $vec(\bar{\mathbf{X}}_y - \mathbf{X})$ in the sample version of optimization function (3.30) as

$$\{\hat{\Sigma}_x - \frac{1}{n_y}\sum_{i \in J_y} vec(\mathbf{X}_i - \bar{\mathbf{X}}_y)vec(\mathbf{X}_i - \bar{\mathbf{X}}_y)^T\} \times \hat{\Sigma}_x^{-\frac{1}{2}}, \tag{3.41}$$

denoted as $\hat{M}_y$ and change dimension of $C_y \in \mathbb{R}^{d_l d_r \times 1}$ to $C_y \in \mathbb{R}^{d_l d_r \times p_l p_r}$ for $\forall y \in \{1, ..., h\}$.

When computing solution for $\hat{C}$, we can directly replace the term $vec(\bar{\mathbf{X}}_y - \mathbf{X})$ by $\hat{M}_y$ in

equation (3.31), that is, the solution for $\hat{C}$ with fixed α and β is:

$$\begin{aligned} &\hat{C} = (\hat{C}_1, ..., \hat{C}_h) \in \mathbb{R}^{d_l d_r \times p_l p_r h} \quad where \\ &\hat{C}_y = ((\beta^T \otimes \alpha^T)\hat{\Sigma}_x^2(\beta \otimes \alpha))^{-1}(\beta^T \otimes \alpha^T)\hat{\Sigma}_x \hat{M}_y \in \mathbb{R}^{d_l d_r \times p_l p_r} \quad y = 1, ..., h. \end{aligned} \tag{3.42}$$

In the alternating algorithm to compute solutions for $\hat{\alpha}$ and $\hat{\beta}$, however, one needs to use another different approach, as $C_y \in \mathbb{R}^{d_l d_r \times p_l p_r}$ is a matrix, and it can no longer be converted as $mat_{d_l}(C_y)$. In fact, for fixed β and C, solution for $\hat{\alpha}$ (or equivalently $vec(\alpha)$) is:

$$\hat{v}ec(\alpha) = (\tilde{X}^T \tilde{W} \tilde{X} + \tau_1 I_{p_l d_l})^{-1} \tilde{X}^T \tilde{W} \tilde{Y} \in \mathbb{R}^{p_l d_l}, \tag{3.43}$$

where $\tilde{Y}$, $\tilde{X}$ and $\tilde{W}$ are defined as:

$$\begin{aligned} \tilde{Y} &= vec\{\hat{M}_1, ..., \hat{M}_h\} \in \mathbb{R}^{{p_l}^2 {p_r}^2 h}, \\ \tilde{X} &= (C^T \otimes \hat{\Sigma}_x)\Pi(vec(\beta) \otimes I_{p_l d_l}) \in \mathbb{R}^{{p_l}^2 {p_r}^2 h \times p_l d_l}, \\ \tilde{W} &= D_f \otimes I_{p_l p_r} \otimes I_{p_l p_r} \in \mathbb{R}^{{p_l}^2 {p_r}^2 h \times {p_l}^2 {p_r}^2 h}. \end{aligned} \tag{3.44}$$

Here Π is a permutation transformation matrix defined previously in (3.13). Similarly, the solution for β while fixing α and C is:

$$\hat{v}ec(\beta) = (\tilde{X}^T \tilde{W} \tilde{X} + \tau_2 I_{p_r d_r})^{-1} \tilde{X}^T \tilde{W} \tilde{Y} \in \mathbb{R}^{p_r d_r}, \tag{3.45}$$

and that $\tilde{Y}$, $\tilde{W}$ keep the same definition, but with $\tilde{X}$ changes to:

$$\tilde{X} = (C^T \otimes \hat{\Sigma}_x)\Pi(I_{p_r d_r} \otimes vec(\alpha)) \in \mathbb{R}^{{p_l}^2 {p_r}^2 h \times p_r d_r}. \tag{3.46}$$

Folded-DR with L_2 regularization

To define a ridge folded-DR estimator, let $(\mathring{\mathbf{X}}_1, \mathring{Y}_1), ..., (\mathring{\mathbf{X}}_n, \mathring{Y}_n)$ be an independent copy of the original data, and similar notation follows for partition $\mathring{J}_1, ..., \mathring{J}_h$ on the new data. Again, if we denote $\bigtriangledown_i = vec(\mathbf{X}_i)$, $\mathring{\bigtriangledown}_j = vec(\mathring{\mathbf{X}}_j)$, $\triangle_{i,j} = \mathring{\bigtriangledown}_i - \bigtriangledown_j$ and $\hat{p}_{kl} = n_k n_l / n$, and use the same abbreviation notation $\hat{M}_{kl}$ as in (3.23), then one can change the sample version of optimization function (3.30) as:

$$\begin{aligned} G_{\tau_1,\tau_2}(\beta, \alpha, C) = \sum_{k=1}^{h} \sum_{l=1}^{h} \hat{p}_{kl} ||\hat{M}_{kl} - \hat{\Sigma}_x(\beta \otimes \alpha) C_{kl}||^2 \\ + \tau_1 vec(\alpha)^T vec(\alpha) + \tau_2 vec(\beta)^T vec(\beta), \end{aligned} \tag{3.47}$$

where $C_{kl} \in \mathbb{R}^{d_l d_r \times p_l p_r}$, $\forall k, l \in \{1, ..., h\}$, so that $C = [C_{11}, ..., C_{1h}, ..., C_{h1}, ..., C_{hh}] \in \mathbb{R}^{d_l d_r \times p_l p_r h^2}$. The optimization for (3.47) over C with fixed α and β can be computed as

$$\begin{aligned} \hat{C} &= [\hat{C}_{11}, ..., \hat{C}_{1h}, ..., \hat{C}_{h1}, ..., \hat{C}_{hh}] \quad where \\ \hat{C}_{kj} &= ((\beta^T \otimes \alpha^T)\hat{\Sigma}_x^2(\beta \otimes \alpha))^{-1}(\beta^T \otimes \alpha^T)\hat{\Sigma}_x \hat{M}_{kj} \quad k, j = 1, ..., h. \end{aligned} \tag{3.48}$$

The alternate solutions for $\hat{\alpha}$ with fixed β and C is:

$$vec(\hat{\alpha}) = (\tilde{X}^T \tilde{W} \tilde{X} + \tau_1 I_{p_l d_l})^{-1} \tilde{X}^T \tilde{W} \tilde{Y} \in \mathbb{R}^{p_l d_l}, \tag{3.49}$$

where $\tilde{Y}$, $\tilde{X}$ and $\tilde{W}$ are defined as:

$$\begin{aligned} \tilde{Y} &= vec\{M_{11}, ..., M_{1h}, ..., M_{h1}, ..., M_{hh}\} \in \mathbb{R}^{{p_l}^2 {p_r}^2 h^2}, \\ \tilde{X} &= (C^T \otimes \hat{\Sigma}_x)\Pi(vec(\beta) \otimes I_{p_l d_l}) \in \mathbb{R}^{{p_l}^2 {p_r}^2 h^2 \times p_l d_l}, \\ \tilde{W} &= D_f \otimes D_f \otimes I_{p_l p_r} \otimes I_{p_l p_r} \in \mathbb{R}^{{p_l}^2 {p_r}^2 h^2 \times {p_l}^2 {p_r}^2 h^2}. \end{aligned} \tag{3.50}$$

Similarly, the solution for β while fixing α and C is:

$$\hat{v}ec(\beta) = (\tilde{X}^T\tilde{W}\tilde{X} + \tau_2 I_{p_r d_r})^{-1}\tilde{X}^T\tilde{W}\tilde{Y} \in \mathbb{R}^{p_r d_r}, \tag{3.51}$$

and that $\tilde{Y}$, $\tilde{W}$ keep the same definition, but with $\tilde{W}$ changed to:

$$\tilde{X} = (C^T \otimes \hat{\Sigma}_x)\Pi(I_{p_r d_r} \otimes vec(\alpha)) \in \mathbb{R}^{{p_l}^2 {p_r}^2 h^2 \times p_r d_r}. \tag{3.52}$$

As for choosing tuning parameters τ_1 and τ_2 for folded-SAVE with L_2 regularization and folded-DR with L_2 regularization, similar GCV statistic as in (3.39) can be derived for both methods, thus are omitted here.

Note that since both folded-SAVE and folded DR with L_2 regularization involve the inverse of covariance matrix $\hat{\Sigma}_x$ in the calculation of $\hat{M}_y$ in (3.41) and M_{kj} in (3.23), one can either use a Moore Penrose generalized inverse or a ridge-regression type inverse to avoid singularity when the dimensions $p_l \times p_r$ exceeds the sample size n. But unlike folded-SIR with L_2 regularization which does not need to calculate the inverse of sample covariance matrix, the estimator of folded-SAVE and folded-DR with L_2 regularization may become unstable since such approximation of inverse covariance matrix $\hat{\Sigma}_x$ is included in the numerical algorithm.

3.3.4 Sufficient dimension folding with both L_1 and L_2 regularizations

Folded-SIR with both L_1 and L_2 regularizations

In addition to properly address the singularity problem for sample covariance matrix and shrink the magnitude of the basis matrices when estimating the *central folding subspace*, the estimated ridge folded-SIR estimator still consists of linear combinations of all the rows and columns of the original matrix predictors. One would desire to pertain fewer number of rows and columns in the reduced space in order to have better interpretations. Thus, following the idea of least absolute shrinkage selection operator (Lasso) (Tibshirani, 1996), we introduce L-1 regularization to the ridge folded-SIR estimator similar to that of Li and Yin (2008) to achieve variable selection purpose. If we denote $(\hat{\alpha}, \hat{\beta}, \hat{C})$ as the ridge folded-SIR estimator, that is, $(\hat{\alpha}, \hat{\beta}, \hat{C}) = argmin_{\alpha,\beta,C} G_{\tau_1,\tau_2}(\alpha, \beta, C)$. We propose a sparse ridge folded-SIR estimator as the the following:

Definition 4 *For a non-negative constant λ_1 and λ_2, let $\hat{\phi} = (\hat{\phi}_1, ..., \hat{\phi}_{p_l}) \in \mathbb{R}^{p_l}$ and $\hat{\gamma} = (\hat{\gamma}_1, ..., \hat{\gamma}_{p_r}) \in \mathbb{R}^{p_r}$ be obtained by minimizing*

$$G_{\lambda_1,\lambda_2}(\phi, \gamma) = \sum_{y=1}^{h} \hat{p}_y ||vec(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) - \hat{\Sigma}_x (diag(\gamma)\hat{\beta} \otimes diag(\phi)\hat{\alpha})\ \hat{C}_y||^2, \tag{3.53}$$

over all possible ϕ and γ, with constraint that $\sum_{i=1}^{p_l} \phi_i \leq \lambda_1$ and $\sum_{i=1}^{p_r} \gamma_i \leq \lambda_2$. Then $Span(diag(\hat{\gamma})\hat{\beta}) \otimes Span(diag(\hat{\phi})\hat{\alpha})$ is defined as the sparse ridge folded-SIR estimator of the central folding subspace $S_{Y|\circ\mathbf{X}\circ}$.

We can estimate the corresponding $\hat{\phi} = (\hat{\phi}_1, ..., \hat{\phi}_{p_l})$ and $\hat{\gamma} = (\hat{\gamma}_1, ..., \hat{\gamma}_{p_r})$ alternately using standard LASSO algorithm. First, note that:

$$\begin{aligned}(diag(\gamma)\hat{\beta}) \otimes (diag(\phi)\hat{\alpha})C_y &= vec(diag(\phi)\hat{\alpha}\ mat_{d_l}(C_y)\hat{\beta}^T diag(\gamma)) \\ &= K_{p_r,p_l} vec(diag(\gamma)\hat{\beta} mat_{d_l}^T(C_y)\hat{\alpha}^T diag(\phi)),\end{aligned} \tag{3.54}$$

where K_{p_r,p_l} is a communication matrix whose definitions and properties can be found in Magnus and Neudecker (1979). Our alternate Lasso algorithm is stated as the following:

1. Generate $(\phi_1, ..., \phi_{p_l})$ and $(\gamma_1, ..., \gamma_{p_r})$ say, from standard normal distributions $MVN_{p_l}(0, I_{p_l})$ and $MVN_{p_r}(0, I_{p_r})$.

2. **For fixed** γ:

   Using the first equation of (3.54), if we denote $\hat{\alpha}\ mat_{d_l}(C_y)\hat{\beta}^T diag(\gamma) \in \mathbb{R}^{p_l \times p_r} = U^{(y)} = [U_1^{(y)}, ..., U_{p_r}^{(y)}]$, so that each $U_i^{(y)} \in \mathbb{R}^{p_l}\ \forall i = 1, ..., p_r$ is a column vector, and notice that

$$\begin{aligned}vec(diag(\phi)[U_1^{(y)}, ..., U(y)_{p_r}]) &= vec(diag(U_1)\phi, ..., diag(U_{p_r})\phi) \\ &= [diag(U_1^{(y)}), ..., diag(U_{p_r}^{(y)})]^T \phi.\end{aligned} \tag{3.55}$$

   For abbreviation, we denote $[diag(U_1^{(y)}), ..., diag(U_{p_r}^{(y)})]^T$ as $P_y \in \mathbb{R}^{p_l p_r \times p_l}\ , y = 1, ..., h$. Then if we write

$$\begin{aligned}\tilde{Y} &= vec(vec(\bar{\mathbf{X}}_1), ..., vec(\bar{\mathbf{X}}_h)) \in \mathbb{R}^{p_l p_r h}, \\ \tilde{X} &= (I_h \otimes \hat{\Sigma}_x)[P_1^T, ..., P_h^T]^T \in \mathbb{R}^{p_l p_r h \times p_l}, \\ \tilde{W} &= D_f \otimes I_{p_l p_r} \in \mathbb{R}^{p_l p_r h \times p_l p_r h}, \\ \tilde{W}^{\frac{1}{2}} &= D_f^{\frac{1}{2}} \otimes I_{p_l p_r} \in \mathbb{R}^{p_l p_r h \times p_l p_r h}.\end{aligned} \tag{3.56}$$

   We can estimate $\hat{\phi}$ by applying standard Lasso algorithms such as ones in Tibshirani

(1996) and Efron (2004) with response variable $\tilde{W}^{\frac{1}{2}}\tilde{Y}$, predictor variable design matrix $\tilde{W}^{\frac{1}{2}}\tilde{X}$.

3. **For fixed** ϕ:

   Using the second equation of (3.54), if we denote $\hat{\beta}mat_{d_l}^T(C_y)\hat{\alpha}^T diag(\phi)) \in \mathbb{R}^{p_r \times p_l} = V^{(y)} = [V_1^{(y)}, ..., V_{p_l}^{(y)}]$, so that each $V_i^{(y)} \in \mathbb{R}^{p_r}\ \forall i = 1, ..., p_l$ is a column vector, and notice that

$$\begin{aligned} vec(diag(\gamma)[V_1^{(y)}, ..., V(y)_{p_l}]) &= vec(diag(V_1)\gamma, ..., diag(V_{p_l})\gamma) \\ &= [diag(V_1^{(y)}), ..., diag(V_{p_r}^{(y)})]^T\gamma. \end{aligned} \tag{3.57}$$

   We denote $[diag(V_1^{(y)}), ..., diag(V_{p_r}^{(y)})]^T$ as $Q_y \in \mathbb{R}^{p_l p_r \times p_r}\ , y = 1, ..., h$. Similarly, if we use the same notations for $\tilde{Y}$ and $\tilde{W}$, but instead define $\tilde{X}$ as:

$$\tilde{X} = (I_h \otimes \hat{\Sigma}_x K_{pr,pl})[Q_1^T, ..., Q_h^T]^T \in \mathbb{R}^{p_l p_r h \times p_r}, \tag{3.58}$$

   then one can estimate $\hat{\gamma}$ again by calling Lasso algorithm with response variable $\tilde{W}^{\frac{1}{2}}\tilde{Y}$ and predictor $\tilde{W}^{\frac{1}{2}}\tilde{X}$.

4. Check convergence.

   Calculate the objective function value at the kth iteration step as $G_{\lambda_1,\lambda_2}(\hat{\phi}^{(k)}, \hat{\gamma}^{(k)})$. Then if the absolute difference of the objective function between two consecutive steps is less than some pre-specified threshold ϵ (for example 10^{-4}), that is,

$$|G_{\lambda_1,\lambda_2}(\hat{\phi}^{(k)}, \hat{\gamma}^{(k)}) - G_{\lambda_1,\lambda_2}(\hat{\phi}^{(k-1)}, \hat{\gamma}^{(k-1)})| \leq \epsilon, \tag{3.59}$$

   then stop the algorithm and use $(\hat{\phi}^{(k)}, \hat{\gamma}^{(k)})$ as the estimation of (ϕ, γ). Otherwise iterate through step 2 to step 4 until it satisfies the convergence criterion.

In our proposed methodology for folded-SIR with both L_1 and L_2 regularizations, there are two tuning parameters λ_1 and λ_2 awaited to be estimated. Similar to Akaike information criterion (AIC; Akaike, 1973), Bayesian information criterion (BIC; Schwarz, 1978), and residual and the criterion function (RIC; Shi and Tsai, 2002), for fixed tuning parameter λ_1, we define AIC, BIC and RIC as the following:

$$\begin{aligned} AIC &= p_l p_r h \times log(\frac{G_{\lambda_1,\lambda_2}(\hat{\phi},\hat{\gamma})}{p_l p_r h}) + 2 \times p_{\lambda_1,\lambda_2}, \\ BIC &= p_l p_r h \times log(\frac{G_{\lambda_1,\lambda_2}(\hat{\phi},\hat{\gamma})}{p_l p_r h}) + log(p_l p_r h) \times p_{\lambda_1,\lambda_2}, \\ RIC &= (p_l p_r h - p_{\lambda_1,\lambda_2}) \times log(\frac{G_{\lambda_1,\lambda_2}(\hat{\phi},\hat{\gamma})}{p_l p_r h - p_{\lambda_1,\lambda_2}}) \\ &\quad + p_{\lambda_1,\lambda_2}(log(p_l p_r h) - 1) + \frac{4}{p_l p_r h - p_{\lambda_1,\lambda_2} - 2}, \end{aligned} \tag{3.60}$$

where $G_{\lambda_1,\lambda_2}(\hat{\phi},\hat{\gamma})$ is the objective function value with fixed tuning parameter λ_1 and λ_2, and p_{λ_1,λ_2} denotes the total number of non-zero elements for vector $\hat{\phi}$ and $\hat{\gamma}$, it penalizes the least square solution so that we will favor sparse solution. In practice, we often search through a grid values of λ_1 and λ_2 and choose the pair such that the proposed AIC (or BIC, or RIC) is minimized.

Extension to other folding methods with both L_1 and L_2 regularizations

Folded-SAVE with both L_1 and L_2 regularizations

To define a folded-SAVE with both L_1 and L_2 regularizations estimator, one only needs to replace the term $vec(\bar{\mathbf{X}}_y - \mathbf{X})$ in the sample version of optimization function (3.53) as

$$\{\hat{\Sigma}_x - \frac{1}{n_y}\sum_{i\in J_y} vec(\mathbf{X}_i - \bar{\mathbf{X}}_y)vec(\mathbf{X}_i - \bar{\mathbf{X}}_y)^T\} \times \hat{\Sigma}_x^{-\frac{1}{2}}, \tag{3.61}$$

denoted as $M_y \in \mathbb{R}^{p_l p_r \times p_l p_r}$ and change dimension of $C_y \in \mathbb{R}^{d_l d_r \times 1}$ to $C_y \in \mathbb{R}^{d_l d_r \times p_l p_r}$ for $\forall y \in \{1, ..., h\}$. $\hat{\alpha}$, $\hat{\beta}$ and $\hat{C}$ are instead, solutions obtained by ridge folded-SAVE estimator.

Notice that the generalization of solutions for computing for $\hat{\phi}$ and $\hat{\gamma}$, however, is not trivial, since $C_y \in \mathbb{R}^{d_l d_r \times p_l p_r}$ is a matrix, and it can no longer be converted as $mat_{d_l}(C_y)$. In fact, first notice that if we decompose basis matrices into their column vectors, that is, $\hat{\alpha} = [\hat{\alpha}_1, ..., \hat{\alpha}_{d_l}] \in \mathbb{R}^{p_l \times d_l}$ and $\hat{\beta} = [\hat{\beta}_1, ..., \hat{\beta}_{d_r}] \in \mathbb{R}^{p_r \times d_r}$, where $\hat{\alpha}_i \in \mathbb{R}^{p_l}$ $\forall i = 1, ..., d_l$ and $\hat{\beta}_j \in \mathbb{R}^{p_r}$ $\forall i = 1, ..., d_r$, then the the following equations hold:

$$\begin{aligned} diag(\phi)\hat{\alpha} &= [diag(\hat{\alpha}_1), ..., diag(\hat{\alpha}_{d_l})]^T \phi, \\ diag(\gamma)\hat{\beta} &= [diag(\hat{\beta}_1), ..., diag(\hat{\beta}_{d_r})]^T \gamma. \end{aligned} \tag{3.62}$$

Therefore, we have that

$$\begin{aligned} &vec\{\hat{\Sigma}_x(diag(\gamma)\hat{\beta} \otimes diag(\phi)\hat{\alpha})\hat{C}_y\} \\ &= (\hat{C_y}^T \otimes \hat{\Sigma}_x)\Pi[vec(diag(\gamma)\beta) \otimes I_{p_l d_l}][diag(\hat{\alpha}_1), ..., diag(\hat{\alpha}_{d_l})]^T \phi \\ &= (\hat{C_y}^T \otimes \hat{\Sigma}_x)\Pi[I_{p_r d_r} \otimes vec(diag(\phi)\alpha)][diag(\hat{\beta}_1), ..., diag(\hat{\beta}_{d_r})]^T \gamma. \end{aligned} \tag{3.63}$$

For fixed γ, the solution for $\hat{\phi}$ can be solved by calling Lasso algorithm with response $\tilde{W}^{\frac{1}{2}}\tilde{Y}$ and predictor matrix $\tilde{W}^{\frac{1}{2}}\tilde{X}$, where

$$\begin{aligned} \tilde{Y} &= vec(\hat{M}_1, ..., \hat{M}_h) \in \mathbb{R}^{p_l p_r h}, \\ \tilde{X} &= (\hat{C}^T \otimes \hat{\Sigma}_x)\Pi[vec(diag(\gamma)\beta) \otimes I_{p_l d_l}][diag(\hat{\alpha}_1), ..., diag(\hat{\alpha}_{d_l})]^T \in \mathbb{R}^{p_l p_r h \times p_l}, \\ \tilde{W}^{\frac{1}{2}} &= D_f{}^{\frac{1}{2}} \otimes I_{p_l p_r} \in \mathbb{R}^{p_l p_r h \times p_l p_r h}. \end{aligned} \tag{3.64}$$

For fixed ϕ, the solution for $\hat{\gamma}$ can also be solved by Lasso algorithm with the same response variable $\tilde{W}^{\frac{1}{2}}\tilde{Y}$, but replacing the term $\tilde{X}$ in the predictor matrix $\tilde{W}^{\frac{1}{2}}\tilde{X}$ with

$$\tilde{X} = (\hat{C}^T \otimes \hat{\Sigma}_x)\Pi[I_{p_r d_r} \otimes vec(diag(\phi)\alpha)][diag(\hat{\beta}_1), ..., diag(\hat{\beta}_{d_r})]^T \gamma \in \mathbb{R}^{p_l p_r h \times p_r}. \tag{3.65}$$

Folded-DR with both L_1 and L_2 regularizations

To define a sparse ridge folded-DR estimator, if we keep the same previous notation that $(\mathring{\mathbf{X}}_1, \mathring{Y}_1), ..., (\mathring{\mathbf{X}}_n, \mathring{Y}_n)$ be an independent copy of the original data, and similar notation follows for partition $\mathring{J}_1, ..., \mathring{J}_h$ on the new data, and that $\bigtriangledown_i = vec(\mathbf{X}_i)$, $\mathring{\bigtriangledown}_j = vec(\mathring{\mathbf{X}}_j)$, $\triangle_{i,j} = \mathring{\bigtriangledown}_i - \bigtriangledown_j$ and $\hat{p}_{kl} = n_k n_l / n$, and same $\hat{M}_{kl}$ $\forall k, j = 1, ..., h$ in (3.23). The definition of sparse ridge folded-DR estimator follows similar definition as sparse ridge folded-SIR estimator in (3.53), that is

$$G_{\lambda_1,\lambda_2}(\beta, \alpha, C) = \sum_{k=1}^{h} \sum_{l=1}^{h} \hat{p}_{kl} ||\hat{M}_{kl} - \hat{\Sigma}_x (diag(\gamma)\hat{\beta} \otimes diag(\phi)\hat{\alpha}) \hat{C}_{kl}||^2, \tag{3.66}$$

with constraint that $\sum_{i=1}^{p_l} |\phi_i| \leq \lambda_1$and $\sum_{i=1}^{p_r} |\gamma_i| \leq \lambda_2$, and $\hat{\alpha}$, $\hat{\beta}$ and $\hat{C}_{kl} \in \mathbb{R}^{d_l d_r \times p_l p_r}$ for $\forall k, l = 1, ..., h$, $\hat{C} = [\hat{C}_{11}, ..., \hat{C}_{1h}, ..., \hat{C}_{h1}, ..., \hat{C}_{hh}] \in \mathbb{R}^{d_l d_r \times {p_l}^2 {p_r}^2}$ are obtained by ridge folded-DR estimator. For fixed γ, the solution for $\hat{\phi}$ can be solved by calling Lasso algorithm with response $\tilde{W}^{\frac{1}{2}} \tilde{Y}$ and predictor matrix $\tilde{W}^{\frac{1}{2}} \tilde{X}$, where

$$\begin{aligned}
\tilde{Y} &= vec\{M_{11}, ..., M_{1h}, ..., M_{h1}, ..., M_{hh}\} \in \mathbb{R}^{{p_l}^2 {p_r}^2 h^2}, \\
\tilde{X} &= (\hat{C}^T \otimes \hat{\Sigma}_x) \Pi [vec(diag(\gamma)\beta) \otimes I_{p_l d_l}][diag(\hat{\alpha}_1), ..., diag(\hat{\alpha}_{d_l})]^T \in \mathbb{R}^{{p_l}^2 {p_r}^2 h^2 \times p_l d_l}, \\
\tilde{W}^{\frac{1}{2}} &= {D_f}^{\frac{1}{2}} \otimes {D_f}^{\frac{1}{2}} \otimes I_{p_l p_r} \otimes I_{p_l p_r} \in \mathbb{R}^{{p_l}^2 {p_r}^2 h^2 \times {p_l}^2 {p_r}^2 h^2}.
\end{aligned} \tag{3.67}$$

For fixed ϕ, the solution for $\hat{\gamma}$ can also be solved by Lasso algorithm with the same response variable $\tilde{W}^{\frac{1}{2}} \tilde{Y}$, but replacing the term $\tilde{X}$ in the predictor matrix $\tilde{W}^{\frac{1}{2}} \tilde{X}$ with

$$\tilde{X} = (\hat{C}^T \otimes \hat{\Sigma}_x) \Pi [I_{p_r d_r} \otimes vec(diag(\phi)\alpha)][diag(\hat{\beta}_1), ..., diag(\hat{\beta}_{d_r})]^T \in \mathbb{R}^{{p_l}^2 {p_r}^2 h^2 \times p_l d_l}. \tag{3.68}$$

In order to select appropriate L_1 regularization tuning parameters λ_1 and λ_2 for sparse ridge folded-SAVE estimator and sparse ridge folded-DR estimator, we can adopt similar AIC, BIC and RIC criteria defined in (3.60), by replacing the objective function value with folded-SAVE and folded-DR methods.

3.4 Estimation of structural dimensions

In the estimation procedure of regularized sufficient dimension folding estimators, we regard structural dimension d_l and d_r as known quantities. As a matter of fact, in practice, we still need to estimate d_l and d_r based on data. We propose two methods in this section to determine structural dimensions. In the first part, we follow an empirical procedure, with the help of AIC, BIC and RIC type statistics to determine structural dimensions. In the second part, we follow the idea of dimension folding principal fitted components proposed by Ding and Cook (2014), as well as Minimum Average Variance Estimation (MAVE) and Outer Product Gradient (OPG) in Xia et al (2002) to construct a OPG type criterion.

3.4.1 Empirical AIC, BIC and RIC

We use folded-SIR as an example. Recall that the solution of folded-SIR estimator can be found by minimizing

$$G(\beta, \alpha, C) = \sum_{y=1}^{h} \hat{p}_y ||\hat{\Sigma}_x^{-\frac{1}{2}}(vec(\bar{\mathbf{X}}_y) - vec(\bar{\mathbf{X}})) - \hat{\Sigma}_x^{\frac{1}{2}}(\beta \otimes \alpha)C_y||^2, \tag{3.69}$$

over $\alpha \in \mathbb{R}^{p_l \times d_l}$, $\beta \in \mathbb{R}^{p_r \times d_r}$ and $C = (C_1, ..., C_h) \in \mathbb{R}^{d_l d_r \times h}$, and that $||\cdot||$ stands for Frobenius norm with respect to the usual inner product. For a given pair of structural dimensions d_l and d_r, denote the objective function under solution $(\hat{\alpha}, \hat{\beta}, \hat{C})$ as $G_{d_l d_r}(\hat{\alpha}, \hat{\beta}, \hat{C})$. We define

the empirical AIC, BIC and RIC as

$$
\begin{aligned}
AIC &= p_l p_r h \times log(\frac{G_{d_l d_r}(\hat{\alpha}, \hat{\beta}, \hat{C})}{p_l p_r h}) + 2 \times p_{d_l, d_r}, \\
BIC &= p_l p_r h \times log(\frac{G_{d_l d_r}(\hat{\alpha}, \hat{\beta}, \hat{C})}{p_l p_r h}) + log(p_l p_r h) \times p_{d_l, d_r}, \\
RIC &= (p_l p_r h - p_{d_l, d_r}) \times log(\frac{G_{d_l d_r}(\hat{\alpha}, \hat{\beta}, \hat{C})}{p_l p_r h - p_{d_l, d_r}}) \\
&\quad + p_{d_l, d_r}(log(p_l p_r h) - 1) + \frac{4}{p_l p_r h - p_{d_l, d_r} - 2},
\end{aligned}
\tag{3.70}
$$

where $p_{d_l d_r}$ denotes the number of parameters to estimate for basis matrices α and β. In our case, $p_{d_l d_r} = p_l \times d_l + p_r \times d_r$. We choose dimensions d_l and d_r which minimizes AIC, or BIC or RIC. The proposed AIC, BIC and RIC in (3.70) to determine structural dimensions for folded-SIR estimator can be easily extend to similar criteria for folded-SAVE and folded-DR, with $G_{d_l d_r}(\hat{\alpha}, \hat{\beta}, \hat{C})$ replaced by their associated estimated objective function values.

3.4.2 Folding with OPG type criterion

Ding and Cook (2014) proposed dimension folding principal fitted Components (DF-PFC) which converts the estimation of *central folding subspace* into regression models. They claimed that, the predictor matrix $\mathbf{X} \in \mathbb{R}^{p_l \times p_r}$ can be written as:

$$
\mathbf{X} = \mu + \mathcal{T}_2 v_2 f(Y) {v_1}^T \mathcal{T}_1^T + \epsilon, \tag{3.71}
$$

where $\mathcal{T}_1 \in \mathbb{R}^{p_r \times d_r}$, $\mathcal{T}_2 \in \mathbb{R}^{p_l \times d_l}$ are semi-orthogonal matrices that reduces the column and row dimensions of predictor matrix $\mathbf{X}$. $\mu \in \mathbb{R}^{p_l \times p_r}$ is the overall mean of $\mathbf{X}$, and error term ϵ is independent of Y. $f(Y) \in R^{q_l \times q_r}$ is a function of response Y. $v_1 \in \mathbb{R}^{d_r \times q_r}$ $(d_r \leq q_r)$, $v_2 \in \mathbb{R}^{d_l \times q_l}$ $(d_l \leq q_l)$ are the latent variables of rank d_r and d_l.

For continuous response Y, one can usually approximate $f(Y)$ by a series of basis functions or piecewise basis functions. For instance when using polynomial approximates, we set $q_l = q_r = r$, and $f(Y)$ as a diagonal matrix with diagonal elements as $Y, Y^2, ..., Y^r$. For categorical response Y with h categories, one can chose $f(Y)$ to be diagonal matrix of dimension $q_l = q_r = h - 1$, and its kth diagonal element as $diag(f(Y))_k = I(Y \in J_k) - n_k/n, k = 1, ..., h-1$, where $I(\cdot)$ is an indicator function, J_k indicates the kth category, and n_k it the number of observations in J_k.

Ding and Cook (2014) further proved that when random error ϵ follows a general matrix normal distribution $N_{p_l \times p_r}(0_{p_l \times p_r}, \Omega, M)$, then the desired *central folding subspace* can be equivalent as $S_{Y|\circ\mathbf{X}\circ} = Span(\Omega^{-1}\mathcal{T}_1) \otimes Span(M^{-1}\mathcal{T}_2)$. If the random error ϵ is isotropic, that is if $\epsilon \sim N_{p_l \times p_r}(0_{p_l \times p_r}, I_{p_r}, I_{p_l})$, then the *central folding subspace* $S_{Y|\circ\mathbf{X}\circ}$ can further be simplified as $Span(\mathcal{T}_1) \otimes Span(\mathcal{T}_2)$. For simplicity, we assume isotropic errors from now on.

Ding and Cook (2014) estimated the desired *central folding subspace* by assuming a global linear model form as in (3.71), such that:

$$E(\mathbf{X}|Y) = \mu + \mathcal{T}_2 v_2 f(Y) {v_1}^T \mathcal{T}_1^T. \tag{3.72}$$

Instead of estimating $E(\mathbf{X}|Y)$ as a global linear model, one can also use a local model to approximate $E(\mathbf{X}|Y)$, and thus estimate *central folding space* accordingly. Similar to the local polynomial smoothing (Fan and Gijbels, 1996) and Minimum Average Variance Estimation (MAVE) proposed by Xia et al (2002), for each data point $(\mathbf{X}_i, Y_i)$, we can approximate $E(\mathbf{X}_i|Y_i)$ at point Y_0 as:

$$E(\mathbf{X}_i|Y_i) = c_0 + \mathcal{T}_{2i} v_{2i} [f(Y_i) - f(Y_0)] {v_1}_i^T \mathcal{T}_{i0}^T \tag{3.73}$$

where $\mathcal{T}_{1i} \in \mathbb{R}^{p_r}$, $\mathcal{T}_{2i} \in \mathbb{R}^{p_l}$, $v_1 \in \mathbb{R}^{q_l}$, $v_{2i} \in \mathbb{R}^{q_r}$ are the latent variables, $c_0 \in \mathbb{R}^{p_l \times p_r}$ is the difference of overall mean of for $\mathbf{X}_i$ at point Y_i and Y_0, and error term ϵ is still independent

of Y. $f(Y) \in R^{q_l \times q_r}$ is a function of response Y that follows the same definition as in Ding and Cook (2014). Thus, one can estimate basis matrices for *central folding subspace* by minimizing

$$E(\mathbf{X} - E(\mathbf{X}|Y)), \tag{3.74}$$

using local approximation of $E(\mathbf{X}|Y)$ in (3.73). As for sample estimation, we can minimize:

$$\sum_{i=1}^{n} ||X_i - c_j - \mathcal{T}_{2j} v_{2j} (f(Y_i) - f(Y_j)) {v_{1j}}^T {\mathcal{T}_{1j}}^T||^2 w_{i,j}, \tag{3.75}$$

over $c_j \in \mathbb{R}^{p_l \times p_r}$, $\mathcal{T}_{1j} \in \mathbb{R}^{p_r}$, $\mathcal{T}_{2j} \in \mathbb{R}^{p_l}$, $v_{1j} \in \mathbb{R}^{1 \times q_r}$ and $v_{2j} \in \mathbb{R}^{1 \times q_l}$, with constraint that $\mathcal{T}_{1j}^T \mathcal{T}_{1j} = 1$, $\mathcal{T}_{2j}^T \mathcal{T}_{2j} = 1$, and $v_{1j} v_{1j}^T = 1$ for all $j = 1, ..., n$. $||\cdot||$ is taken as Frobenius norm with respect to usual inner product. w_{ij} are kernel weights centered at Y_i, Y_j with $\sum_{ij} w_{ij} = 1$, they satisfy that

$$w_{ij} = \frac{K_h(diag(Y_i) - diag(Y_j))}{\sum_{h=1}^{n} K_h(diag(Y_i) - diag(Y_j))}. \tag{3.76}$$

Here, for any $v \in \mathbb{R}^p$, $K_h(v) = h^{-p} K(||v||/h)$, where $K(\cdot)$ is the chosen kernel function. In our study, we simply take Gaussian kernel. We apply a similar iterate least square algorithm proposed by Xue and Yin (2014) in their folded-OPG method to obtain estimates of $\hat{\mathcal{T}}_{1j} \in \mathbb{R}^{p_r}$, $\hat{\mathcal{T}}_{2j} \in \mathbb{R}^{p_l}$, $\hat{v_{1j}} \in \mathbb{R}^{1 \times r}$, $\hat{v_{2j}} \in \mathbb{R}^{1 \times r}$ and $\hat{c}_j \in \mathbb{R}^{p_l \times p_r}$ for $j = 1, ..., n$. We illustrate the algorithm as the following:

For each $j = 1, ..., n$, we can iteratively estimate $\mathcal{T}_{1j}$, $\mathcal{T}_{2j}$, v_{1j}, v_{2j} and c_j by:

1. Generate the initial values of $\hat{\mathcal{T}}_{2j} \in \mathbb{R}^{p_l}$, $\hat{v_{1j}} \in \mathbb{R}^{1 \times r}$ and $\hat{v_{2j}} \in \mathbb{R}^{1 \times r}$ for $j = 1, ..., n$ from a sample of say, standard normal distribution $N(0, 1)$ variables.

2. For fixed $\mathcal{T}_{1j}$, v_{1j} and v_{2j}, minimize (3.75) over c_j and $\mathcal{T}_{1j}$. If we denote

$$\begin{aligned}
\tilde{Y} &= vec(\mathbf{X}_1, ..., \mathbf{X}_n) \in \mathbb{R}^{p_l p_r n}, \\
\tilde{X}_j &= \begin{pmatrix} I_{p_l p_r}, [I_{p_r} \otimes \mathcal{T}_{2j} v_{2j} (f(Y_1) - f(Y_j)) {v_{1j}}^T)], \\ ..., \\ I_{p_l p_r}, [I_{p_r} \otimes \mathcal{T}_{2j} v_{2j} (f(Y_n) - f(Y_j)) {v_{1j}}^T)] \end{pmatrix} \in \mathbb{R}^{(p_l p_r n) \times (p_l + 1) p_r}, \\
\tilde{W}_j &= diag(w_{1j}, ..., w_{nj}) \otimes I_{p_l p_r} \in \mathbb{R}^{p_l p_r n \times p_l p_r n},
\end{aligned} \tag{3.77}$$

then solution of $\hat{c}_j$ and $\hat{\mathcal{T}}_{2j}$, or equivalently,

$$\begin{pmatrix} vec(\hat{c}_j) \\ \hat{\mathcal{T}}_{1j} \end{pmatrix} = (\tilde{X}_j^T \tilde{W}_j \tilde{X}_j)^{-1} \tilde{X}_j^T \tilde{W}_j \tilde{Y} \in \mathbb{R}^{(p_l + 1) p_r}. \tag{3.78}$$

Standardize $\hat{\mathcal{T}}_{1j}$ so that $\hat{\mathcal{T}}_{1j}^T \hat{\mathcal{T}}_{1j} = 1$.

3. For fixed $\mathcal{T}_{1j}$, $\mathcal{T}_{2j}$, v_{2j} and c_j, minimize (3.75) over v_{1j}. If we keep the same notation of $\tilde{W}_j$, but replace $\tilde{Y}$ and $\tilde{X}_j$ as:

$$\begin{aligned}
\tilde{Y} &= vec(\mathbf{X}_1 - c_j, ..., \mathbf{X}_n - c_j) \in \mathbb{R}^{p_l p_r n}, \\
\tilde{X}_j &= \begin{pmatrix} \mathcal{T}_{1j} \otimes \mathcal{T}_{2j} v_{2j} (f(Y_1) - f(Y_j)), \\ ..., \\ \mathcal{T}_{1j} \otimes \mathcal{T}_{2j} v_{2j} (f(Y_n) - f(Y_j)) \end{pmatrix} \in \mathbb{R}^{(p_l p_r n) \times r},
\end{aligned} \tag{3.79}$$

then solution of $\hat{v_{1j}}$, or equivalently,

$$vec(\hat{v_{1j}}) = ({\tilde{X}_j}^T \tilde{W}_j \tilde{X}_j)^{-1} {\tilde{X}_j}^T \tilde{W}_j \tilde{Y} \in \mathbb{R}^r. \tag{3.80}$$

Standardize $\hat{v_{1j}}$ so that $\hat{v_{1j}} \hat{v_{1j}}^T = 1$.

4. For fixed $\mathcal{T}_{1j}$, v_{1j}, v_{2j} and c_j, minimize (3.75) over $\mathcal{T}_{2j}$. If we keep the same notation of $\tilde{W}$, and change $\tilde{X}$, $\tilde{Y}$ as:

$$\begin{aligned} \tilde{Y} &= vec(\mathbf{X}_1, ..., \mathbf{X}_n) \in \mathbb{R}^{p_l p_r n}, \\ \tilde{X}_j &= \begin{pmatrix} \mathcal{T}_{1j} v_{1j} (f(Y_1) - f(Y_j))^T {v_{2j}}^T \otimes I_{p_l}, \\ ..., \\ \mathcal{T}_{1j} v_{1j} (f(Y_n) - f(Y_j))^T {v_{2j}}^T \otimes I_{p_l} \end{pmatrix} \in \mathbb{R}^{(p_l p_r n) \times p_l}, \end{aligned} \tag{3.81}$$

then solution of $\hat{\mathcal{T}}_{2j}$ is:

$$\hat{\mathcal{T}}_2 = (\tilde{X}_j^T \tilde{W}_j \tilde{X}_j)^{-1} \tilde{X}_j^T \tilde{W}_j \tilde{Y} \in \mathbb{R}^{(p_l+1)p_r}. \tag{3.82}$$

Standardize $\hat{\mathcal{T}}_{2j}$ so that ${\hat{\mathcal{T}}_{2j}}^T \hat{\mathcal{T}}_{2j} = 1$.

5. For fixed $\mathcal{T}_{1j}$, $\mathcal{T}_{2j}$, v_{1j} and c_j, minimize (3.75) over v_{2j}. If we keep the same notation of $\tilde{W}_j$ and $\tilde{Y}$ but replace $\tilde{X}_j$ as

$$\tilde{X}_j = \begin{pmatrix} \mathcal{T}_{1j} v_{1j} (f(Y_1) - f(Y_j))^T \otimes \mathcal{T}_{2j}, \\ ..., \\ \mathcal{T}_{1j} v_{1j} (f(Y_n) - f(Y_j))^T \otimes \mathcal{T}_{2j} \end{pmatrix} \in \mathbb{R}^{(p_l p_r n) \times r}, \tag{3.83}$$

then solution of $\hat{v_{2j}}$, or equivalently,

$$vec(\hat{v_{2j}}) = ({\tilde{X}_j}^T \tilde{W}_j \tilde{X}_j)^{-1} {\tilde{X}_j}^T \tilde{W}_j \tilde{Y} \in \mathbb{R}^r. \tag{3.84}$$

6. Check convergence. If we denote the sample version of objective function as $G_j(\hat{\mathcal{T}}_{1j}, \hat{\mathcal{T}}_{2j}, \hat{v_{1j}}, \hat{v_{2j}}) = \sum_{i=1}^n ||\mathbf{X}_i - \hat{c}_j - \mathcal{T}_{2j} v_{2j} (f(Y_i) - f(Y_j)) \hat{v_{1j}}^T {\hat{\mathcal{T}}_{1j}}^T ||^2 w_{ij}$, then we stop the algorithm if the absolute difference between kth step objective function value

and $k-1$th step objective function value is smaller than some pre-specified tolerance value ϵ, i.e.,

$$|G_j(\hat{\mathcal{T}}_{1j}^{(k)}, \hat{\mathcal{T}}_{2j}^{(k)}, \hat{v_1}_j^{(k)}, \hat{v_2}_j^{(k)}) - G_j(\hat{\mathcal{T}}_{1j}^{(k-1)}, \hat{\mathcal{T}}_{2j}^{(k-1)}, \hat{v_1}_j^{(k-1)}, \hat{v_2}_j^{(k-1)})| \leq \epsilon. \tag{3.85}$$

Otherwise, we iterate between Step 2 and Step 5 until convergence.

In practice, we choose h as proportional to $n^{-\frac{1}{r+4}}$ according to Silverman (1986). Based on the estimated $\hat{\mathcal{T}}_{1j},\ \forall j = 1, .., p_l$ and $\hat{\mathcal{T}}_{2j},\ \forall j = 1, .., p_r$, we ensemble them as $\hat{\Theta}_\alpha = \sum_{j=1}^{p_l} \hat{\mathcal{T}}_{1j} \hat{\mathcal{T}}_{1j}^{\ T}$ and $\hat{\Theta}_\beta = \sum_{j=1}^{p_r} \hat{\mathcal{T}}_{2j} \hat{\mathcal{T}}_{2j}^{\ T}$. Then we apply eigenvalue decomposition on $\hat{\Theta}_\alpha$ and $\hat{\Theta}_\beta$ to obtain a descending ordered eigenvalues as $(\hat{\lambda}_1, ..., \hat{\lambda}_{p_l})$ and $(\hat{\gamma}_1, ..., \hat{\gamma}_{p_r})$. Finally, we calculate the ratios of two consecutive eigenvalues as $\hat{r}_i = \hat{\lambda}_i / \hat{\lambda}_{i+1}$ for $i = 1, ..., p_l - 1$ and $\hat{q}_j = \hat{\gamma}_j / \hat{\gamma}_{j+1}$ for $j = 1, ..., p_r - 1$, and apply the maximal eigenvalue ration criterion proposed in contour projected dimension reduction (MERC; Luo, Wang and Tsai, 2009) to determine structural dimensions as

$$\begin{aligned} \hat{d}_l &= argmax_{1 \leq i \leq d_{lmax}} \{\hat{r}_i\} \\ \hat{d}_r &= argmax_{1 \leq j \leq d_{rmax}} \{\hat{q}_j\} \end{aligned} \tag{3.86}$$

In practice, one can set $d_{lmax} = d_{rmax} = 5$ as suggested by Luo, Wang and Tsai (2009).

3.5 Numerical Study

In this section, we compare folded SIR with L_1 regularization and folded SIR with both L_1 and L_2 regularizations, and least absolute shrinkage selection operator (Lasso; Tibshirani, 1996) on simulated non-linear models. We consider the case where sparseness exist on the coordinates of the basis matrices α, β of the corresponding *left folding subspace* $S_{Y|\circ\mathbf{X}}$ and *right folding subspace* $S_{Y|\mathbf{X}\circ}$, that is, the *central folding subspace* concludes with a much

smaller sub-predictor matrix which contains all the information about response variable Y. As our methodology part suggests, both folded SIR with L_1 regularization and folded SIR with both L_1 and L_2 regularizations are capable of achieving such variable selection, or more precisely, row selection and column selections from the predictor matrix. But with the later also imposing L_2 regularizations, it should be more robust dealing with unstable estimate of inverse of covariance matrix. For example, cases where sample size are less than the number of predictors in the predictor matrix, and where predictors are correlated with each other. We further break down into the examples with two types of sparsenesses. We assume that

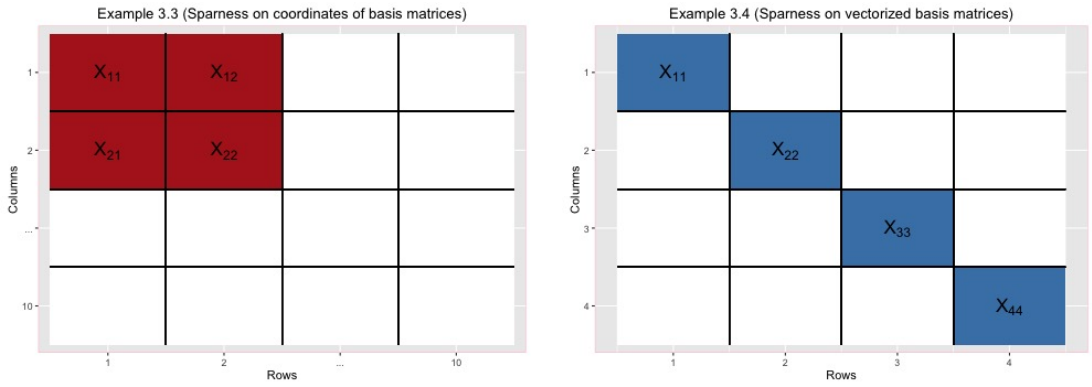

Figure 3.2: Visualization of true central folding space in Example 3.3 and Example 3.4

predictor matrix $\mathbf{X} \in \mathbb{R}^{p_l \times p_r}$, and that $p_l = p_r = p$, where p takes different values in order to account for different settings. We calculate True Positive Rate and False Positive Rate as performance measure for our methods. An ideal method should yield True Positive Rate close to 100% and False Positive Rate close to 0%. In the very first example, we focus on the forward model, where response variable Y_1 is connected with the predictor matrix as the following forward model:

Example 3.3

$$Y_1 = \mathbf{X}_{1,1} \times (\mathbf{X}_{1,2} + \mathbf{X}_{2,1} + 1) + 0.2\epsilon.$$

The vectorized predictor matrix follows a standard normal distributions, that is $vec(\mathbf{X}) \sim MVN_{p^2}(0, I)$. Error term ϵ also follows a standard univariate normal distribution. We further set $p = 10$ and sample size $n = 80$ so that there are $p^2 = 100$ elements in the predictor matrix, exceeding sample size.

Note that, in our example, the *folding central subspace* $S_{Y|\circ\mathbf{X}\circ}$ is larger than the *central subspace* $S_{Y|vec(\mathbf{X})}$ for vectorized data. The smallest sub-matrix that contains the entire information about response variable Y is the one with elements $\mathbf{X}_{1,1}$, $\mathbf{X}_{1,2}$, $\mathbf{X}_{2,1}$ and $\mathbf{X}_{2,2}$. Thus the desired *central folding space* $S_{Y|\circ\mathbf{X}\circ} = \beta \otimes \alpha$, where $\beta = \alpha = (e_1, e_2)$, and $e_j \in \mathbb{R}^p$ is a vector where its jth element is 1 and other elements 0. On the other hand, if we vecotorize the predictor, the corresponding *central subspace* $S_{Y|vec(\mathbf{X})} = span(B)$ where $B = (e_1, e_2 + e_{p+1}) \in \mathbb{R}^{p^2 \times 2}$ and $e_j \in \mathbb{R}^{p^2}$ is a vector where its jth element is 1 and other elements 0. Therefore, one can not directly compare *central folding subspace* $S_{Y|\circ\mathbf{X}\circ}$ and *central subspace* $S_{Y|vec(\mathbf{X})}$ since they are different spaces indeed. Note that this example is where sparseness exists in terms of both the coordinates of basis matrices, as well as the vectorized basis matrices. Both folded-SIR with L_1 regularization and folded-SIR with both L_1 and L_2 are capable to achieve such sparse solutions.

As for Lasso, note that it is an least square based method which imposes L_1 regularizations on the vectorized predictor $vec(\mathbf{X})$. According to its least square formulation, in terms of population solution, Lasso targets on space $\Sigma_x^{-1} \times Cov(vec(\mathbf{X}), Y_1) = \Sigma^{-1} \times E(vec(\mathbf{X}), Y_1) = \Sigma_x^{-1} \times e_1$, where $e_1 = (1, 0, ..., 0) \in \mathbb{R}^{p^2}$. Therefore, the OLS - based Lasso would fail to identify $X_{1,2}$, $X_{2,1}$ and $X_{2,2}$. In consequence, we should expect to see low TPR for all directions except $\beta_1 \otimes \alpha_1$.

Table 3.1 summarizes the results of TPR and FPR for three methods on Example 3.1. We apply a grid search strategy for selecting tuning parameters λ_α, λ_β in folded-SIR with L_1 regularization, ranging from $log_{10}(\lambda_\alpha) = -4$ to $log_{10}(\lambda_\alpha) = -1$ (same range for λ_β) with 0.2 increase in each step. The results indicated that AIC, BIC and RIC all agree on the same

optimal tuning parameters $log_10(\lambda_\alpha) = log_{10}(\lambda_\beta) = -2.6$. We take similar strategy when searching for tuning parameter τ_1, τ_2 in folded-SIR with L_2 regularization, as well as tuning parameter λ_1, λ_2 in folded-SIR with both L_1 and L_2 regularizations. GCV in folded-SIR with L_2 regularization points to $log_{10}(\tau_1) = log_{10}(\tau_2) = -2.2$ as the best combinations, and AIC, BIC and RIC all agree on $log(\lambda_1) = log(\lambda_2) = -1$ as the optimal choices for folded-SIR with both L_1 and L_2 regularizations.

In Table 3.1, we observe that traditional Lasso model performs great in terms of TPR when estimating the linear term $\mathbf{X}_{11}$ indicated by direction $\beta_1 \otimes \alpha_1$, but Lasso does not always recover the other two directions $\beta_2 \otimes \alpha_1$ and $\alpha_2 \otimes \beta_2$ as their corresponding TPR are as low as 3%. On the other hand, folded-SIR with L_1 regularization provide better TPR on the estimation of the associated non-linear terms in the model indicated by direction $\beta_2 \otimes \alpha_1$, $\beta_2 \otimes \alpha_2$, by sacrificing a larger FPR on other directions and smaller TPR on direction $\beta_1 \otimes \alpha_1$. folded-SIR with both L_1 and L_2 regularizations performs similarly, but with higher TPR on the true directions, and a smaller FPR compared with folded-SIR with L_1 regularization. This is probably due to the fact that folded-SIR with both L_1 and L_2 regularizations is especially designed to achieve sparseness on the coordinates of basis matrices, and it also requires much smaller number of unknown parameters $(p_l + p + r)$ compared with folded-SIR with L_1 regularization $(p_l \times d_l + p_r \times d_r)$.

Table 3.1: Average of TPR and FPR based on 100 replications for example 3.3

n=80, p=10, uncorrelated	TPR				FPR
Method	$\beta_1 \otimes \alpha_1$	$\beta_1 \otimes \alpha_2$	$\beta_2 \otimes \alpha_1$	$\beta_2 \otimes \alpha_2$	Average
folded-SIR L_1	43.00%	31.00%	32.00%	25.00%	24.23%
folded-SIR, L_1 + L_2	83.00%	62.00%	60.00%	46.00%	18.64%
Lasso	100.00%	96.00%	3.00%	11.00%	6.26%

Example 3.4

$$Y_1 = \mathbf{X}_{1,1} \times (\mathbf{X}_{2,2} + \mathbf{X}_{3,3} + 1) + \mathbf{X}_{4,4} + 0.2\epsilon.$$

We also consider Example 3.4 where sparseness exists in the vectorized basis matrices, but not in the coordinates of basis matrices. The smallest sub-matrix containing the predictor elements that are associated with the response variable is as the same as the original matrix predictors, i.e., $S_{Y|\circ\mathbf{X}\circ} = \beta \otimes \alpha$, where $\beta = \alpha = (e_1, e_2, e_3, e_4) \in \mathbb{R}^{p\times 4}$. Folded-SIR with L_1 regularization in theory should be capable of estimating such sparse vectorized basis matrices. Folded-SIR with both L_1 and L_2 on the other hand, aiming at suppress some of the coordinates of basis matrices as 0, may fail to estimate the desired *central folding subspace* $S_{Y|\circ\mathbf{X}\circ}$ correctly.

Table 3.2: Average of TPR and FPR based on 100 replications for example 3.4

n=80, p=10, uncorrelated	TPR				FPR
Method	$\beta_1 \otimes \alpha_1$	$\beta_2 \otimes \alpha_2$	$\beta_3 \otimes \alpha_3$	$\beta_4 \otimes \alpha_4$	Average
folded-SIR L_1	79.00%	0.00%	59.00%	98.00%	51.92%
folded-SIR, L_1 + L_2	18.00%	0.00%	0.00%	100.00%	3.00%
Lasso	100.00%	29.00%	24.00%	23.00%	29.33%

As shown in Table 3.2, Lasso again provides 100% TPR on the estimation of $\beta_1 \otimes \alpha_1$, corresponding to the linear term $\mathbf{X}_{1,1}$ in Example 3.4. Because Lasso is an OLS formulated estimation method, it provides much lower TPR on non-linear terms $\mathbf{X}_{1,1} \times \mathbf{X}_{3,2}$ and $\mathbf{X}_{1,1} \times \mathbf{X}_{3,3}$, indicated by directions $\beta_2 \otimes \alpha_2$ and $\beta_3 \otimes \alpha_3$. Surprisingly, Lasso also fails to provide high TPR on the estimation of another linear term $\mathbf{X}_{4,4}$ in the model, as its TPR ($\beta_4 \otimes \alpha_4$) is as low as 23%. Compared with Lasso, folded-SIR with L_1 performs much better on the estimation of $\beta_3 \otimes \alpha_3$, $\beta_4 \otimes \alpha_4$, which corresponds to non-linear term $\mathbf{X}_{1,1} \times \mathbf{X}_{3,3}$ and linear term $\mathbf{X}_{4,4}$ in Example 3.4. Folded-SIR with L_1 regularization does not recover $\mathbf{X}_{1,1} \times \mathbf{X}_{2,2}$ at all in the simulation. The better TPR performance of folded-SIR with L_1 regularization comes with the price of higher FPR as its average FPR is 51.92% which is larger than that of Lasso method. Last but not least, folded-SIR with both L_1 and L_2 regularizations performs the worst, as it completely overlook directions $\beta_2 \otimes \alpha_2$ and $\beta_3 \otimes \alpha_3$ and provide

low TPR on $\beta_1 \otimes \alpha_1$ as well. It does achieve 100% TPR on $\beta_4 \otimes \alpha_4$ and a much lower FPR compared with the other two methods. Our simulations in Example 3.3 and Example 3.4 seem to provide evidence to suggest implementing folded-SIR with L_1 regularizations when sparseness exist in vectorized basis matrices, and folded-SIR with both L_1 and L_2 regularizations when sparseness exist in coordinates of basis matrices.

Table 3.3: Percentages of estimated AIC, BIC and RIC structural dimensions

AIC ($d_l = d_r = 2$)	$\hat{d}_r = 1$	$\hat{d}_r = 2$	$\hat{d}_r = 3$	$\hat{d}_r = 4$	$\hat{d}_r = 5$
$\hat{d}_l = 1$	0%	0%	0%	0%	0%
$\hat{d}_l = 2$	0%	0%	0%	0%	0%
$\hat{d}_l = 3$	0%	0%	0%	0%	0%
$\hat{d}_l = 4$	0%	0%	0%	10%	0%
$\hat{d}_l = 5$	0%	0%	0%	**90%**	0%
BIC ($d_l = d_r = 2$)	$\hat{d}_r = 1$	$\hat{d}_r = 2$	$\hat{d}_r = 3$	$\hat{d}_r = 4$	$\hat{d}_r = 5$
$\hat{d}_l = 1$	0%	0%	0%	0%	0%
$\hat{d}_l = 2$	0%	0%	1%	0%	0%
$\hat{d}_l = 3$	0%	0%	3%	3%	0%
$\hat{d}_l = 4$	0%	0%	3%	22%	0%
$\hat{d}_l = 5$	0%	0%	0%	**68%**	0%
RIC ($d_l = d_r = 2$)	$\hat{d}_r = 1$	$\hat{d}_r = 2$	$\hat{d}_r = 3$	$\hat{d}_r = 4$	$\hat{d}_r = 5$
$\hat{d}_l = 1$	36%	2%	0%	0%	0%
$\hat{d}_l = 2$	5%	**57%**	0%	0%	0%
$\hat{d}_l = 3$	0%	0%	0%	0%	0%
$\hat{d}_l = 4$	0%	0%	0%	0%	0%
$\hat{d}_l = 5$	0%	0%	0%	0%	0%

Finally, we examine the estimation of structural dimensions d_l and d_r through criteria including empirical AIC, BIC and RIC. We alter Example 3.3 with the same model but reduce the dimensions to $p_l = p_r = p = 5$. In this setting, the true structural dimensions are $d_l = d_r = 2$. Table 3.3 lists the proportion of estimated structural dimensions d_l and d_r based on 100 iterations using empirical AIC, BIC and RIC criteria. One can immediately recognize that AIC tends to be over-conservative on estimating the dimension, as the majority of the estimations (90%) locate the structural dimensions as $d_l = 5$ and $d_r = 4$. BIC tends to behave similarly as AIC, with its majority of the estimation (68%) with dimensions $d_l = 5$ and $d_r = 4$ as well. RIC, on the other hand, provides the best performance as most of its

iterations (57%) agree on $d_l = d_r = 2$ as structural dimensions. In practice, one may consider using empirical RIC as a standard criterion for estimating structural dimensions d_l and d_r for folding methods.

3.6 Application

3.6.1 Primary Biliary Cirrhosis longitudinal data

We continue to use the longitudinal Primary Biliary Cirrhosis (PBC) dataset provided by Mayo Clinic, as we did in the second chapter. The dataset was collected in a follow up experiment conducted between 1974 and 1984 with 312 patients participated in the experiment. We convert the longitudinal dataset as matrix predictor by regarding the repeated measurements of the the bio-markers collected at different visit time points into matrix format for each patient. According to Xue and Yin (2014), visits between 90 days and 270 days are treated as 6-month, 270 - 550 days as 1 year, 550 - 910 days as 2 years and 910 -1275 as 3 years. The columns of the predictor matrix consist of three types of bio-markers including bilirubin, albumin and prothrombin time. The variables of interest include:

Response variable: Univariate, years between patients' registration to the earlier of transplanting or death.

Predictor variable: 3×4 matrix predictor, including columns of the predictor as the discrete visit time from subjects, and rows of as three types of bio-markers measurement.

We apply both folded-SIR with L_1 regularizations and folded-SIR with both L_1 and L_2 regularization on PBC data to acheieve sparse estimates on the corresponding *central fold-*

ing subspace. As suggested in Xue and Yin (2014), we restricted the structural dimension as $d_l = d_r = 1$. The following equation summarizes the estimated directions from the two regularized models: folded-SIR with L_1 regularizations and folded-SIR with both L_1 and L_2 regularizations. We included results from folded-SIR as a benchmark comparison.

Folded-SIR with L_1 regularization:

$$\begin{aligned} \hat{\alpha} &= (0.0418, -0.8950, 0)^T \\ \hat{\beta} &= (0, 0, 0, -0.9789)^T \end{aligned} \tag{3.87}$$

Folded-SIR with both L_1 and L_2 regularization:

$$\begin{aligned} \hat{\alpha} &= (0.0723, -0.7869, 0)^T \\ \hat{\beta} &= (0, 0, 0, -0.9693)^T \end{aligned} \tag{3.88}$$

Folded-SIR:

$$\begin{aligned} \hat{\alpha} &= (0.0530, -0.9948, -0.0868)^T \\ \hat{\beta} &= (-0.0598, -0.0631, -0.2657, -0.9601)^T \end{aligned} \tag{3.89}$$

The extracted non-zero coordinates for directions α and β are α_1, α_2 and β_3 for both folded-SIR with L_1 regularization and folded-SIR with both L_1 and L_2 regularizations. The non-zero coordinates can be interpreted that the combinations of the first bio-marker (bilirubin bio-marker) and second bio-marker (albumin bio-marker) at the fourth time point (3 years) have major positive effects on patients' time between registration to the earlier of transplanting or death.

Our results are somewhat consistent with the results from Xue and Yin (2014), as they applied folded-OPG in this analysis with a slightly larger sample size. In their bootstrap

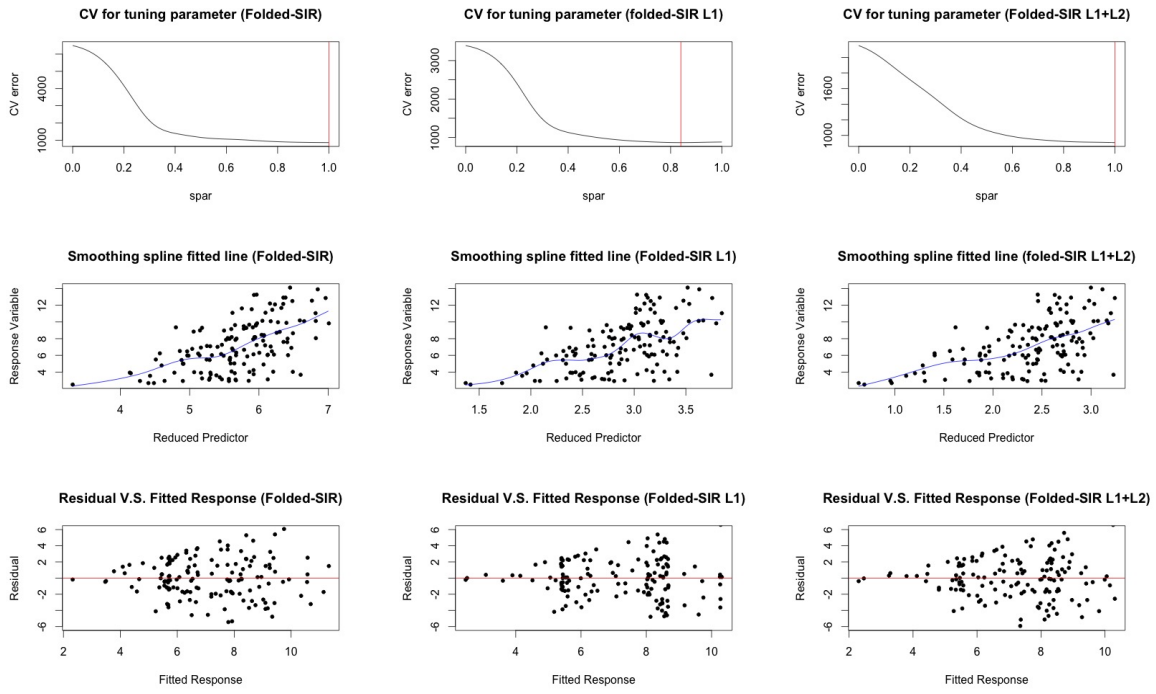

Figure 3.3: Summary of smoothing splines with reduced predictors for three models

confidence interval for the directions α and β, only the intervals for coordinates α_1, α_2 and β_4 do not contain 0, which correspond exactly to the non-zeros coordinates we estimate through folded-SIR with both L_1 regularization and with both L_1 and L_2 regularizations.

Finally, we implement the sparse directions to calculate the reduced predictor as $\hat{\alpha}^T \mathbf{X} \hat{\beta} \in \mathbb{R}^{1\times 1}$, then apply smoothing splines to fit the reduced predictor against the response variable. The following plots summarize the fitted models for folded-SIR, folded-SIR with L_1 regularization and folded-SIR with both L_1 and L_2 regularizations. The first row of the figures provides the cross-validation process to search for the optimal tuning parameter in smoothing splines models. The second row provides plots as residual versus fitted response and the last row plot out the fitted non-linear splines. Both folded-SIR with L_1 regularizations and folded-SIR with both L_1 and L_2 regularizations achieve similar accuracy as folded-SIR, as

indicated by the residual plots, while using smaller number of predictors from the original predictor matrix.

3.7 Discussion

In this chapter, we provide an extension of regularized dimension folding methods according to its least square formulation. We introduce two types of sparseness for matrix predictors and develop method accordingly. In particular, we develop dimension folding with L_1 regularization to facilitate sparse estimate on the desired *central folding subspace*. Dimension folding with L_2 regularization is then introduced to address cases when dimensions of predictor $p_l p_r$ is large than the sample size n, or when high correlation exists in predictor matrix. Dimension folding with both L_2 and L_1 regularization is also implemented to simultaneously impose reduction estimation as well as variable selection. Our simulation experiences and real data application indicates the methods tend to reduce the matrix predictor into a smaller sub-matrix compared with other traditional methods like Lasso. We believe that such methodology similar framework can also be adapted to other folding related estimators such as folded-OPG (Xue and Yin, 2014) and dimension folding principal fitted components (DF-PFC) (Ding and Cook, 2014).

3.8 Appendix

3.8.1 Derivation of folded-SIR with L_2 regularization when $B = \hat{\Sigma}_x^{-1}$

We demonstrate in the following result that even if we use the original B as $B = \hat{\Sigma}_x^{-1}$ in the objective function (3.28) from Li, Kim and Altman (2010), the actual computation of L_2 regularized solution for folded-SIR does not involve the calculation of inverse $\hat{\Sigma}_x^{-1}$. First

of all, instead of the definition of (3.30) defined previously, we still use $\hat{\Sigma}_x^{-1}$ in the objective function, and define that,

Definition 5 *For non-negative constant τ_1 and τ_2, let*

$$\begin{aligned} G_{\tau_1,\tau_2}(\alpha,\beta,C) = &\sum_{y=1}^{h} \hat{p}_y\{(vec(\bar{\mathbf{X}}_y) - vec(\bar{\mathbf{X}})) - \hat{\Sigma}_x(\beta\otimes\alpha)C_y\}^T \\ &\times \hat{\Sigma}_x^{-1}\{(vec(\bar{\mathbf{X}}_y) - vec(\bar{\mathbf{X}})) - \hat{\Sigma}_x(\beta\otimes\alpha)C_y\} \\ &+ \tau_1 vec(\alpha)^T vec(\alpha) + \tau_2 vec(\beta)^T vec(\beta). \end{aligned} \tag{3.90}$$

Let $(\hat{\alpha},\hat{\beta},\hat{C}) = argmin_{\alpha,\beta,C} G_{\tau_1,\tau_2}(\alpha,\beta,C)$, then $span(\hat{\beta}\otimes\hat{\alpha})$ is called a ridge folded-SIR estimator of the central folding subspace *with respect to (τ_1,τ_2).*

Thus, under this definition, the alternating least-square algorithm become that:

For fixed α and β

$$\begin{aligned} &\hat{C} = (\hat{C}_1,...,\hat{C}_h), \quad where \\ &\hat{C}_y = ((\beta^T\otimes\alpha^T)\hat{\Sigma}_x(\beta\otimes\alpha))^{-1}(\beta^T\otimes\alpha^T)vec(\bar{\mathbf{X}}_y - \bar{\mathbf{X}}) \quad y = 1,...,h. \end{aligned} \tag{3.91}$$

For fixed β and C

$$\begin{aligned} vec(\hat{\alpha}) &= (\tilde{X}^T\tilde{W}\tilde{X} + \tau_1 I_{p_l d_l})^{-1}\tilde{X}^T\tilde{W}\tilde{Y} \\ &= (U^T(D_f\otimes\hat{\Sigma}_x)U + \tau_1 I_{p_l d_l})^{-1}U^T(D_f\otimes I_{p_l p_r})\tilde{Y}, \end{aligned} \tag{3.92}$$

where U, D_f, $\tilde{Y}$ and $\tilde{X}$ keep the same notation previously, but we change $\tilde{W}$ as

$$\tilde{W} = D_f\otimes\hat{\Sigma}_x^{-1}. \tag{3.93}$$

Thus, instead of calculating the inverse of a $p_l p_r \times p_l p_r$ matrix $\hat{\Sigma}_x$, $U^T(D_f \otimes \hat{\Sigma}_x)U + \tau_1 I_{p_l d_l}$ is instead a $p_l d_l \times p_l d_l$ matrix, which is usually invertible when the desired structural dimensions d_l and d_r are far less than p_l and p_r, the ridge regression type term $\tau_1 I_{p_l d_l}$ also ensure the inverse calculation to remain stable.

For fixed α and C

The solution for β as

$$\begin{aligned} vec(\hat{\beta}) &= (\tilde{\tilde{X}}^T \tilde{W} \tilde{\tilde{X}} + \tau_2 I_{p_r d_r})^{-1} \tilde{\tilde{X}}^T \tilde{W} \tilde{Y} \\ &= (V^T(D_f \otimes \hat{\Sigma}_x)V + \tau_2 I_{p_r d_r})^{-1} V^T (D_f \otimes I_{p_l p_r}) \tilde{Y}, \end{aligned} \tag{3.94}$$

where $\tilde{Y}$ and $\tilde{W}$ follows the same as the solution for α, and V, D_f, transformation matrix K_{p_d,d_r} keep the same notation previously defined, and finally $\tilde{\tilde{X}} = (I_h \otimes \Sigma_x)V$. Similarly, the $p_r d_r \times p_r d_r$ matrix $V^T(D_f \otimes \hat{\Sigma}_x)V + \tau_2 I_{p_r d_r}$ contain much smaller dimensions compared with $\hat{\Sigma}_x$, and its ridge regression type term $\tau_2 I_{p_r d_4}$ makes sure the stability when inverting it.

We still compute solution for $\hat{C}$, $\hat{\alpha}$, and $\hat{\beta}$ until the objective function stabilizes. The GCV statistic also changes when selecting tuning parameters τ_1 and τ_2. In fact, if we further changes $\tilde{W}^{\frac{1}{2}}$ in the previous section, as $\tilde{W}^{\frac{1}{2}} = D_f^{\frac{1}{2}} \otimes \hat{\Sigma}_x^{-\frac{1}{2}}$, then we define a generalized cross-validation criterion (GCV) as the following

$$GCV = \frac{||(I_{p_l p_r h} - S_{\tau_1})\tilde{W}^{1/2}\tilde{Y}||^2 + ||(I_{p_l p_r h} - S_{\tau_2})\tilde{W}^{1/2}\tilde{Y}||^2}{2p_l p_r h\{1 - \frac{trace(S_{\tau_1}) + trace(S_{\tau_2})}{2p_l p_r h}\}^2}, \tag{3.95}$$

where

$$\begin{aligned} S_{\tau_1} &= (D_f \otimes \hat{\Sigma}_x^{\frac{1}{2}})U(U^T(D_f \otimes \hat{\Sigma}_x)U + 2\tau_1 I_{p_l d_l})^{-1} U^T (D_f \otimes \hat{\Sigma}_x^{\frac{1}{2}}), \\ S_{\tau_2} &= (D_f \otimes \hat{\Sigma}_x^{\frac{1}{2}})V(V^T(D_f \otimes \hat{\Sigma}_x)V + 2\tau_2 I_{p_r d_r})^{-1} V^T (D_f \otimes \hat{\Sigma}_x^{\frac{1}{2}}). \end{aligned} \tag{3.96}$$

We select values of tuning parameters τ_1 and τ_2 that minimizes (3.95) as the optimal choice for folded-SIR with L_2 regularization.

3.8.2 Derivation of GCV statistic

First notice that the equivalence of the following two types of penalized regression problems. That is, for response variable $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, a diagonal weight matrix $\Lambda \in \mathbb{R}^{p\times p}$ with all its diagonal elements non-negative, a non-negative tuning parameter $\lambda \geq 0$, and unknown $B \in \mathbb{R}^p$, minimizing

$$\frac{1}{n}(Y - XB)^T(Y - XB) + \lambda B^T \Lambda B, \tag{3.97}$$

over β is equivalent to minimizing

$$\frac{1}{n}(Y - XB)^T(Y - XB), \tag{3.98}$$

with constraint that $B^T \Lambda B \leq \lambda$. And the solution for B can be stated as:

$$\hat{B}(X^T X + n\lambda\Lambda)^{-1} X^T Y. \tag{3.99}$$

Therefore, keeping the same notation of U, V, $\tilde{Y}$, $\tilde{X}$, $\tilde{\tilde{X}}$ and $\tilde{W}$. If we denote

$$\begin{aligned}
\mathring{Y} &= \begin{pmatrix} \tilde{Y} \\ \tilde{Y} \end{pmatrix} \in \mathbb{R}^{2p_l p_r h}, \\
\mathring{X} &= \begin{pmatrix} \tilde{X} & 0_{p_l p_r h \times p_r d_r} \\ 0_{p_l p_r h \times p_l d_l} & \tilde{\tilde{X}} \end{pmatrix} \in \mathbb{R}^{2p_l p_r h \times (p_l d_l + p_r d_r)}, \\
\mathring{W} &= \begin{pmatrix} \tilde{W} & 0_{p_l p_r h \times p_l p_r h} \\ 0_{p_l p_r h \times p_l p_r h} & \tilde{W} \end{pmatrix} \in \mathbb{R}^{2p_l p_r h \times 2p_l p_r h}, \\
\mathring{W}^{\frac{1}{2}} &= \begin{pmatrix} \tilde{W}^{\frac{1}{2}} & 0_{p_l p_r h \times p_l p_r h} \\ 0_{p_l p_r h \times p_l p_r h} & \tilde{W}^{\frac{1}{2}} \end{pmatrix} \in \mathbb{R}^{2p_l p_r h \times 2p_l p_r h}, \\
\mathring{B} &= [vec(\alpha)^T, vec(\beta)^T]^T \in \mathbb{R}^{p_l d_l + p_r d_r}, \\
\mathring{\Lambda} &= diag(\tau_1, ..., \tau_1, \tau_2, ...\tau_2) \in \mathbb{R}^{(p_l d_l + p_r d_r) \times (p_l d_l + p_r d_r)},
\end{aligned} \tag{3.100}$$

where diagonal matrix $\mathring{\Lambda}$ has its first $p_l d_l$ diagonal elements as τ_1 and its last $p_r d_r$ diagonal elements as τ_2. We can rewrite the objective function (3.30) as:

$$(\mathring{Y} - \mathring{X}\mathring{B})^T \mathring{W} (\mathring{Y} - \mathring{X}\mathring{B}) + 2\mathring{B}^T \mathring{\Lambda} \mathring{B}. \tag{3.101}$$

Further note that in Golub, Heath and Wahb(1979), they defined the generalized cross validation (GCV) statistic as

$$V(\lambda) = \frac{1/n||I_n - A(\lambda)Y||^2}{[1/n \quad Trace(I_n - A(\lambda))]^2}, \tag{3.102}$$

where $A(\lambda) = X(X^TX + n\lambda I_p)^{-1}X^T$. Similarly to $A(\lambda)$, if we replace X by $\mathring{W}^{\frac{1}{2}}\mathring{X}$ and $n\lambda I_p$ by $2\mathring{\Lambda}$, then we can define

$$\begin{aligned} S_{\tau_1,\tau_2} &= \mathring{X}(\mathring{X}^T\mathring{X} + 2\mathring{\Lambda})^{-1}\mathring{X}^T \\ &= \begin{pmatrix} \tilde{W}^{\frac{1}{2}}\tilde{X}(\tilde{X}^T\tilde{W}\tilde{X} + 2\tau_1 I_{p_l d_l})^{-1}\tilde{X}^T\tilde{W}^{\frac{1}{2}} & 0_{p_l p_r h \times p_l p_r h} \\ 0_{p_l p_r h \times p_l p_r h} & \tilde{W}^{\frac{1}{2}}\tilde{X}(\tilde{X}^T\tilde{W}\tilde{X} + 2\tau_1 I_{p_l d_l})^{-1}\tilde{X}^T\tilde{W}^{\frac{1}{2}} \end{pmatrix}. \end{aligned} \tag{3.103}$$

If we denote,

$$\begin{aligned} S_{\tau_1} &= (D_f \otimes \hat{\Sigma}_x^{\frac{1}{2}})U(U^T(D_f \otimes \hat{\Sigma}_x^2)U + 2\tau_1 I_{p_l d_l})^{-1}U^T(D_f \otimes \hat{\Sigma}_x^{\frac{1}{2}}), \\ S_{\tau_2} &= (D_f \otimes \hat{\Sigma}_x^{\frac{1}{2}})V(V^T(D_f \otimes \hat{\Sigma}_x^2)V + 2\tau_2 I_{p_r d_r})^{-1}V^T(D_f \otimes \hat{\Sigma}_x^{\frac{1}{2}}), \end{aligned} \tag{3.104}$$

then S_{τ_1,τ_2} can be written as

$$S_{\tau_1,\tau_2} = \begin{pmatrix} S_{\tau_1} & 0_{p_l p_r h \times p_l p_r h} \\ 0_{p_l p_r h \times p_l p_r h} & S_{\tau_2} \end{pmatrix} \in \mathbb{R}^{2p_l p_r h \times 2p_l p_r h}. \tag{3.105}$$

Thus, by replacing sample size n by $2p_l p_r h$, we can also define GCV statistic as

$$\begin{aligned} GCV &= \frac{||I_{2p_l p_r h} - S_{\tau_1,\tau_2}\mathring{W}^{\frac{1}{2}}\mathring{Y}||^2}{2p_l p_r h(1 - trace(S_{\tau_1,\tau_2})/2p_l p_r h)^2} \\ &= \frac{||(I_{p_l p_r h} - S_{\tau_1})\tilde{W}^{1/2}\tilde{Y}||^2 + ||(I_{p_l p_r h} - S_{\tau_2})\tilde{W}^{1/2}\tilde{Y}||^2}{2p_l p_r h\{1 - \frac{trace(S_{\tau_1}) + trace(S_{\tau_2})}{2p_l p_r h}\}^2}. \end{aligned} \tag{3.106}$$

3.8.3 Alternative estimation of structural dimensions

We propose an alternative estimation of structural dimensions d_l and d_r, similar to that of Li and Yin (2008). Following the methodology part of folded-OPG criterion, assume that

we have acquired the estimates of $\hat{\mathcal{T}}_{1j}$ and $\hat{\mathcal{T}}_{2j}$ for $j = 1, ..., n$. If we denote:

$$\begin{aligned} \hat{\Omega}_\alpha &= \sum_{j=1}^{n} \hat{\mathcal{T}}_{1j} \hat{\mathcal{T}}_{1j}^{\ T} + I_{p_l}, \\ \hat{\Omega}_\beta &= \sum_{j=1}^{n} \hat{\mathcal{T}}_{2j} \hat{\mathcal{T}}_{2j}^{\ T} + I_{p_r}, \end{aligned} \tag{3.107}$$

and further denote the the eigenvalues of $\hat{\Omega}_\alpha$ as $\hat{\delta}_1, ..., \hat{\delta}_{p_l}$, and that the eigenvalues of $\hat{\Omega}_\beta$ as $\hat{\theta}_1, ..., \hat{\theta}_{p_r}$. Following Li and Yin (2008) and Zhu, Miao and Peng (2006), we can use the following BIC type estimator for d_l and d_r:

$$\begin{aligned} \hat{d}_l &= \underset{m \in \{0,1...,p_l-1\}}{arg\ max} \{ \frac{n}{2} \sum_{i=1+min(k,m)}^{p_l} (log(\hat{\delta}_i) + 1 - \hat{\delta}_i) - \frac{C_n m(2p_l - m + 1)}{2} \}, \\ \hat{d}_r &= \underset{m \in \{0,1...,p_r-1\}}{arg\ max} \{ \frac{n}{2} \sum_{i=1+min(k,m)}^{p_r} (log(\hat{\theta}_i) + 1 - \hat{\theta}_i) - \frac{C_n m(2p_r - m + 1)}{2} \}, \end{aligned} \tag{3.108}$$

where C_n is a penalty constant which takes various forms according to Zhu, Miao and Peng (2006). They claim that $C_n = O(n^a)$, and a should be within certain range to ensure the weak convergence and strong convergence for their proposed estimator. In simulation studies, they try different versions of C_n, and take the one with best performance. Similarly, we also take $C_n = W_n h/n$, where $W_n = 0.1log(n), 0.5log(n), 0.1n^{1/3}, 0.5n^{1/3}, (0.5log(n) + 0.1n^{1/3})$, and $(0.5log(n) + 0.1n^{1/3})/2$.

Bibliography

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory, B. N. Petrov and F. Csaki (eds)*, 267-281, 1973.

[2] X. Chen, C. Zou and R. D. Cook. Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics* **38**, 36963723, 2010.

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

[4] R. D. Cook. Testing predictor contribution in sufficient dimension reduction. *The Annals of Statistics* **32**, 1062-1092, 2004.

[5] R. D. Cook and S. Weisberg. Discussion of "Sliced inverse regression for dimension reduction". *Journal of the American Statistical Association* **86**, 28-33, 1991.

[6] S. Ding and R. D. Cook. Dimension folding PCA and PFC for matrix-valued predictors. *Statistica Sinica* **24**, 463-492, 2014.

[7] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani. Least angle regression. *The Annals of Statistics* **32**, 407-499, 2004.

[8] J. Fan and I. Gijbels. Local polynomial modeling and its application. *Chapman and Hall*, 1996.

[9] G.H. Golub, M. Heath and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215-223, 1979.

[10] I.T. Jolliffe, N.T. Trendafilov and M. Uddin. A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics* **12**, 531-547, 2003.

[11] B. Li, M. Kim and N. Altman. On dimension folding of matrix- or array-valued statistical objects. *The Annals of Statistics* **38**, 1094-1121, 2010.

[12] B. Li, and S. Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association* **102**, 2143-2172, 2007.

[13] K.-C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**, 316-342, 1991.

[14] L. Li. Sparse sufficient dimension reduction. *Biometrika* **94**, 603-613, 2007.

[15] L. Li. and C. J. Nachtsheim. Sparse sliced inverse regression. *Technometrics* **48**, 503510, 2006.

[16] L. Li and X. Yin. Sliced inverse regression with regularizations. *Biometrics* **64**, 124-131, 2008.

[17] L. Ni, R. D. Cook and C. L. Tsai. A note on shrinkage sliced inverse regression. *Biometrika* **92**, 242247, 2005.

[18] P. Shi and C.-L. Tsai. Regression model selection - a residual likelihood approach. *Journal of the Royal Statistical Society, Series B* **64**, 237- 252, 2002.

[19] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics* **6**, 461-464, 1978.

[20] S. Shan, B. Cao, Y. Su, L. Qing, X. Chen and W. G. Unified principal component analysis with generalized covariance matrix for face recognition. *IEEE Conf. on Comp. Vis. and Pat. Recog* **13**, 1-7, 2008.

[21] B. W. Silverman. Density Estimation for Statistics and Data Analysis. *Chapman and Hall*, 1986.

[22] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical Society, Series B* **58**, 267-288, 1996.

[23] Q. Wang and X. Yin A nonlinear multi-dimensional variable selection method for high dimensional data: sparse MAVE. *Computational Statistics and Data Analysis* **52**, 45124520, 2008.

[24] Y. Xia, H. Tong, W.K. Li and L. Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society, Series B* **64**, 363-410, 2002.

[25] Y. Xue and X. Yin. Sufficient dimension folding for regression mean function. *Journal of Computational and Graphical Statistics* **39**, 1028-1043, 2014.

[26] J. Yang, D. Zhang, A.F. Frangi and J. Yang. Two dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Trans. Patter Anal. Mach. Intell* **69**, 224-231, 2005.

[27] X. Yin and H. Hilafu. Sequential sufficient dimension reduction for large p, small n problems. *Journal of the Royal Statistical Society, Series B* **4**, 879-892, 2015.

[28] D. Zhang and Z. Zhou. (2D)2PCA: Two-directional two-dimensional PCA for efficient face representation and recognition. *Neurocomp* **26**, 131-137, 2004.

[29] J. Zhou and X. He. Dimension reduction based on constrained canonical correlation and variable filtering. *Annals of Statistics* **36**, 16491668, 2008.

[30] L. Zhu, B. Miao and H. Peng. On sliced inverse regression with large dimensional covariates. *Journal of American Statistical Association* **101**, 630-643, 2006.

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

Chapter 4

Testing Variable Contributions in Sufficient Dimension Folding

Abstract

Matrix/array predictors have aroused recent interest for its complicated structure and wide application in real world data such as images and videos. How to effectively reduce dimensions of predictor matrix/array, provide variable selection, as well as assess predictor variables' contribution to response variable remain challenging tasks. For vector predictor variable, two types of methodologies are proven to perform such tasks. One framework falls into the category of regularized sufficient dimension reduction, which produces sparse estimates on *central subspace* by introducing regularizations in estimation procedure, thus achieves variable selection purpose. Another approach is through hypothesis testing framework developed by Cook (2004). Based on sufficient dimension folding designed for reducing dimensions for matrix/array data, we extend hypothesis testing methods into folding methods to help evaluate predictors' contribution to response variable. We define three tests, namely, marginal dimensional hypothesis, marginal and conditional coordinate hypotheses.

We provide future directions to construct associated test statistics and asymptotic theories for these tests.

4.1 Introduction

Big Data often come with both high dimensions and complicated structures. Examples of such includes images and videos whose inner structures are matrix and higher dimensional arrays. How to effectively reduce the number of predictors while preserving the data structure remains a challenging task. While traditional methods seek to reduce the dimensions by converting the matrix/array data into vector data, Li, Kim and Altman (2010) on the other hand, pointed out that such simplification may lose sufficient information, as well as making it difficult to interpret. They proposed the framework of sufficient dimension folding, aligned with sufficient dimension reduction, for matrix/array-valued objects and developed folded-SIR, folded-SAVE and folded-DR correspondingly.

Nevertheless, the estimated *central folding subspace* (Li, Kim and Altman, 2010) for matrix/array data still contain linear combinations of folds from the predictor matrix/array, making it difficulty to interpret and provide insights on fewer numbers of predictor variables. Also, there lacks an effective method to estimate structural dimensions for the *central folding subspace*, except an AIC type test statistics proposed by Ding and Cook (2014) which focused on the maximum likelihood estimates on *central folding subspace*, and a cross-validation based method described in folded-OPG method by Xue and Yin (2014). To address these problems, two types of frameworks have proven to be useful. One framework is regularized sufficient dimension reduction. For vector predictor, methods proposed by Ni, Cook and Tsai (2005), Li and Nachtsheim (2006), Li (2007), Zhou and He (2008), Wang and Yin (2008), Chen, Zou and Cook (2010) and Li and Yin (2008) fall into such framework. Yin and Hilafu (2015) provided an alternative approach called sufficient variable selection which focuses

on selecting variables that are associated with response variable, and proposed two paths to estimate relevant predictor variables especially for large p small n problems. Another approach to tackle this problem is through hypothesis testing framework proposed by Cook (2004) including three tests: marginal dimension hypothesis test, marginal and conditional coordinate hypotheses. In this chapter, we extend the hypothesis testing method into matrix predictors.

The remaining contents of this paper are organized as follows: In Section 4.2, we briefly review on the methodology of testing predictor contributions in sufficient dimension reduction (Cook, 2004). Section 4.3 extends the definitions of three tests in Cook (2004) for matrix predictor. We end with Section 4.3 that summarizes our future work including constructing corresponding test statistics, building up asymptotic properties of the test statistics, etc.

4.2 Review on testing predictor contribution in sufficient dimension reduction

We first focus on dimension reduction for vector data $X \in \mathbb{R}^p$, and provide a review on how to test predictors contributions in sufficient dimension reduction from Cook (2004). Cook (2004) summarized three types of tests for testing either dimension or predictor contribution for the reduced data using SDR methods. They are:

- Marginal dimension hypotheses

  $H_0 : d = m$ against $H_1 : d > m$;

- Marginal coordinate hypotheses

  $H_0 : P_{\mathcal{H}} S_{Y|X} = \mathcal{O}_p$ against $H_1 : P_{\mathcal{H}} S_{Y|X} \neq \mathcal{O}_p$;

- Conditional coordinate hypotheses

  $H_0 : P_{\mathcal{H}} S_{Y|X} = \mathcal{O}_p$ against $H_1 : P_{\mathcal{H}} S_{Y|X} \neq \mathcal{O}_p$ given d.

The only difference between marginal coordinate hypotheses test and conditional coordinate hypotheses test is that the latter test assumes the structural dimension$d$ is given. In the defined tests, $\mathcal{H}$ is a user-selected subspace of the original predictor space, $\mathcal{O}_p$ is the origin in R_p. $P_{\mathcal{H}} S_{Y|\mathbf{X}} = \mathcal{O}_p$ simply says user-selected subspace $\mathcal{H}$ has no "contribution" towards the response variable Y. In practice, if one needs to test whether the r selected predictors X_2 from the original data partition $X^T = (X_1^T, X_2^T)$ contribute to the regression, one can denote $\mathcal{H} = span((0, I_r)^T)$, and test if $Y \perp\!\!\!\perp X|\eta_1^T X_1$, which is equivalent to test if $P_{\mathcal{H}} S_{Y|X} = \mathcal{O}_p$. Here η_1 corresponds to the first part of the partition of the $p \times d$ matrix η as $\eta^T = (\eta_1^T, \eta_2^T)$ according to the partition of X, and η is the basis matrix of $S_{Y|X}$ .Note that the user-selected subspace $\mathcal{H}$ does not have to correspond to a subset of predictors, it can also be specified to test the contribution from certain linear combinations of the predictors.

Cook (2004) then discussed the detailed formulation of test statistics for the three types of tests in the context of Sliced Inverse Regression. It essentially involved incorporating coordinate restrictions by re-deriving it as the solution to a multivariate non-linear least squares problem.

4.2.1 SIR with least square formulation and marginal dimension hypothesis

Cook (2004) reformulated the estimation of sliced inverse regression (SIR; Li, 1991) as OLS solution. Under standardization that $Z = \Sigma^{-\frac{1}{2}}(X - E(X))$, and $S_{Y|Z} = \Sigma^{\frac{1}{2}} S_{Y|X}$, Cook (2004) stated that the basis matrix B of $S_{Y|Z}$ can be estimated by minimizing least square loss function:

$$L_d(\mathbf{B}, C_y) = \sum_{y=1}^{h} \sum_{j=1}^{n_y} ||\hat{Z}_{yj} - BC_y||^2, \tag{4.1}$$

over $B \in \mathbb{R}^{p \times d}$ and over $C_y \in \mathbb{R}^d$ with constraint that $\sum_y \hat{p_y} C_y = 0$, where $\hat{p_y} = \frac{n_y}{n}$ is the observed slice proportion. Following similar logic of ANOVA decomposition in linear

regression, one can further decompose the loss function in the format of:

$$L_d(B, C_y) = \sum_{y,j} ||\hat{Z}_{yj} - \bar{Z}_y||^2 + ||\bar{Z}_y - BC_y||^2. \tag{4.2}$$

Then the minimization of loss function $L_d(\mathbf{B}, C_y)$ is equivalent to the minimization of $||\bar{Z}_y - BC_y||^2$. Cook (2004) indicated that the estimation of C_y and B follows by an alternate algorithm with one step estimating C_y as a usual regression solution, and the other step of estimating B by eigenvalue decomposition. Combining the two steps, estimations of basis matrix $\hat{\gamma}$ will be the eigenvectors of $\hat{M}$ and estimation of the projection $\hat{\rho}_y = \hat{\gamma}^T \bar{Z}_y$. Then, if we plug in estimation $\hat{\gamma}$ and $\hat{\rho}_y$, we can rewrite the loss function as residual sum of squares:

$$\begin{aligned} \hat{L}_d &= \sum_{y=1}^{h} \sum_{j=1}^{n_h} ||\hat{Z}_{yj} - \bar{Z}_y||^2 + n \sum_{j=d+1}^{p} \hat{\lambda}_j \text{ if } d \leq p-1, \\ \hat{L}_p &= \sum_{y=1}^{h} \sum_{j=1}^{n_h} ||\hat{Z}_{yj} - \bar{Z}_y||^2 \text{ if } d \leq p-1. \end{aligned} \tag{4.3}$$

Then test for *structural dimension* d with hypotheses that $H_0 : d = m$ against $H_1 : d > m$ can be formulated using test statistics as the difference of the residual sum of squares under two hypotheses:

$$T_n(m) = \hat{L}_m - \hat{L}_p = n \sum_{j=m+1}^{p} \hat{\lambda}_j. \tag{4.4}$$

Li (1991) indicated that under normality assumption of X (both linearity condition and coverage condition will be satisfied automatically) $T_n(m)$ follows a chi-square distribution with $(p-d)(h-d-1)$ degrees of freedom asymptotically.

4.2.2 Test Statistics

Cook (2004) showed that the two types of hypothesis tests Marginal coordinate hypotheses and Conditional coordinate hypotheses share the same test statistics, which is essentially

the difference of residual sum of squares under null and alternative hypotheses. Under null hypothesis, the test statistics is:

$$L_m^{'}(B, C_y) = \sum_{yj} ||\hat{Z}_{yj} - \bar{Z}_y||^2 + \sum_{yj} ||P_{\hat{\mathcal{H}}} \bar{Z}_y||^2 + T_n^{'}(m), \tag{4.5}$$

where $T_n^{'}(m) = n \sum_{j=m+1}^{p} \hat{\lambda}_j^{'}$ ($T_n^{'}(p) = 0$) and $\hat{\lambda}_1^{'} \geq \cdots \geq \hat{\lambda}_p^{'}$ are eigenvalues of $Q_{\hat{\mathcal{H}}} \hat{M} Q_{\hat{\mathcal{H}}}$. We describe the detailed testing statistics for marginal and conditional coordinate tests in the following.

4.2.3 Marginal coordinate hypotheses testing

Marginal coordinate hypotheses testing allows one to test predictor contributions without specifying the *structural dimension* d. The test statistics is constructed as the difference of residual sum of squares under null hypothesis that $d = p$ and $P_{\mathcal{H}} S_{Y|\mathbf{Z}} = \mathcal{O}_p$ and under alternative hypothesis that $d = p$, that is

$$T_n(\mathcal{H}) = \hat{L}_p^{'} - \hat{L}_p = n \; trace(P_{\hat{\mathcal{H}}} \hat{M} P_{\hat{\mathcal{H}}}) = ||\sqrt{n} vec(\hat{\alpha}^T Z_n)||^2. \tag{4.6}$$

One can intuitively regard test statistics $n \; trace(P_{\hat{\mathcal{H}}} \hat{M} P_{\hat{\mathcal{H}}})$ as the size of the projection of $\hat{M}$ onto the subspace characterized by the null hypothesis. A large size means the projection is still high thus cannot be neglected, i.e., the contribution of the predictors in interest is significantly different from 0.

Under linearity condition and null hypothesis that $P_{\mathcal{H}} S_{Y|\mathbf{Z}} = \mathcal{O}_p$,

$$\sqrt{n} vec(\hat{\alpha}^T Z_n) \xrightarrow{D} MVN(0, \Omega_{\mathcal{H}}), \tag{4.7}$$

and consequently,

$$T_n(\mathcal{H}) \xrightarrow{\mathcal{L}} \sum_{i=1}^{hr} \omega_i \chi_i^2(1), \tag{4.8}$$

where covariance matrix $\Omega_{\mathcal{H}} = E(D_g^{-1}\epsilon\epsilon^T D_g^{-1} \otimes \alpha^T \mathbf{Z}\mathbf{Z}^T\alpha)$ and $\omega_1 \geq \omega_2 \geq \cdots \omega_{hr}$ are then eigenvalues of $\Omega_{\mathcal{H}}$. Under linearity condition, coverage condition and constant covariance condition and the null hypothesis

$$\sqrt{n} vec(\hat{\alpha}^T \mathbf{Z}_n) \xrightarrow{D} MVN(0, \Omega_{\mathcal{H}}), \tag{4.9}$$

and consequently,

$$T_n(\mathcal{H}) \xrightarrow{\mathcal{L}} \sum_{i=1}^{h-1} \delta_j \chi_j^2(r), \tag{4.10}$$

where covariance matrix $\Omega_{\mathcal{H}} = (Q_g - \mu^T\mu) \otimes I_r$ and $\delta_1 \geq \delta_2 \geq \cdots \delta_h$ are then eigenvalues of $Q_g - \mu^T\mu$. When implementing the test statistics in terms of sample data, one needs to replace the associated terms with sample estimate such as $\hat{\alpha}$, $\hat{Z}_{yj}$, $D_{\hat{g}}$, etc. The P-values for this test can be found be comparing the test statistics $T_n(\mathcal{H})$ with the percentage points of $\sum_{i=1}^{hr} \hat{\omega}_i \chi_i^2(1)$ using Satterthwaite's approximation.

4.2.4 Conditional coordinate hypotheses testing

Conditional coordinate hypotheses testing is essentially similar to Marginal coordinate hypotheses, except that one need to specify the structural dimension d. The test statistics is constructed as the difference of residual sum of squares under null hypothesis that the structural dimension is d and $P_{\mathcal{H}} S_{Y|\mathbf{Z}} = \mathcal{O}_p$ and under alternative hypothesis that the structural dimension is d, that is

$$T_n(\mathcal{H}|d) = \hat{L}_d^{'} - \hat{L}_d \tag{4.11}$$

A similar asymptotic distribution as (4.9) and (4.10) are developed in Cook (2004).

4.3 Testing predictor contribution in sufficient dimension folding

4.3.1 Definitions of three tests

Following Cook's (2004) definitions of three types of tests, we can define similar tests with regard to matrix predictors $\mathbf{X} \in \mathbb{R}^{p_l \times p_r}$. The three tests are defined as:

- Marginal dimension hypotheses:

  $H_0 : d_l = m_l \; and \; d_r = m_r$ against $H_1 :$ at least one of d_l and d_r is larger than the hypothesized dimension;

- Marginal coordinate hypotheses:

  $H_0 : P_{\mathcal{H}_l} S_{Y|\circ \mathbf{X} \circ} P_{\mathcal{H}_r} = \mathcal{O}_{p_l p_r}$ against $H_1 : P_{\mathcal{H}_l} S_{Y|\circ \mathbf{X} \circ} P_{\mathcal{H}_r} \neq \mathcal{O}_{p_l p_r}$;

- Conditional coordinate hypotheses:

  $H_0 : P_{\mathcal{H}_l} S_{Y|\circ \mathbf{X} \circ} P_{\mathcal{H}_r} = \mathcal{O}_{p_l p_r}$ against $H_1 : P_{\mathcal{H}_l} S_{Y|\circ \mathbf{X} \circ} P_{\mathcal{H}_r} \neq \mathcal{O}_{p_l p_r}$ given d_l and d_r.

4.3.2 Folded-SIR with least square formulation

As Li, Kim and Altman (2010) indicated, the estimation of *central folding subspace* can be constructed through the Kronecker envelope of a usual *central subspace* estimator $U \in \mathbb{R}^{p_l p_r \times k}$, which in turn can be estimated by minimizing:

$$E||AU - A(\beta \otimes \alpha)C_y||^2, \tag{4.12}$$

over $\alpha \in \mathbb{R}^{p_l \times d_l}$, $\beta \in \mathbb{R}^{p_r \times d_r}$ and $C_y \in \mathbb{R}^{d_l d_r \times k}$, and that d_l and d_r are structural dimensions assumed to be known in the estimation. By specifying $\Sigma_x = Cov(vec(\mathbf{X}))$, $A = \Sigma_x^{\frac{1}{2}}$ and $U = \Sigma^{-1}[E(vec(\mathbf{X}|Y) - E(vec(\mathbf{X})] \in \mathbb{R}^{p_l p_r}$, the minimizer of (4.12) as $span(\hat{\beta} \otimes \hat{\alpha})$ forms

a population estimator of Kronecker envelope of SIR, defined as folded-SIR (Li, Kim and Altman, 2010). As for sample estimates, one only needs to replace U by sliced mean of $vec(X)$, Σ_X by the sample covariance matrix of vectorized predictor.

To construct test statistics for three tests, we can follow similar logic from Cook (2004), and further decompose the residual sum of squares into two parts. First notice that for fixed α and C_y, the residual sum of squares with respect to β as unknown parameter can be written as:

$$E||\Sigma_x^{-\frac{1}{2}}[E(vec(\mathbf{X}|Y) - E(vec(\mathbf{X})] - ({C_y}^T \otimes \Sigma_x^{\frac{1}{2}})\Pi[I_{p_r d_r} \otimes vec(\alpha)]vec(\beta)||^2; \tag{4.13}$$

For fixed β and C_y, the residual sum of squares with respect to unknown parameter α can be written as:

$$E||\Sigma_x^{-\frac{1}{2}}[E(vec(\mathbf{X}|Y) - E(vec(\mathbf{X})] - ({C_y}^T \otimes \Sigma_x^{\frac{1}{2}})\Pi[vec(\beta) \otimes I_{p_l d_l}]vec(a)||^2; \tag{4.14}$$

For fixed α and β, the residual sum of squares with respect to unknown parameters $C_y,\ y = 1, ..., h$ can be written as

$$E||\Sigma_x^{-\frac{1}{2}}[E(vec(\mathbf{X}|Y) - E(vec(\mathbf{X})] - ({C_y}^T \otimes \Sigma_x^{\frac{1}{2}})\Pi[I_{p_l p_r} \otimes \Sigma_x^{\frac{1}{2}}(\beta \otimes \alpha)]C_y||^2. \tag{4.15}$$

Here Π is a transformation matrix such that $vec(\beta \otimes \alpha) = \Pi vec(\beta) \otimes vec(\alpha)$. Its details can be found in Li, Kim and Altman (2010).

4.3.3 Future work

The future work on testing predictor contribution on dimension folding will focus on the following parts:

1. Construct test statistics for three types of tests, namely, marginal dimension hypotheses,

marginal coordinate hypotheses and conditional coordinate hypotheses tests.

2. Derive asymptotic distributions for the three test statistics.

3. Generate numeric studies to compare variable selection results using test based method derived here and regularized dimension folding method derived in the previous chapter. Apply these methods to real data.

Bibliography

[1] E. Bura and R. D. Cook. Extending sliced inverse regression: The weighted chi-squared test. *Journal of the American Statistical Association* **96**, 9961003.

[2] X. Chen, C. Zou and R. D. Cook. Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics* **38**, 36963723, 2010.

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

[4] R. D. Cook. Testing predictor contribution in sufficient dimension reduction. *The Annals of Statistics* **32**, 1062-1092, 2004.

[5] R. D. Cook and S. Weisberg. Discussion of "Sliced inverse regression for dimension reduction". *Journal of the American Statistical Association* **86**, 28-33, 1991.

[6] S. Ding and R. D. Cook. Dimension folding PCA and PFC for matrix-valued predictors. *Statistica Sinica* **24**, 463-492, 2014.

[7] B. Li, M. Kim and N. Altman. On dimension folding of matrix- or array-valued statistical objects. *The Annals of Statistics* **38**, 1094-1121, 2010.

[8] B. Li, and S. Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association* **102**, 2143-2172, 2007.

[9] K.-C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**, 316-342, 1991.

[10] L. Li. Sparse sufficient dimension reduction. *Biometrika* **94**, 603-613, 2007.

[11] L. Li. and C. J. Nachtsheim. Sparse sliced inverse regression. *Technometrics* **48**, 503510, 2006.

[12] L. Li and X. Yin. Sliced inverse regression with regularizations. *Biometrics* **64**, 124-131, 2008.

[13] L. Ni, R. D. Cook and C.-L. Tsai. A note on shrinkage sliced inverse regression. *Biometrika* **92**, 242247, 2005.

[14] Q. Wang and X. Yin A nonlinear multi-dimensional variable selection method for high dimensional data: sparse MAVE. *Computational Statistics and Data Analysis* **52**, 45124520, 2008.

[15] Y. Xue and X. Yin. Sufficient dimension folding for regression mean function. *Journal of Computational and Graphical Statistics* **39**, 1028-1043, 2014.

[16] X. Yin and H. Hilafu. Sequential sufficient dimension reduction for large p, small n problems. *Journal of the Royal Statistical Society, Series B* **4**, 879-892, 2015.

[17] J. Zhou and X. He. Dimension reduction based on constrained canonical correlation and variable filtering. *Annals of Statistics* **36**, 16491668, 2008.