Sufficient Dimension Folding Theory and Methods

by

Yuan Xue

(Under the Direction of Xiangrong Yin)

Abstract

This dissertation undertakes the theory and methods of sufficient dimension folding for matrix-/array-valued objects. Traditionally, researchers reduced the dimensions of matrix-/array-valued data by collapsing the data into vectorization. Nonetheless, analysis based on the vectorized data loses the crucial structural information carried by the data. Keeping the structure is critical in many fields. Dimension folding is a cutting-edge technology for capturing the critical essence of those structured data, reducing their dimensions as much as possible, yet preserving their intrinsic structure.

We first consider sufficient dimension folding for the regression mean function when predictors are matrix- or array-valued. A new concept *central mean folding subspace* and its two local estimation methods: folded outer product of gradients estimation (*folded-OPG*) and folded minimum average variance estimation (*folded-MAVE*) are proposed. The asymptotic property for *folded-MAVE* is established. A modified BIC criterion is used to determine the dimensions of the *central mean folding subspace*. Performances of the two local estimation methods are evaluated by simulated examples and the efficacy is demonstrated in finite samples. The *folded-MAVE* method is adopted to analyze a primary biliary cirrhosis data set.

Second, we focus on sufficient dimension folding for regression on robustness for matrix- or array-valued objects. The *central functional folding subspace* and a class of estimation methods on robust estimators are introduced. Special attention is paid to the central quantile dimension folding subspace, a widely interesting case of the *central functional folding subspace*. The performances of the proposed estimation methods on estimating the central quantile folding dimension subspace are evaluated by simulated models. We also apply our method using quantile regression to the primary biliary cirrhosis data set.

Third, we introduce our future work. A class of dimension folding estimators based on an ensemble of *folded-MAVE* is introduced to characterize the central folding subspace (CFS). The ensemble estimators can exhaustively estimate the central folding subspace without imposing restrictive conditions on the predictors. A cross validation criterion is proposed to determine the dimensions of CFS. Theoretical properties and numerical performance of the proposed method will be studied in the future.

INDEX WORDS: Central Folding Subspace, Central Mean Folding Subspace, Central Quantile Folding Subspace, Folded Minimum Average Variance Estimation, Folded MAVE ensemble, Modified BIC criterion, Cross Validation.

Sufficient Dimension Folding Theory and Methods

Yuan Xue

B.S., Central University for Nationalities, China 2006

M.S., Mississippi State University 2008

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Doctor of Philosophy

Department of Statistics

Athens, Georgia

2012

Sufficient Dimension Folding Theory and Methods

by

Yuan Xue

Approved:

Major Professor:	Xiangrong Yin
Committee:	William McCormick Cheolwoo Park Jaxk Reeves Lily Wang

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2012

Sufficient Dimension Folding Theory and Methods

Yuan Xue

Acknowledgments

I am deeply indebted to my major professor, Dr. Xiangrong Yin, for his generous help, encouragement and tremendous patience for the completion of this dissertation. Dr. Yin's endless patience and kindness provide me with a pleasurable studying environment and numerous opportunities to engage with the statistical research community, such as attending workshops and conferences, referring on scientific articles and etc. Dr. Yin's assistance extended beyond the usual requirement of the dissertation and provided significant help to my professional development.

I would like to extend my deepest thanks to my committee members Dr. Yehua Li, Dr. William McCormick, Dr. Cheolwoo Park, Dr. Jaxk Reeves and Dr. Lily Wang, for serving on my committee and their valuable suggestions and timely assistance in supporting my study at UGA.

My sincere appreciation is also extended to the faculty, staff and students of the Department of Statistics, who helped me in many ways and encouraged me to pursue my achievements in academics. In addition, I extend my special thanks to the faculty committee for providing me teaching assistantship and bring me to this great department in

2009. I am thankful for tremendous assistance provided by administrative staff at the department, with special thanks to Tim Cheek, Julie Davis and Daphney Smith.

Furthermore, I would like to thank Dr. William McCormick, Dr. Xiangrong Yin and Dr. Lily Wang for writing recommendation letters to support my job applications and giving me constructive career advice. I am also grateful for the tremendous support on my teaching assistant job provided by Dr. Jeongyoun Ahn, Dr. Gauri Datta, Dr. Nicole Lazar, Dr. Abhyuday Mandal, Dr. William McCormick, Dr. John Stufken, Dr. Xiangrong Yin, Dr. Lily Wang and many others. And I am grateful for the excellent teaching given by Dr. Daniel Hall, Dr. William McCormick, Dr. Lynne Seymour, Dr. Kimberly Love-Myers and many others. Their helps provide me solid foundation in methodology and theory in statistics and encourage me to attain this degree.

Finally, I would like to thank my family members for their encouragement, love and solicitude.

Contents

Chapter 1

Sufficient Dimension Folding for Regression Mean Function

1.1 Introduction

Modern data often have complex structures. For instance, matrix-valued predictors as in pictures, and array-valued predictors as in videos. Those data sets are large and structured, with each dimension representing different information in nature. Many traditional approaches are to vectorize matrix-/array-valued data so that methods that can efficiently analyze vector-valued predictors can be directly used. However, Li, Kim and Altman (2010) used electroencephalography (EEG) data to illustrate that vectorizing matrix-valued/array-valued predictors may lose sufficient information, data structure and related interpretation. In practice, treating the predictor as a matrix not only can preserve the original matrix structure of the predictor and important aspects of interpretation, but also can reduce the number of parameters in dimension reduction estimation, which enhances the estimation accuracy. Dimension folding aims to reduce the matrix-/array-valued predictors as many as possible while preserving the structure interpretation of the underlying predictors. The sufficient dimension reduction subspace is a subset of dimension folding subspace. Except the information carried in dimension reduction subspace, the folding subspace also covers the information on data structure. Li, Kim and Altman (2010) introduced central dimension folding subspace for matrix- or array-valued objects and proposed three estimation techniques: folded-SIR, folded-SAVE and folded-DR. In this Chapter, we are interested in dimension folding for the regression mean function.

Our motivation is a primary biliary cirrhosis data set, available at http://lib.stat.cmu.edu/datasets/pbcseq. A group of predictors are repeatedly measured over time for 312 patients. If we view the group of predictors as one dimension of a matrix and the time line as another, then we would have a matrix formed predictor and each patient is a sample. In dimension folding we are interested in reducing the dimensions of such a matrix predictor along the row and column directions simultaneously. Thus, the correlations between the multivariate predictors across time are protected naturally.

To attain a more accurate estimator when the conditional mean is our interest, we introduce the *central mean folding subspace* (CMFS), which aims at the regression mean function only. We propose two local estimation methods for dimension folding: folded outer product of gradients estimation (*folded-OPG*) and folded minimum average variance estimation (*folded-MAVE*). The *folded-MAVE* does not require a strong assumption on the distribution of predictor; it can exhaustively recover the *central mean folding subspace*, and the estimation procedure can be broken down into iterations between several quadratic optimization steps, each of which has an explicit solution. A modified BIC criterion is used to determine the dimensions of the *central mean folding subspace*.

The remainder of the Chapter is organized as follows. In Section 1.2,s we introduce the *central mean folding subspace* along with its properties. In Section 1.3, we introduce several estimation methods for the *central mean folding subspace*. In particular, we study the *folded-MAVE* and its property. Simulations and application are included in Section 1.4 and Section 1.5, respectively. A brief remark on array-valued predictors is presented in Section 1.6, followed by a short discussion in Section 1.7. We delay proofs to the Appendix.

1.2 Central mean dimension folding subspace

Let $\mathcal{S}(M)$ denote a subspace spanned by the columns of a matrix M and let P_M denote the orthogonal projection onto $\mathcal{S}(M)$, that is, $P_M = M(M^\intercal M)^{-1}M^\intercal$. We consider the regression of a univariate response Y on a $p \times q$ random matrix $\mathbf{X}$ and assume the data $\{y_i, \mathbf{x}_i\}$, $i = 1, \ldots, n$ are iid observations on $(Y, \mathbf{X})$.

1.2.1 Overview of central folding subspace

Suppose there are two matrices $A \in \mathbb{R}^{p\times d}$ and $B \in \mathbb{R}^{q\times r}$, $d \leq p$ and $r \leq q$, such that

$$Y \perp\!\!\!\perp \mathbf{X} | A^{\intercal}\mathbf{X}B. \tag{1.2.1}$$

Then Y depends on $\mathbf{X}$ only through $A^{\intercal}\mathbf{X}B$. The subspaces $\mathcal{S}(A)$ and $\mathcal{S}(B)$ are called a left- and right- dimension folding subspace for $Y|\mathbf{X}$, respectively (Li, Kim and Altman, 2010). Under mild regularity conditions, the intersection of two left- or two right-dimension folding subspaces for $Y|\mathbf{X}$ is itself a left- or right- dimension folding subspace. Let $\mathcal{S}_{Y|\circ\mathbf{X}}$ or $\mathcal{S}_{Y|\mathbf{X}\circ}$ be the intersection of all left- or right– dimension folding subspaces for $Y|\mathbf{X}$. The subspace $\mathcal{S}_{Y|\mathbf{X}\circ}\otimes\mathcal{S}_{Y|\circ\mathbf{X}}$ is defined as the central dimension folding subspace (CFS) denoted by $\mathcal{S}_{Y|\circ\mathbf{X}\circ}$, where "$\otimes$" is the Kronecker product.

Let vec$(\cdot)$ be a vector operator defined by stacking the columns of a matrix "$\cdot$" into a vector, so that (1.2.1) is equivalent to $Y \perp\!\!\!\perp \text{vec}(\mathbf{X})|(B\otimes A)^{\intercal}\text{vec}(\mathbf{X})$. Let $\mathcal{S}_{Y|\text{vec}(\mathbf{X})}$ be the central subspace (CS; Cook, 1994, 1996) of Y with respect to the vector predictor vec$(\mathbf{X})$, so that $\mathcal{S}_{Y|\text{vec}(\mathbf{X})} \subseteq \mathcal{S}_{Y|\circ\mathbf{X}\circ}$. However, the opposite relation usually does not hold. Li, Kim and Altman (2010) claimed that $\mathcal{S}_{Y|\circ\mathbf{X}\circ}$ is the best way to reduce matrix-valued predictors if one hopes to preserve the matrix structure of $\mathbf{X}$.

Li, Kim and Altman (2010) proposed the concept of Kronecker envelope. That is, for a random matrix $U \in \mathbb{R}^{(d_R d_L)\times k}$, suppose there are subspaces $\mathcal{S}_{\circ U} \in \mathbb{R}^{d_R}$ and $\mathcal{S}_{U\circ} \in \mathbb{R}^{d_L}$

satisfying

$$\mathcal{S}(U) \subseteq \mathcal{S}_{U\circ} \otimes \mathcal{S}_{\circ U} \text{ almost surely.} \tag{1.2.2}$$

Then the smallest Kronecker product of the two subspaces that satisfy (1.2.2) is defined as the Kronecker envelope of U (Li, Kim and Altman, 2010). The existence of a central dimension folding subspace follows from the results on the Kronecker envelope (Li, Kim and Altman, 2010; Theorems 1 and 2) and the existence of a central subspace (Cook, 1998; Yin, Li and Cook, 2008). Based on the Kronecker envelope, they constructed a general objective function and evaluated it using three different methods: folded-SIR, folded-SAVE and folded-DR, extending from the usual SIR (Li, 1991), SAVE (Cook and Weisberg, 1991) and DR (Li and Wang, 2007).

1.2.2 Central mean dimension folding subspace

Consider a regression model for dimension folding with matrix-valued predictors $\mathbf{X}$ as

$$Y = f(A^{\intercal}\mathbf{X}B) + \epsilon, \tag{1.2.3}$$

where f is an unknown link function, A and B are $p \times d$ and $q \times r$ orthogonal matrices with $d \leq p$ and $r \leq q$, and $E(\epsilon|\mathbf{X}) = 0$ almost surely. The idea of dimension folding for a regression mean function can be briefly described as finding such matrices A and B that all the information about Y carried in $E(Y|\mathbf{X})$ is covered by $A^{\intercal}\mathbf{X}B$. We formalize the model as follows.

Definition 1.1. *Suppose there are matrices* $A \in \mathbb{R}^{p\times d}$ *and* $B \in \mathbb{R}^{q\times r}$, $d \leq p$ *and* $r \leq q$, *such that*

$$Y \perp\!\!\!\perp E(Y|\mathbf{X})|A^{\intercal}\mathbf{X}B. \tag{1.2.4}$$

Then the space $\mathcal{S}(A)$ *or* $\mathcal{S}(B)$ *is called a left- or right- mean dimension folding subspace.*

If the intersection of all left- or right- mean dimension folding subspaces is itself a left- or right- mean dimension folding subspace, we can define the *central mean folding subspace* (CMFS) as:

Definition 1.2. *Let* $\mathcal{S}_{E(Y|\circ\mathbf{X})}$ *and* $\mathcal{S}_{E(Y|X\circ)}$ *be the intersection of all left- or right- mean folding subspaces and itself is a left- or right- mean folding subspace for* $E(Y|\mathbf{X})$, *then* $\mathcal{S}_{E(Y|\circ\mathbf{X})}$ *and* $\mathcal{S}_{E(Y|X\circ)}$ *are defined as the central left- and right- mean dimension folding subspace respectively. Let* $\mathcal{S}_{E(Y|\circ\mathbf{X}\circ)}$ *denote the sufficient central mean dimension folding subspace. Then*

$$\mathcal{S}_{E(Y|\circ\mathbf{X}\circ)} = \mathcal{S}_{E(Y|\mathbf{X}\circ)} \otimes \mathcal{S}_{E(Y|\circ\mathbf{X})}. \tag{1.2.5}$$

The *central mean folding subspace* does not always exist, as the intersection of two left- or right- mean dimension folding subspaces is not always a left- or right- mean dimension folding subspace. Based on the conditions developed by Cook (1998), or Yin, Li and Cook (2008), together with Theorems 1 and 2 of Li, Kim and Altman (2010), one can show that the existence conditions of CMFS are the same as those of central mean subspace (CMS; Cook and Li, 2002). For instance, if the domain of vec($\mathbf{X}$) is open and convex, then the CMFS exists and is unique. Since $Y \perp\!\!\!\perp \mathbf{X}|A^{\intercal}\mathbf{X}B$ implies

$Y \perp\!\!\!\perp E(Y|\mathbf{X})|A^{\intercal}\mathbf{X}B$, a dimension folding subspace is necessarily a mean dimension folding subspace. Once the CMFS exists, $\mathcal{S}_{E(Y|\circ\mathbf{X}\circ)} \subseteq \mathcal{S}_{Y|\circ\mathbf{X}\circ}$ and $\mathcal{S}_{E(Y|\text{vec}(\mathbf{X}))} \subseteq \mathcal{S}_{E(Y|\circ\mathbf{X}\circ)}$, where $\mathcal{S}_{E(Y|\text{vec}(\mathbf{X}))}$ is the CMS for vec($\mathbf{X}$). From here on, we assume $A \in \mathbb{R}^{p\times d}, d \leq p$, is a basis matrix of $\mathcal{S}_{E(Y|\circ\mathbf{X})}$ or $\mathcal{S}_{Y|\circ\mathbf{X}}$ and $B \in \mathbb{R}^{q\times r}, r \leq q$, is a basis matrix of $\mathcal{S}_{E(Y|\mathbf{X}\circ)}$ or $\mathcal{S}_{Y|\mathbf{X}\circ}$ distinguished by whether the regression mean function is in focus or not.

The following proposition gives equivalent conditions for the conditional independence using Definition 1.1. We delay its proof to the Appendix.

Proposition 1.1. *The following statements are equivalent:*

1. $Y \perp\!\!\!\perp E(Y|\mathbf{X})|A^{\intercal}\mathbf{X}B$

2. *Cov*$[(Y, E(Y|\mathbf{X}))|A^{\intercal}\mathbf{X}B] = 0$

3. $E(Y|\mathbf{X})$ *is a function of* $A^{\intercal}\mathbf{X}B$

The *central mean folding subspace* is not invariant under one-to-one transformation of the response variable, because $\mathcal{S}_{E(Y|\circ\mathbf{X}\circ)}$ does not equal $\mathcal{S}_{E(T(Y)|\circ\mathbf{X}\circ)}$ in general. However, the following proposition shows that under affine linear transformation of the predictors, the central mean dimension folding subspace is invariant. Its proof is also delayed to the Appendix.

Proposition 1.2. *Let* $\mathbf{Z} = A_0^{\intercal}\mathbf{X}B_0$ *where* A_0 *and* B_0 *are full rank,* $p \times p$ *and* $q \times q$ *matrices respectively. Then* $S_{E(Y|\circ\mathbf{Z}\circ)} = (B_0^{-1} \otimes A_0^{-1})S_{E(Y|\circ\mathbf{X}\circ)}$.

1.3 Estimation of $S_{E(Y|\circ\mathbf{X}\circ)}$

In this section, we propose two main local estimation methods for the CMFS: *folded-OPG* and *folded-MAVE*. Before doing so, we introduce a technique to obtain a Kronecker product from a semi-orthogonal matrix, which will help us to obtain either an alternative solution if $\mathcal{S}_{E(Y|\text{vec}(\mathbf{X}))} = \mathcal{S}_{E(Y|\circ\mathbf{X}\circ)}$, or better initial estimates in iterative folding approaches used later.

1.3.1 Matrix decomposition to a Kronecker product

Let I_k be the k-dimensional identity matrix. Suppose $\eta \in \mathbb{R}^{pq\times dr}$ is a semi-orthogonal matrix (that is, $\eta^{\intercal}\eta = I_{dr}$) such that $\eta = B^* \otimes A^*$, where $A^* \in \mathbb{R}^{p\times d}$ and $B^* \in \mathbb{R}^{q\times r}$ with ${B^*}^{\intercal}B^* = I_r$ for identifiability. We will establish a decomposition method to obtain matrices $A^* \in \mathbb{R}^{p\times d}$ and $B^* \in \mathbb{R}^{q\times r}$. This decomposition technique bridges a way to estimate $S_{Y|\circ\mathbf{X}}$ and $S_{Y|\mathbf{X}\circ}$ or $S_{E(Y|\circ\mathbf{X})}$ and $S_{E(Y|\mathbf{X}\circ)}$ through $S_{Y|\mathrm{vec}(\mathbf{X})}$ or $S_{E(Y|\mathrm{vec}(\mathbf{X}))}$. Suppose such η is a basis matrix for $\mathcal{S}_{Y|\mathrm{vec}(\mathbf{X})}$, or $\mathcal{S}_{E(Y|\mathrm{vec}(\mathbf{X}))}$. In the first step, we estimate η of $S_{Y|\mathrm{vec}(\mathbf{X})}$ or $S_{E(Y|\mathrm{vec}(\mathbf{X}))}$; in the second step, we use the decomposition technique to estimate A^* and B^*. When $\mathcal{S}_{Y|\mathrm{vec}(\mathbf{X})} = \mathcal{S}_{Y|\circ\mathbf{X}\circ}$ and $\mathcal{S}_{E(Y|\mathrm{vec}(\mathbf{X}))} = \mathcal{S}_{E(Y|\circ\mathbf{X}\circ)}$, our algorithm can provide a solution of $\mathcal{S}_{Y|\circ\mathbf{X}\circ}$ or $\mathcal{S}_{E(Y|\circ\mathbf{X}\circ)}$.

Let $||\cdot||$ be the Frobenius norm, so that the objective functions below are equivalent to ordinary least squares. Let h_{ij} be the i-dimensional vector whose j-th element equals 1 and otherwise 0. Our procedure is as follow.

1. Generate the initial value of $A^*_{(0)} \in \mathbb{R}^{p\times d}$ from a sample of the $N(0,1)$ variables.

2. Given $A^*_{(k)}$, the estimate of A^* in the k-th iteration, the ij-th element of $B^*_{(k)}$ in k-th iteration is the minimizer, $\hat{b}$, of the matrix norm

$$||(h_{qi}^{\intercal} \otimes I_p)\eta(h_{rj} \otimes I_d) - bA^*_{(k)}||.$$

   Then normalize $\hat{b}$ so that ${B^*}^{\intercal}_{(k)}{B^*}_{(k)} = I_r$.

3. Given $B^*_{(k)}$, the minimizer $\hat{a}$ of the objective function,

$$||(h_{pi}^{\mathsf{T}} \otimes I_q)(\mathbb{K}_{p,q}\eta\mathbb{K}_{r,d})(h_{dj} \otimes I_r) - aB^*_{(k)}||,$$

   is the ij-th element of $A^*_{(k+1)}$ in the $(k+1)$-th iteration, where $\mathbb{K}$ is a commutation matrix. Its explicit form and the properties can be found in Magnus and Neudecker (1999). Here we use the property that: if $A^* \in \mathbb{R}^{p\times d}$ and $B^* \in \mathbb{R}^{q\times r}$, then $A^*\otimes B^* = \mathbb{K}_{p,q}(B^* \otimes A^*)\mathbb{K}_{r,d}$.

4. Check the convergence. Let $\tau_{(k)} = B^*_{(k)} \otimes A^*_{(k)}$, $\tau_{(k-1)} = B^*_{(k-1)} \otimes A^*_{(k-1)}$ be the estimates in the k-th and the $(k-1)$-th iteration. If the discrepancy, $||\tau_{(k)}\tau_{(k)}^{\mathsf{T}} - \tau_{(k-1)}\tau_{(k-1)}^{\mathsf{T}}||$, is smaller than some pre-specified tolerance value, such as 10^{-6}, then stop the iteration and set $\hat{A}^* = A^*_{(k)}$, $\hat{B}^* = B^*_{(k)}$; Otherwise, set $k = k + 1$ and proceed Step 2.

1.3.2 Local estimation methods of CMFS

At this stage, we assume that d and r are known. The key idea underlying the subsequent development of *folded-OPG* and *folded-MAVE* is as follows. Let $u = A^{\mathsf{T}}\mathbf{X}B$. Then, based on elementary calculations (Schott 1997), we have:

$$\begin{aligned}\frac{\partial E(Y|\mathbf{X}=\mathbf{x})}{\partial \mathbf{X}} &= \frac{\partial \text{vec}(u)^{\mathsf{T}}}{\partial \text{vec}(\mathbf{X})} \cdot \frac{\partial E(Y|A^{\mathsf{T}}\mathbf{X}B = u)}{\partial \text{vec}(u)} \\ &= (B \otimes A) \cdot \frac{\partial E(Y|A^{\mathsf{T}}\mathbf{X}B = u)}{\partial \text{vec}(u)}.\end{aligned}$$

From the above, we know that $\frac{\partial E(Y|\mathbf{X}=\mathbf{x})}{\partial \mathbf{X}} \in \mathcal{S}(B \otimes A) = \mathcal{S}_{E(Y|\circ\mathbf{X}\circ)}$. Thus, $\mathcal{S}_{E(Y|\circ\mathbf{X}\circ)}$ can be recovered by estimating the gradient of $E(Y|\mathbf{X}=\mathbf{x})$.

folded-OPG

The local polynomial smoothing (Fan and Gijbels, 1996) can be used to estimate the gradients. Here, for each $j = 1, \dots n$, we consider the local linear fitting by minimizing the objective function,

$$\sum_{i=1}^{n} [y_i - c_j - a_j^{\intercal}(\mathbf{x}_i - \mathbf{x}_j) b_j \}]^2 w_{ij}, \tag{1.3.1}$$

over $(c_j, a_j, b_j) \in \mathbb{R}^1 \times \mathbb{R}^p \times \mathbb{R}^q$, subject to $b_j^{\intercal} b_j = 1$ where $w_{ij} \geq 0$ is the kernel weight centered at $\mathbf{x}_i - \mathbf{x}_j$ with $\sum_{i=1}^{n} w_{ij} = 1$. We use the usual kernel weight:

$$w_{ij}(h) = K_h(\text{vec}(\mathbf{x}_i - \mathbf{x}_j)) / \sum_{j=1}^{n} K_h(\text{vec}(\mathbf{x}_i - \mathbf{x}_j)).$$

For any $v \in R^p$, $K_h(v) = h^{-p} K(||v||/h)$, where $K(\cdot)$ is the chosen kernel function. In this chapter, we use the Gaussian kernel.

The *folded-OPG* based on minimizing (1.3.1) can be viewed as "weighted least squares" approach. For each $j = 1, \dots n$, we can iteratively estimate a_j, b_j and c_j. We suggest the following *folded-OPG* algorithm:

1. Generate the initial values of $a_j \in \mathbb{R}^p$ for $j = 1, \cdots, n$ from a sample of $N(0, 1)$ variables.

2. For fixed $a_j \in \mathbb{R}^p$, minimize (1.3.1) over c_j and b_j for $j = 1, \ldots, n$. The solution is

$$\begin{pmatrix} \hat{c}_j \\ \\ \hat{b}_j \end{pmatrix} = \left[\sum_{i=1}^{n} w_{ij}(h) \Delta_{ij}(a_j) \Delta_{ij}^{\mathsf{T}}(a_j) \right]^{\text{-1}} \left[\sum_{i=1}^{n} w_{ij}(h) \Delta_{ij}(a_j) y_i \right],$$

   where $\Delta_{ij}(a_j) = (1, (\text{vec}(\mathbf{x}_i - \mathbf{x}_j))^{\mathsf{T}}(I_q \otimes a_j))^{\mathsf{T}}$. Normalize $\hat{b}_j$: $\hat{b}_j^{\mathsf{T}} \hat{b}_j = 1$.

3. For fixed c_j and b_j, let:

$$\hat{a}_j = \left[\sum_{i=1}^{n} w_{ij}(h) \Delta_{ij}(b_j) \Delta_{ij}^{\mathsf{T}}(b_j) \right]^{\text{-1}} \left[\sum_{i=1}^{n} w_{ij}(h) \Delta_{ij}(b_j)(y_i - c_j) \right],$$

   where $\Delta_{ij}(b_j) = (b_j \otimes I_p)^{\mathsf{T}} \text{vec}(\mathbf{x}_i - \mathbf{x}_j)$.

4. Check convergence. Let $A_{(k)}$, $B_{(k)}$ be the first d or r eigenvectors according to the d or r largest eigenvalues of $\sum_{j=1}^{n} \hat{a}_j \hat{a}_j^{\mathsf{T}}$ or $\sum_{j=1}^{n} \hat{b}_j \hat{b}_j^{\mathsf{T}}$ obtained in the k-th iteration respectively. Let $\tau_{(k)} = B_{(k)} \otimes A_{(k)}$, $\tau_{(k-1)} = B_{(k-1)} \otimes A_{(k-1)}$. If $||\tau_{(k)}\tau_{(k)}^{\mathsf{T}} - \tau_{(k-1)}\tau_{(k-1)}^{\mathsf{T}}||$ is smaller than some pre-specified tolerance value, such as 10^{-6}, stop the iteration and set $A = A_{(k)}$, $B = B_{(k)}$; Otherwise, set $k = k + 1$ and proceed Step 2.

folded-MAVE **method**

The matrices $A \in \mathbb{R}^{p \times d}$ and $B \in \mathbb{R}^{q \times r}$ satisfying (1.2.3) are the minimizers of

$$E[Y - E(Y|A^{\mathsf{T}}\mathbf{X}B)]^2, \tag{1.3.2}$$

over A and B, subject to $A^\intercal A = I_d$ and $B^\intercal B = I_r$. The conditional variance given $A^\intercal \mathbf{X} B$ is

$$\sigma^2_{A,B}(A^\intercal \mathbf{X} B) = E[\{Y - E(Y|A^\intercal \mathbf{X} B)\}^2 | A^\intercal \mathbf{X} B]. \tag{1.3.3}$$

Thus,

$$\min_{A,\ B} E[Y - E(Y|A^\intercal \mathbf{X} B)]^2 = \min_{A,\ B} E\{\sigma^2_{A,B}(A^\intercal \mathbf{X} B)\}. \tag{1.3.4}$$

At the sample level, for any given $\mathbf{x}_0$, $\sigma^2_{A,B}(A^\intercal \mathbf{x}_0 B)$ can be approximated using local linear smoothing as

$$\begin{aligned} \sigma^2_{A,B}(A^\intercal \mathbf{x}_0 B) &\approx \sum_{i=1}^{n} \{y_i - E(y_i|A^\intercal \mathbf{x}_i B)\}^2 w_{i0} \\ &\approx \sum_{i=1}^{n} [y_i - c_0 - a_0^\intercal A^\intercal (\mathbf{x}_i - \mathbf{x}_0) B b_0]^2 w_{i0}, \end{aligned}$$

where $c_0 + a_0^\intercal A^\intercal (\mathbf{x}_i - \mathbf{x}_0) B b_0$ is the local linear expansion of $E(y_i|A^\intercal \mathbf{x}_i B)$ at the point $\mathbf{x}_0$. Finding A and B is equivalent to solving the quadratic minimization

$$\sum_{j=1}^{n} \sum_{i=1}^{n} \rho_j [y_i - c_j - a_j^\intercal A^\intercal (\mathbf{x}_i - \mathbf{x}_j) B b_j]^2 w_{ij}, \tag{1.3.5}$$

over $(c_j, a_j, b_j, A, B) \in \mathbb{R}^1 \times \mathbb{R}^d \times \mathbb{R}^r \times \mathbb{R}^p \times \mathbb{R}^q$, subject to $b_j^\intercal b_j = 1$, $A^\intercal A = I_d$, $B^\intercal B = I_r$, for $i = 1, \ldots, n$ and $j = 1, \ldots, n$. We adopt coefficients $\{\rho_j : j = 1, \ldots, n\}$ to exclude unreliable samples with too few observations around $\rho(v) > 0$ if $v > v_0$ and $\rho(v) = 0$ if

$v \leq v_0$ for some small $v_0 > 0$. We take $\rho_j = \rho(n^{-1}\sum_{i=1}^{n} K_{h_{(d,r)}}(\text{vec}(\mathbf{x}_i - \mathbf{x}_j))$. Formula (1.3.5) is equivalent to

$$\sum_{j=1}^{n}\sum_{i=1}^{n} \rho_j [y_i - c_j - (b_j \otimes a_j)^{\intercal}(B \otimes A)^{\intercal}\text{vec}(\mathbf{x}_i - \mathbf{x}_j)]^2 w_{ij}. \tag{1.3.6}$$

Minimizing the objective function (1.3.5) can be broken down into the following six-step iterative algorithm. And each minimization step is a quadratic optimization problem having an explicit solution. The $\hat{A}$ and $\hat{B}$ estimated by *folded-OPG* method can be used as the initial value for *folded-MAVE*. We set the kernel weights as

$$w_{ij} = K_{h_{(d,r)}}(\text{vec}(\hat{A}^{\intercal}(\mathbf{x}_i - \mathbf{x}_j)\hat{B}))/\sum_{j=1}^{n} K_{h_{(d,r)}}(\text{vec}(\hat{A}^{\intercal}(\mathbf{x}_i - \mathbf{x}_j)\hat{B})). \tag{1.3.7}$$

The bandwidth $h_{(d,r)}$ is set to be proportional to $n^{-1/(dr+4)}$ (Silverman, 1986). Once an estimate of A or B is known, we use the existing estimate to update the bandwidth by formula (1.3.7) to reduce the dimension of the kernel function, which helps to carry out the smoothing over a lower dimension. That is, our *folded-MAVE* method is similar to the refined MAVE (rMAVE; Xia, et al. 2002). More specifically, we define the *folded-MAVE* algorithm as follows.

1. Generate the initial values of $a_j \in \mathbb{R}^d$ for $j = 1, \cdots, n$ from a sample of $N(0, 1)$ variables. Set the estimates $\hat{A}$ and $\hat{B}$ from *folded-OPG* procedure as the initial values of A and B.

2. For fixed $a_j \in \mathbb{R}^d$, $A \in \mathbb{R}^{p\times d}$, $B \in \mathbb{R}^{q\times r}$, minimize (1.3.5) over c_j, b_j for $j = 1, \ldots, n$ subject to $b_j^{\intercal} b_j = 1$. The solution is

$$\begin{pmatrix} \hat{c}_j \\ \\ \hat{b}_j \end{pmatrix} = \left[\sum_{i=1}^{n} \rho_j w_{ij}(h) \Delta_{ij}(a_j, A, B) \Delta_{ij}^{\intercal}(a_j, A, B)\right]^{\text{-1}} \left[\sum_{i=1}^{n} \rho_j w_{ij}(h) \Delta_{ij}(a_j, A, B) y_i\right],$$

where $\Delta_{ij}(a_j, A, B) = (1, (\mathrm{vec}(\mathbf{x}_i - \mathbf{x}_j))^{\intercal}(B \otimes A)(I_r \otimes a_j))^{\intercal}$.

3. For fixed $c_j \in \mathbb{R}^1$, $b_j \in \mathbb{R}^r$, $A \in \mathbb{R}^{p\times d}$, $B \in \mathbb{R}^{q\times r}$, minimize (1.3.5) over a_j for $j = 1, \ldots, n$. Then,

$$\hat{a}_j = \left[\sum_{i=1}^{n} \rho_j w_{ij}(h) \Delta_{ij}(b_j, A, B) \Delta_{ij}^{\intercal}(b_j, A, B)\right]^{\text{-1}} \left[\sum_{i=1}^{n} \rho_j w_{ij}(h) \Delta_{ij}(b_j, A, B)(y_i - c_j)\right],$$

where $\Delta_{ij}(b_j, A, B) = ((\mathrm{vec}(\mathbf{x}_i - \mathbf{x}_j))^{\intercal}(B \otimes A)(b_j \otimes I_d))^{\intercal}$.

4. For fixed c_j, a_j, b_j and A, the $\hat{B}$ that minimizes (1.3.5) is

$$\mathrm{vec}(\hat{B}) = \left[\sum_{i,j=1}^{n} \rho_j w_{ij}(h) \Delta_{ij}(a_j, b_j, A) \Delta_{ij}^{\intercal}(a_j, b_j, A)\right]^{\text{-1}} \left[\sum_{i,j=1}^{n} \rho_j w_{ij}(h) \Delta_{ij}(a_j, b_j, A)(y_i - c_j)\right],$$

where $\Delta_{ij}(a_j, b_j, A) = [I_r \otimes ((\mathbf{x}_i - \mathbf{x}_j)^{\intercal} A)](b_j \otimes a_j)$. Normalize $\hat{B}$ so that $\hat{B}^{\intercal}\hat{B} = I_r$.

5. For fixed c_j, a_j, b_j and B, the $\hat{A}$ that minimizes (1.3.5) is

$$\mathrm{vec}(\hat{A}^{\intercal}) = \left[\sum_{i,j=1}^{n} \rho_j w_{ij}(h) \Delta_{ij}(a_j, b_j, B) \Delta_{ij}^{\intercal}(a_j, b_j, B)\right]^{\text{-1}} \left[\sum_{i,j=1}^{n} \rho_j w_{ij}(h) \Delta_{ij}(a_j, b_j, B)(y_i - c_j)\right],$$

where $\Delta_{ij}(a_j, b_j, B) = [((\mathbf{x}_i - \mathbf{x}_j)B) \otimes I_d](b_j \otimes a_j)$. Normalize $\hat{A}$ so that $\hat{A}^{\intercal}\hat{A} = I_d$.

6. Check convergence. Let $A_{(k)}$, $B_{(k)}$ be the estimator of A and B obtained in the k-th iteration, respectively. Let $\tau_{(k)} = B_{(k)} \otimes A_{(k)}$, $\tau_{(k-1)} = B_{(k-1)} \otimes A_{(k-1)}$. If $||\tau_{(k)}\tau_{(k)}^{\intercal} - \tau_{(k-1)}\tau_{(k-1)}^{\intercal}||$ is smaller than some pre-specified tolerance value, such as 10^{-6}, stop the iteration and set $A = A_{(k)}$, $B = B_{(k)}$; Otherwise, set $k = k + 1$ and proceed Step 2.

1.3.3 Sampling property

In this section, we will investigate the sampling property of our estimator, similar to those made in Xia (2007), and Wang and Xia (2008). In particular, we follow the proofs in Yin and Li (2011).

Let $g(\cdot)$ be a generic density. Let A and B be the corresponding semi-orthogonal bases for the cental left- and right- mean folding subspace. Suppose that A_* and B_* are generic p-row and q-row matrices, respectively. We assume the following conditions:

(C1) [Marginal distribution of vec($\mathbf{X}$)] The covariate vec($\mathbf{X}$) is bounded; its density function $g(\text{vec}(\mathbf{X}))$ has bounded second order derivatives; functions

$$\mu_{B_*\otimes A_*}(u) = E(\text{vec}(\mathbf{X})|(B_* \otimes A_*)^{\intercal}\text{vec}(\mathbf{X}) = u),$$

$$\omega_{B_*\otimes A_*}(u) = E(\text{vec}(\mathbf{X})\text{vec}(\mathbf{X})^{\intercal}|(B_* \otimes A_*)^{\intercal}\text{vec}(\mathbf{X}) = u)$$

have bounded derivatives with respect to u, and $B_* \otimes A_*$ for $B_* \otimes A_* \in \{B_* \otimes A_* : ||(B_* \otimes A_*)(B_* \otimes A_*)^{\intercal} - (B \otimes A)(B \otimes A)^{\intercal}|| \leq c^* \}$ for some $c^* > 0$.

(C2) [Conditional distribution function of Y given $(B_* \otimes A_*)^\intercal \text{vec}(\mathbf{X})$] The conditional density function $g(Y|u)$ has bounded fourth order derivatives with respect to $\text{vec}(\mathbf{X})$, u and $B_* \otimes A_*$ as $B_* \otimes A_*$ is in a small neighbor of $B \otimes A$.

(C3) [Central mean folding subspace] For any semi-orthogonal $p \times d_*$ matrix A_* and semi-orthogonal $q \times r_*$ matrix B_* and constant $c^* > 0$, if $B_* \otimes A_* : ||(B_* \otimes A_*)(B_* \otimes A_*)^\intercal - (B \otimes A)(B \otimes A)^\intercal|| \geq c^*$, then

$$\inf_{B_* \otimes A_*} E[E(Y|A_*^\intercal \mathbf{X} B_*) - E(Y|A^\intercal \mathbf{X} B)]^2 > 0.$$

(C4) [Kernel function] Function $K_0(\cdot)$ is a symmetric univariate density function with bounded second order derivative and compact supports.

(C5) [Bandwidths] For working dimension d_* and r_*, the bandwidths $\{h_k : k = 0, 1, \ldots\}$ satisfy $h_0 \propto n^{-1/(pq+4)}$, $h_t = \max\{\varsigma h_{t-1}, \hbar\}$ with $1/2 < \varsigma < 1$, and $\hbar \propto n^{-1/(d_* r_* + 4)}$.

The general idea of the proof is the following: Note that (1.3.5) is equivalent to

$$\sum_{j=1}^{n} \sum_{i=1}^{n} \rho_j [y_i - c_j - (b_j \otimes a_j)^\intercal (B \otimes A)^\intercal \text{vec}(\mathbf{x}_i - \mathbf{x}_j)]^2 w_{ij}. \tag{1.3.8}$$

Defining $\mathbf{B} = B \otimes A$, $d_j = b_j \otimes a_j$, $X = \text{vec}(\mathbf{X})$, $X_i = \text{vec}(\mathbf{x}_i)$ and $X_{ij} = \text{vec}(\mathbf{x}_i - \mathbf{x}_j)$, (1.3.8) transforms to

$$\sum_{j=1}^{n} \sum_{i=1}^{n} \rho_j [y_i - c_j - d_j^\intercal \mathbf{B}^\intercal X_{ij}]^2 w_{ij}. \tag{1.3.9}$$

Note that formulation (1.3.9) is exactly what Yin and Li (2011) had for the objective function, if we set, $f(y_i) = y_i$ and $m = 1$. In such a case, when the central mean subspace of vec($\mathbf{X}$) is the same as that of the folding subspace, estimate $\hat{\mathbf{B}}$ is the weighted least squares solution as in Yin and Li (2011). Hence, the asymptotic results follow immediately. However, when the central mean subspace of vec($\mathbf{X}$) is not the same as that of the folding subspace, estimating $\mathbf{B}$ directly by the weighted least squares method will return basis matrix for other subspaces rather than the *central mean folding subspace*, generally a subspace of folding space if dimensions are assumed correctly to the ones for the central mean subspace. To prevent the latter, we notice that the Kronecker product structure of $\mathbf{B}$ can be recovered by using the iteration methods we proposed previously. Thus the final estimate, $\hat{\mathbf{B}} = \hat{B} \otimes \hat{A}$, is then the weighted least squares solution of (1.3.9) with such a structure. Therefore, we can prove the asymptotic result exactly following proof in Yin and Li (2011). Suppose $\hat{A}_*$ and $\hat{B}_*$ are the corresponding estimates of A_* and B_*. Then the following theorem gives the convergence rate of *folded-MAVE* when $d_* = d$ and $r_* = r$, whose proof will be provided upon request.

Theorem 1.1. *Suppose conditions (C1)–(C5) hold, $d_* = d$, $r_* = r$ and the final bandwidth is $\hbar$, then the* folded-MAVE *estimator $\hat{B} \otimes \hat{A}$ is consistent with,*

$$\begin{aligned} &||(\hat{B} \otimes \hat{A})(\hat{B} \otimes \hat{A})^{\intercal} - (B \otimes A)(B \otimes A)^{\intercal}|| \\ &= O_P\{\hbar^4 + \log n/(n\hbar^{dr}) + n^{-1/2}\}. \end{aligned} \tag{1.3.10}$$

1.3.4 Estimation of the CMFS dimension

Previously, we assumed that d and r are known, but typically we don't know either d or r. Thus, an estimation method for d and r is needed. In this section, we propose a modified BIC criterion to estimate d and r for *folded-MAVE.*

For vector predictors, Xia et al. (2002) proposed a Cross Validation (CV) criterion. Zhu, Miao and Peng (2006) proposed several BIC criteria to determine the dimension of CS. Wang and Yin (2008) suggested a modified BIC criterion for Sparse MAVE. Following Wang and Yin (2008), we propose our modified BIC criterion to estimate d and r as follows:

$$BIC_{(d_*,r_*)} = \log(\frac{RSS_{(d_*,r_*)}}{n}) + \frac{C_n \times d_* r_*}{n h_{(d_*,r_*)}^{d_* r_*}}, \tag{1.3.11}$$

where $1 \leq d_* \leq p, 1 \leq r_* \leq q$, $C_n > 0$, and $RSS_{(d_*,r_*)}$ is the residual sum of squares from the local linear smoothing using semi-orthogonal $p \times d_*$ matrix $\hat{A}_{d_*}$ and semi-orthogonal $q \times r_*$ matrix $\hat{B}_{r_*}$. That is, let $\mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j$, then

$$RSS_{(d_*,r_*)} = \sum_{j=1}^{n} \sum_{i=1}^{n} (y_i - \hat{c}_j - \hat{a}_j^\intercal \hat{A}_{d_*}^\intercal \mathbf{x}_{ij} \hat{B}_{r_*} \hat{b}_j)^2 K_{h_{(d_*,r_*)}}(\text{vec}(\hat{A}_{d_*}^\intercal \mathbf{x}_{ij} \hat{B}_{r_*})). \tag{1.3.12}$$

The estimated dimensions are then

$$(\hat{d}, \hat{r}) = \min\{(d_*, r_*) : (d_*, r_*) = \arg \min_{\substack{1 \leq d_* \leq p, \\ 1 \leq r_* \leq q}} \{BIC_{(d_*,r_*)}\}\}.$$

We have the following result, and its proof is in the Appendix.

Proposition 1.3. *Under the assumptions in the Appendix, $(\hat{d}, \hat{r}) \to (d, r)$, in probability.*

We use the bandwidth, $h_{(d*,r*)} = n^{-1/(4+d*r*)}$, similar to that of Wang and Yin (2008). For the BIC criterion, we used $C_n = W_n = (.5\log(n) + .1n^{1/3})/2$ (Zhu, Miao and Peng, 2006). In such a case if $dr < 8$, Proposition 1.3 holds, so that $(\hat{d}, \hat{r}) \to (d, r)$, in probability. Indeed, in our simulated models, $dr = 4$ or $dr = 2$, so results showed what we expected.

1.3.5 The standardization of X

When the scales of the elements of **X** are different, estimates using the original scales may not be accurate. To eliminate such an effect, we may need to standardize **X** before estimating its central folding subspace or *central mean folding subspace.* However, since **X** is a matrix, the standardization encounters some difficulties. It involves the estimation of the Kronecker product structure of the covariance matrix, $\Sigma = \text{Cov}(\text{vec}(\mathbf{X}))$. That is, Σ is said to be $(q, p)-$separable (Lu and Zimmerman, 2005) if $\Sigma = \Sigma_r \otimes \Sigma_l$, where Σ_r is $q \times q$ and positive definite, and Σ_l is $p \times p$ and positive definite. If $\text{vec}(\mathbf{X})$ follows a multivariate normal distribution with mean $\text{vec}(\mathbf{U})$, where **U** is the mean of **X**, and the covariance matrix $\Sigma = \Sigma_r \otimes \Sigma_l$, we use the notation

$$\mathbf{X} \sim N_{p,q}(\mathbf{U}, \Sigma_l, \Sigma_r)$$

to stand for the distribution of the random matrix $\mathbf{X}$ (Srivastava, von Rosen and von Rosen, 2008).

Under the matrix normal distribution model $N_{p,q}(\mathbf{U}, \Sigma_l, \Sigma_r)$, Dutilleul (1999), Lu and Zimmerman (2005) and Roy and Khattree (2005) proposed the maximum likelihood estimation (MLE) for $\mathbf{U}$, Σ_l and Σ_r using an iterative "flip-flop" scheme (Mardia and Goodall, 1993; Dutilleul, 1999; Brown, Kenward and Bassett, 2001), alternating between the estimates of Σ_l and Σ_r. However, since $\Sigma = (c\Sigma_r) \otimes (c^{-1}\Sigma_l)$ for any non-zero constant c, the estimates of Σ_l and Σ_r are not unique if there is no restriction on Σ_l or Σ_r except for positive definiteness (Lee, Dutilleul and Roy, 2010).

Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are n iid sample from $N_{p,q}(\mathbf{U}, \Sigma_l, \Sigma_r)$ with $n > \max(p, q)$. Let $x_{(\cdot i)j}$ be the i-th column vector and $x_{(k\cdot)j}$ be the k-th row for the j-th sample $\mathbf{x}_j$. Then we can rewrite $\mathbf{x}_j$ into $\mathbf{x}_j = (x_{(\cdot 1)j}, \ldots, x_{(\cdot q)j})$, where $x_{(\cdot 1)j} \ldots, x_{(\cdot q)j}$ are $p \times 1$ column vectors, or $\mathbf{x}_j = (x^{\top}_{(1\cdot)j}, \ldots, x^{\top}_{(p\cdot)j})^{\top}$, where $x_{(1\cdot)j}, \ldots, x_{(p\cdot)j}$ are $1 \times q$ row vectors, for $j = 1, \ldots, n$. Let $\bar{\mathbf{X}} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{x}_j$ then $\bar{\mathbf{X}} = (\bar{x}_{(\cdot 1)}, \ldots, \bar{x}_{(\cdot q)})$ or $\bar{\mathbf{X}} = (\bar{x}^{\top}_{(1\cdot)}, \ldots, \bar{x}^{\top}_{(p\cdot)})^{\top}$. By constraining the on-diagonal elements in Σ_r to be 1, Srivastava, von Rosen and von Rosen (2008) constructed the unique MLE of $\mathbf{U}$, Σ_l, and Σ_r as: $\hat{\mathbf{U}} = \bar{\mathbf{X}}, \hat{\Sigma}_l = S = \frac{1}{nq}\sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{X}})(\mathbf{x}_j - \bar{\mathbf{X}})^{\top}$. For the (i, k)-th entry, $\sigma_{i,k}$, in Σ_r, where $i \neq k$:

$$\hat{\sigma}_{i,k} = \frac{1}{nq}\sum_{j=1}^{n} tr(S^{-1}(x_{(\cdot i)j} - \bar{x}_{(\cdot i)})(x_{(\cdot i)j} - \bar{x}_{(\cdot i)})^{\top}),$$

where $tr(\cdot)$ stands for the trace of a matrix. Note that $\frac{n}{n-1}S$ is an unbiased and consistent estimator of Σ_l, and $\hat{\sigma}_{i,k}$ is a consistent estimator of $\sigma_{i,k}$.

Pfeiffer, Forzani and Bura (2011) provided different estimators for Σ_l and Σ_r. Assuming there are no missing value, in $\mathbf{X}$, then their estimators are:

$$\hat{\Sigma}_{lt} = \frac{1}{n}\sum_{j=1}^{n}(x_{(\cdot t)j} - \bar{x}_{(\cdot t)})(x_{(\cdot t)j} - \bar{x}_{(\cdot t)})^{\intercal}, \hat{\Sigma}_l = \frac{1}{q}\sum_{t=1}^{q}\hat{\Sigma}_{lt}.$$

and

$$\hat{\Sigma}_{rk} = \frac{1}{n}\sum_{j=1}^{n}(x_{(k\cdot)j} - \bar{x}_{(k\cdot)})^{\intercal}(x_{(k\cdot)j} - \bar{x}_{(k\cdot)}), \hat{\Sigma}_r = \frac{1}{p}\sum_{k=1}^{p}\hat{\Sigma}_{lk}.$$

Likelihood ratio tests for the separability of Σ can be found in Lu and Zimmerman (2005), Roy and Khattree (2005) and Srivastava, von Rosen and von Rosen (2008, 2009). Our simulations show that the above two methods for estimating the Kronekcer product of the covariance matrix are equivalent, and we adopt Pfeiffer's method. Without confusion, let $\mathbf{Z}$ be the standardization of $\mathbf{X}$ and $\mathbf{z}_i$ be the i-th sample of $\mathbf{Z}$. We normalize the i-th sample of vec($\mathbf{X}$) as $\text{vec}(\mathbf{z}_i) = \hat{\Sigma}^{-1/2}(\text{vec}(\mathbf{x}_i) - \text{vec}(\bar{\mathbf{x}})) = (\hat{\Sigma}_r \otimes \hat{\Sigma}_l)^{-1/2}(\text{vec}(\mathbf{x}_i) - \text{vec}(\bar{\mathbf{x}}))$. Thus $\mathbf{z}_i = \hat{\Sigma}_l^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}})\hat{\Sigma}_r^{-1/2}$. Suppose the bases for the corresponding central left- and right- mean dimension folding subspaces of $\mathbf{Z}$ are $A_{\mathbf{Z}}$ and $B_{\mathbf{Z}}$, then we transform the bases back to the $\mathbf{X}$ scale as $\hat{\Sigma}_l^{-1/2}A_{\mathbf{Z}}$ and $\hat{\Sigma}_r^{-1/2}B_{\mathbf{Z}}$ for the central left- and central right mean dimension folding subspaces, respectively.

1.4 Numerical Study

To evaluate the accuracy of estimates of our methods, we use the distance proposed by Li, Zha and Chiaromonte (2005),

$$\Delta(B \otimes A, \hat{B} \otimes \hat{A}) = ||P_{B\otimes A} - P_{\hat{B}\otimes\hat{A}}||.$$

We use the Frobenius norm $||\cdot||$ and denote it by Δ_f. The smaller the Δ_f is, the better the estimate is. In addition, we also use the benchmark distance provided by Li, Wen and Zhu (2008) to understand the accuracy of the estimates. Let α and β be $s \times t$ random matrices whose entries are independent random variables each with a standard normal distribution, satisfying $\alpha \perp\!\!\!\perp \beta$. Let P_α and P_β be the projections onto the column space of α and β separately. The benchmark distance, $E(||P_\alpha - P_\beta||)$, is determined only by the values of s and t. We estimate the benchmark distance by running 10,000 simulations and denote it by Δ_B.

We use $p = q = 5$ in the simulated examples. The error ϵ is independent of $\mathbf{X}$ and follows a standard normal distribution. In order to compare with the fold-SIR, folded-SAVE and folded-DR, all the models we construct satisfy $\mathcal{S}_{Y|\circ\mathbf{X}\circ} = \mathcal{S}_{E(Y|\circ\mathbf{X}\circ)}$. In the first four examples, the predictor $\text{vec}(\mathbf{X}) \sim N_{pq}(0, I_{pq})$. In Example 1.5, we consider a correlated predictor model, where $\text{vec}(\mathbf{X}) \sim N_{pq}(0, \Sigma_{\text{vec}(\mathbf{X})})$ and $\Sigma_{\text{vec}(\mathbf{X})}$ is a $pq \times pq$ positive definite matrix with (j_1, j_2)-th entry $0.5^{|j_1 - j_2|}$. We run 100 replicates for each model and compute the average of Δ_f and its standard error, and report the accuracy as: mean ± standard deviation. Let $e_1 = (1, 0, 0, 0, 0)^\intercal$ and $e_2 = (0, 1, 0, 0, 0)^\intercal$. Further, we write X_{ij} as the random predictor in the ij-th position of $\mathbf{X}$.

Example 1.1. *:* $Y = X_{11} \times (X_{12} + X_{21} + 1) + 0.2 \times \epsilon$.

In this example, $\mathcal{S}_{Y|\circ\mathbf{X}\circ} = \mathcal{S}_{E(Y|\circ\mathbf{X}\circ)} = \mathcal{S}(e_1 \otimes e_1,\ e_1 \otimes e_2,\ e_2 \otimes e_1, e_2 \otimes e_2)$ with $d = r = 2$; while $\mathcal{S}_{Y|\text{vec}(\mathbf{X})} = \mathcal{S}_{E(Y|\text{vec}(\mathbf{X}))} = \mathcal{S}(e_1 \otimes e_1,\ e_1 \otimes e_2 + e_2 \otimes e_1)$. Obviously, $\mathcal{S}_{Y|\text{vec}(\mathbf{X})} = \mathcal{S}_{E(Y|\text{vec}(\mathbf{X}))} \subset \mathcal{S}_{Y|\circ\mathbf{X}\circ} = \mathcal{S}_{E(Y|\circ\mathbf{X}\circ)}$. For this example, the two-step decomposition method in Section 1.3.1 can not exhaustively recover the CMFS. To demonstrate the performance of the two-step method in estimating CMFS of $\mathbf{X}$, we first apply the OPG method to vec($\mathbf{X}$) and then decompose the estimated base into a Kronecker product of two matrices as described in Section 1.3.1. Table 1 shows that the two-step decomposition method ("two-step") fails in recovering the *central mean folding subspace*. The rMAVE for vec($\mathbf{X}$) also fails to recover the CMFS as shown in Table 1.1. The *folded-OPG* and *folded-MAVE* dominate the folded-SIR, folded-SAVE and folded-DR, while *folded-MAVE* performs the best among all these methods. With sample size increasing, the accuracy of estimating CMFS is improved. The benchmark distance is $\Delta_B = 2.586$.

Table 1.1: Example 1.1: Accuracy of estimates

n	Δ_f						
	two-step	rMAVE	*folded-OPG*	*folded-MAVE*	folded-SIR	folded-SAVE	folded-DR
200	1.8963	1.9269	0.9511	0.3980	1.6039	2.2181	1.3186
	± 0.3324	± 0.0580	± 0.2979	± 0.3403	± 0.4439	± 0.3263	± 0.3762
400	1.8759	1.9160	0.5629	0.1789	1.0461	1.5495	0.7722
	± 0.2584	± 0.0615	± 0.1272	± 0.0805	± 0.3765	± 0.4716	± 0.2188
600	1.8200	1.9133	0.4039	0.1087	0.7804	1.1168	0.5641
	± 0.2630	± 0.0519	± 0.0931	± 0.0330	± 0.2203	± 0.3783	± 0.1680

Example 1.2. *:* $Y = X_{11}/\{0.5 + (X_{21} + 1.5)^2\} + 0.5 \times \epsilon$.

Example 1.2 is a model whose corresponding vector version is favored by SIR (Li, 1991; Xia, et al. 2002). Here, $\mathcal{S}_{Y|\circ\mathbf{X}\circ} = \mathcal{S}_{E(Y|\circ\mathbf{X}\circ)} = \mathcal{S}_{E(Y|\text{vec}(\mathbf{X}))} = \mathcal{S}_{E(Y|\text{vec}(\mathbf{X}))} = \mathcal{S}(e_1 \otimes e_1, e_1 \otimes e_2)$ and $d = 2$, $r = 1$. Thus, unlike Example 1, estimating the CMFS through the two-step decomposition method or CMS of the vec($\mathbf{X}$) by rMAVE can recover the dimension folding subspace. And again, we use the two-step method based on OPG, and rMAVE using vec($\mathbf{X}$). Since more parameters need to be estimated, the estimates of "two-step" approach and rMAVE are less accurate than *folded-OPG* and *folded-MAVE* shown in Table 1.2, as expected. The benchmark distance for this example is $\Delta_B = 1.916$. And once again, *folded-MAVE* dominates other folded methods.

Table 1.2: Example 1.2: Accuracy of estimates

n	Δ_f						
	two-step	rMAVE	*folded-OPG*	*folded-MAVE*	folded-SIR	folded-SAVE	folded-DR
200	0.7518	0.9477	0.6751	0.2938	0.6377	1.8523	0.8348
	± 0.2165	± 0.1805	± 0.2092	± 0.1059	± 0.1602	± 0.2088	± 0.3018
400	0.4242	0.5093	0.4064	0.1747	0.4205	1.6929	0.5347
	± 0.1020	± 0.0989	± 0.0947	± 0.0379	± 0.1129	± 0.3299	± 0.1537
600	0.3276	0.3639	0.3199	0.1375	0.3340	1.4734	0.4156
	± 0.0683	± 0.0600	± 0.0836	± 0.0330	± 0.0854	± 0.3440	± 0.1156

Example 1.3. *: $Y = X_{11} + (X_{21} + X_{22})^2 + 0.5 \times \epsilon$.*

Let $e_{12} = e_1 + e_2$. In this example, $\mathcal{S}_{Y|\circ\mathbf{X}^\circ} = \mathcal{S}_{E(Y|\circ\mathbf{X}^\circ)} = \mathcal{S}(e_1 \otimes e_1,\ e_1 \otimes e_2,\ e_{12} \otimes e_1, e_{12} \otimes e_2) = \mathcal{S}(e_1 \otimes e_1,\ e_1 \otimes e_2,\ e_2 \otimes e_1, e_2 \otimes e_2)$. This model contains both a linear term and a quadratic term, $d = r = 2$ and $\Delta_B = 2.586$. Table 1.3 shows that *folded-MAVE* is the best, followed by *folded-OPG* and folded-DR. It appears that folded-SIR and folded-SAVE miss the folding subspace, which may be expected, as folded-SIR will miss the quadratic term while folded-SAVE may miss the linear term.

Table 1.3: Example 1.3: Accuracy of estimates

n	Δ_f				
	folded-OPG	*folded-MAVE*	folded-SIR	folded-SAVE	folded-DR
200	1.2780	0.8492	2.2559	2.2327	1.5927
	± 0.3763	± 0.4659	± 0.1856	± 0.1646	± 0.4533
400	0.8231	0.5526	2.2169	2.2159	1.1061
	± 0.2708	± 0.4058	± 0.2004	± 0.1881	± 0.4423
600	0.6416	0.3674	2.1608	2.1930	0.8203
	± 0.1878	± 0.2441	± 0.2879	± 0.2120	± 0.2692

Example 1.4. *:* $Y = X_{11} + 2 \times X_{21}^2 + 3 \times X_{12}^2 + 4 \times X_{22}^2 + 0.2 \times \epsilon$.

This example is a modified model of Wang and Yin (2008). Here, $d = r = 2$ and $\Delta_B = 2.586$, and $\mathcal{S}_{Y|\circ\mathbf{X}^\circ} = \mathcal{S}_{E(Y|\circ\mathbf{X}^\circ)} = \mathcal{S}(e_1 \otimes e_1,\ e_1 \otimes e_2,\ e_2 \otimes e_1, e_2 \otimes e_2)$. Since there are more quadratic terms in the model, it is expected that folded-SAVE's performance will be better than that of folded-SIR. In fact, as shown in Table 1.4, folded-SAVE and folded-DR are better than that of the *folded-OPG*, but *folded-MAVE* is still the best.

Table 1.4: Example 1.4: Accuracy of estimates

n	Δ_f				
	folded-OPG	*folded-MAVE*	folded-SIR	folded-SAVE	folded-DR
200	1.4537 ± 0.4121	0.3849 ± 0.4027	2.4870 ± 0.2177	1.2641 ± 0.4013	1.1567 ± 0.5869
400	0.9487 ± 0.2819	0.1277 ± 0.0507	2.3854 ± 0.1976	0.6262 ± 0.1648	0.5869 ± 0.1615
600	0.7452 ± 0.2742	0.0781 ± 0.0270	2.2716 ± 0.2676	0.5090 ± 0.1344	0.4338 ± 0.1158

Example 1.5. *:* $Y = X_{11} \times (X_{12} + X_{21} + 1) + 0.2 \times \epsilon$.

This is Example 1.1, but with $\text{vec}(\mathbf{X}) \sim N_{pq}(0, \Sigma_{\text{vec}(\mathbf{X})})$, where $\Sigma_{\text{vec}(\mathbf{X})}$ is a $pq \times pq$ positive definite matrix with (j_1, j_2)-th entry $0.5^{|j_1 - j_2|}$. Thus the predictor variables are moderately correlated. Table 1.5 indicates that *folded-OPG* and *folded-MAVE* work well, and both are better than the folded-SIR, folded-SAVE and folded-DR when $\mathcal{S}_{Y|\circ\mathbf{X}^\circ}$ coincides $\mathcal{S}_{E(Y|\circ\mathbf{X}^\circ)}$. The respective estimates, are as expected, found to be less accurate than that of the model with independent variables shown in Example 1.1.

Table 1.5: Example 1.5: Accuracy of estimates

n	Δ_f				
	folded-OPG	*folded-MAVE*	folded-SIR	folded-SAVE	folded-DR
200	1.3609 ± 0.4006	0.5792 ± 0.4455	1.8173 ± 0.4242	2.2927 ± 0.2464	1.6098 ± 0.3927
400	0.8230 ± 0.2622	0.2512 ± 0.1196	1.2087 ± 0.3780	1.9869 ± 0.4146	1.0097 ± 0.3381
600	0.6389 ± 0.1773	0.2191 ± 0.0733	0.9707 ± 0.3002	1.6963 ± 0.4638	0.7293 ± 0.2676

Example 1.6. *: Signal-ratio study*

We check the changes of the performance of our *folded-MAVE* method based on the signal-noise ratio for Example, 1.1, 1.2, 1.3 and 1.4. The error term ϵ follows a standard normal distribution. We apply a series of signal-noise ratios to the first four models and fit the change trend. We run Example 1.1 and 1.3 with the noise terms: 0.2ϵ, 0.5ϵ, ϵ, 1.5ϵ, 2ϵ, 2.5ϵ, 3ϵ, 4ϵ, 6ϵ, 8ϵ, 10ϵ. And for Example 1.2, we choose noise terms: 0.2ϵ, 0.5ϵ, ϵ, 1.5ϵ, 2ϵ, 2.5ϵ, 3ϵ, 4ϵ, 6ϵ, 8ϵ, 9ϵ. The estimation for Example 1.4 is more accurate comparing to the other 3 models for large errors. Thus we apply noise terms 0.5ϵ, ϵ, 2ϵ, 4ϵ, 6ϵ, 8ϵ, 10ϵ, 12ϵ, 16ϵ, 20ϵ and 24ϵ to Example 1.4. Figure 1.1 shows as the error increases, the average of Δ_f increases, which indicates less accuracy. And lager sample size n helps to enhance the estimation accuracy.

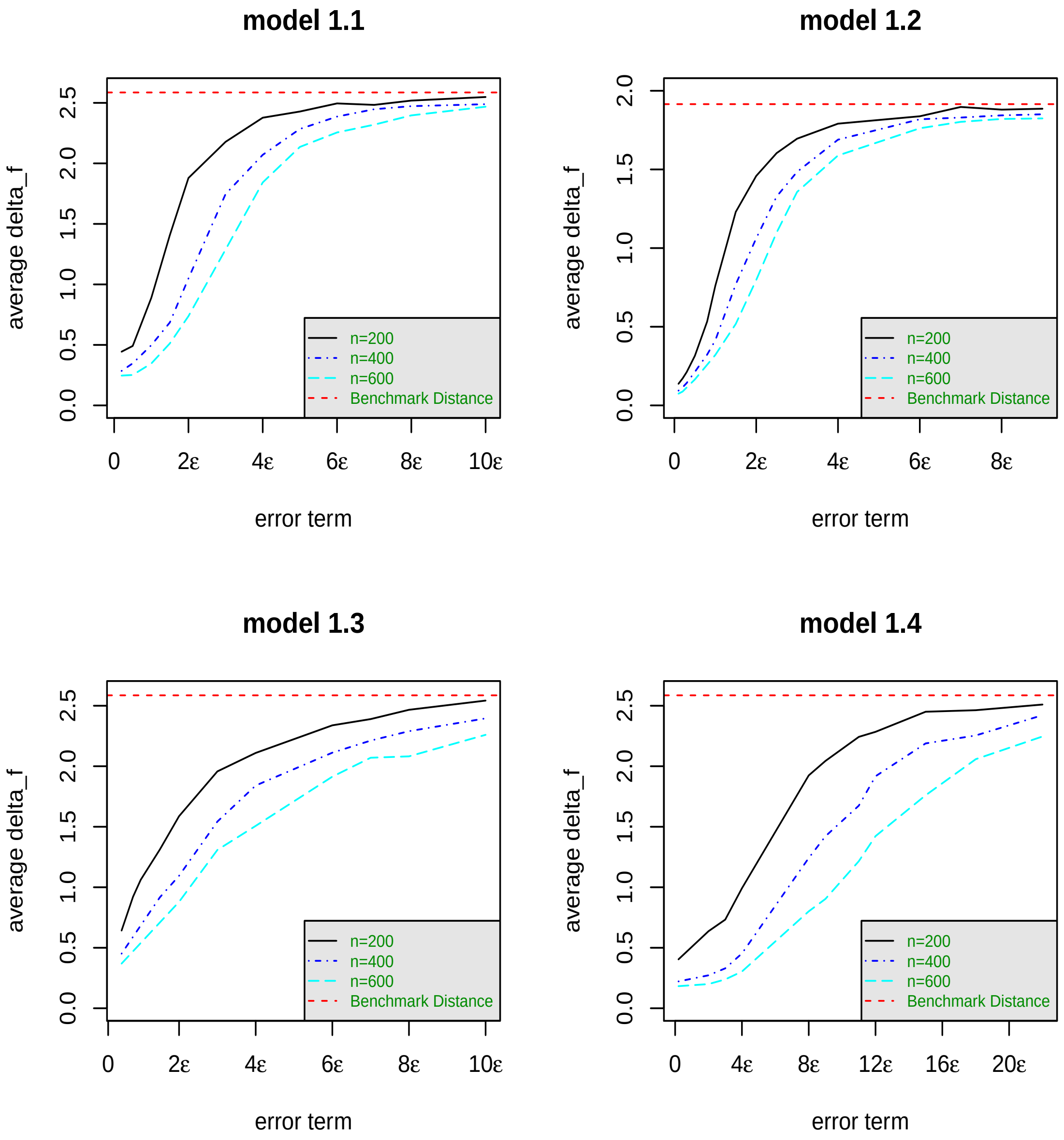

Figure 1.1: Change of the average of Δ_f as error increases

Table 1.6 reports the percentage of the correctly estimated dimensions by our modified BIC criterion. With sample size increasing, for each example, the accuracy of correctly estimated (d, r) is improved. The performance of the modified BIC is also less accurate for correlated data (Example 1.5) than that in the corresponding model with independent variables (Example 1.1). Nevertheless, the modified BIC criterion works well in these limited simulations.

Table 1.6: Percentage of correctly estimate d and r using the modified BIC criterion.

n	Example				
	1.1	1.2	1.3	1.4	1.5
200	88%	93%	54%	81%	62%
400	97%	100%	68%	92%	76%
600	100%	100%	69%	97%	88%

1.5 Application

In this section, we analyze part of the longitudinal data from the follow-up to a Mayo Clinic trial on primary biliary cirrhosis (PBC). The follow-up was conducted between 1974 and 1984 and its description can be found in Fleming & Harrington (1991) and Murtaugh et al. (1994). The data we used are available at http://lib.stat.cmu.edu/datasets/pbcseq. The data set contains multiple laboratory results for 312 patients. Primary biliary cirrhosis is a chronic, progressive cholesteric liver disease of adults that leads to liver failure and the need of transplantation or death (Talwalkar and K. D Lindor, 2003). Several biomarkers such as biliribin, albumin and prothrombin time are adopted to diagnose PBC. Müller (2005) investigated the PBC data from the Mayo Clinic trail to predict the long-term survival as function of the repeated longitudinally recorded level of serum bilirubin. Albert and Shih (2010) proposed an approach for jointly modeling multiple longitudinal measurements and discrete time-to-event data. They applied their method to verify whether the biomarkers are prognostic for the time to either liver transplantation needed or death. Those two papers worked on the log-transformed measurements of the covariates. Kim et al. (2000) presented an abbreviated model to assess the PBC disease risk score based on the measured level of bilirubin, albumin and other markers.

We work on the PBC data from the aspect of dimension folding. We treat the time baseline as one fold of covariates, and the multivariate predictors repeatedly measured over time as another fold of covariates and thus we have a matrix formed predictor. *folded-MAVE* can reduce the dimensions on both folds at the same time as much as possible without losing any information and protect the longitudinal structure of the predictors as well. We concentrate our attention on the measurement of biliribin, albumin level and prothrombin time at time point: 6-month, 1-year, 2-year and 3-year so that the predictor is a 3×4 random matrix.

The response is the time in years between registration and the earlier of transplanting or death. We treat visits between day 90 and day 270 from the enrollment as in the group at time point 6-months. Visits between day 270 and day 550, between day 550 and day 910, between day 910 and day 1275 from the enrollment are identified as at time point 1-year, 2-year and 3-year, respectively. There are 187 patients who had full record at those four time points. However, some patients whose transplant-free or survival time approximates but is less than 3 years are also considered in the analysis, as long as they visited at the four time points.

Sufficient dimension reduction for the analysis of longitudinal data was first used by Li and Yin (2009). They treated time as a discrete random variable. Conditioning on time, they suggested a partial ordinary least squares method, which is an analogy of the partial OLS method in Li, Cook and Chiaromonte (2003), to estimate the partial central mean subspace and then combined all the partial central mean subspaces together. However, their method missed the correlation structure among different time points. Pfeiffer, Forzani and Bura (2011) proposed a longitudinal version of sliced inverse regression (LSIR) to reduce the dimensions of longitudinally measured predictor by assuming that the first and second moments of the predictors can be decomposed into a time and a marker component via a Kronecker product structure. LSIR takes the correlation across time into account, but it requires the assumption of a linearity condition and it can not exhaustively estimate the directions for binary outcomes. The *folded-MAVE* method considers the longitudinal data as a matrix, and the correlations among the structure.

By applying the modified BIC criterion to the data, we have $\hat{d} = \hat{r} = 1$. The estimated bases for the left- and right- mean folding subspaces in the $\mathbf{X}$-scale are $A_{\mathbf{X}} = (\alpha_1, \alpha_2, \alpha_3)^\intercal = (0.11029205, -0.99246655, -0.05334607)^\intercal$ and $B_{\mathbf{X}} = (\beta_1, \beta_2, \beta_3, \beta_4)^\intercal = (-0.2277369, -0.2320542, -0.1827723, -0.9278367)^\intercal$. To test the significance of each coefficient in $A_{\mathbf{X}}$ and $B_{\mathbf{X}}$, we compute the 95% bootstrap confidence intervals of $A_{\mathbf{X}}$ and $B_{\mathbf{X}}$ with 1000 bootstrap samples. Table 1.7 indicates the confidence intervals that show, at 0.05 level, serum biliribin and albumin level at year 3 significantly affect the length of the time between registration and the earlier of transplanting or death.

The plot of the top panel in Figure 1.2 shows the summary plot. That is, the response versus the reduced predictor. Taking the minus sign in $B_{\mathbf{X}}$ into account, biliribin has a negative relationship and albumin level has a positive relationship with the transplant-free or survival time, which is consistent with the already known medical outcome (Shapiro, Smith and Schaffner, 1979). Time point year 3 contributed significantly to the time component, which shows the progressive nature of the disease.

The *folded-MAVE* allows us to assess longitudinal regression in a low-dimensional circumstance. Next we fit a regression model based on the reduced data using the smoothing spline method with degree of freedom 4 and the smoothing parameter $\lambda = 0.01842554$ (Fan and Yao, 2003). The fitted smoothing spline is the imposed curve on the top panel in Figure 1, while the residual against the fitted response is on the bottom panel in Figure 1.2. The regression fit appears adequate and the residual plot shows the model is very reasonable.

Table 1.7: 95% bootstrap confidence interval

	lower bound	upper bound
α_1	0.06994139	0.17586169
α_2	-0.99700777	-0.97023943
α_3	-0.18019987	0.09014473
β_1	-0.4526979	0.11500118
β_2	-0.6973135	0.02418339
β_3	-0.6057525	0.15929250
β_4	-0.9745220	-0.64108263

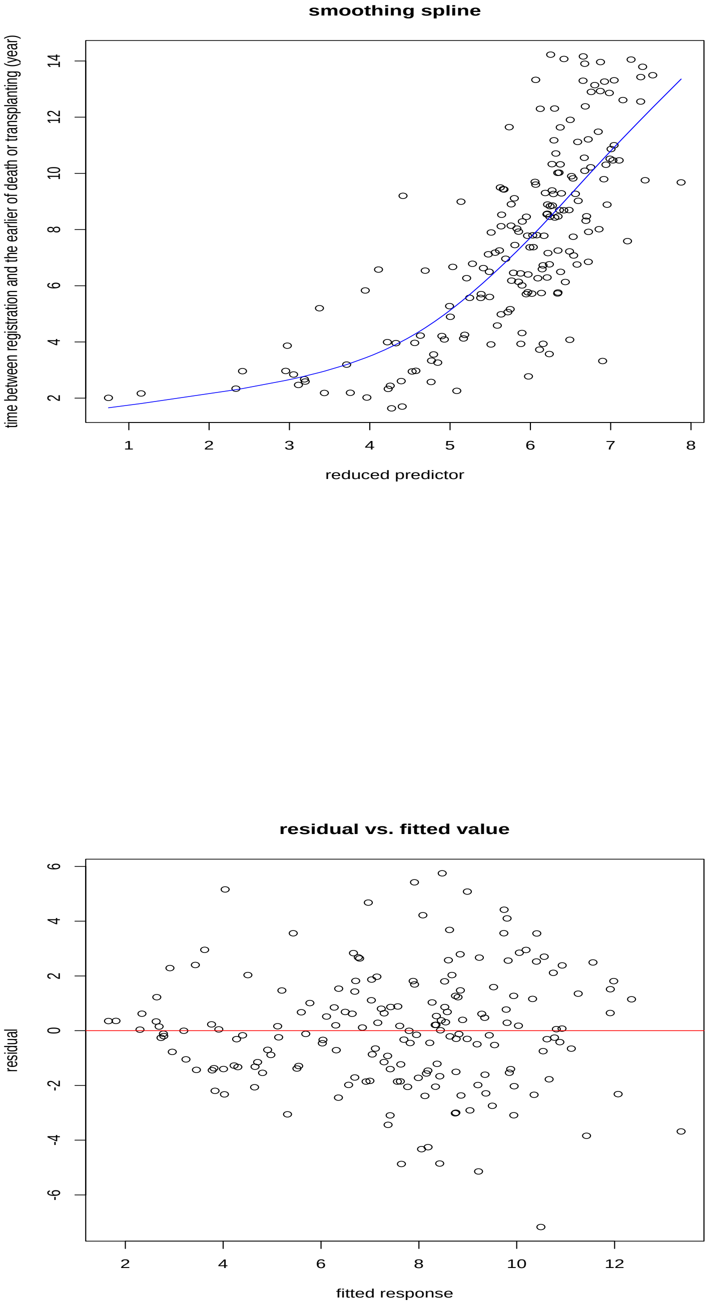

Figure 1.2: Top panel: Summary plot; Bottom panel: Residual plot.

1.6 Generalization to array-valued predictors

In this section, we briefly discuss how to extend the theory and methodology for the matrix-valued predictors to array-valued predictors. Let $\mathbf{X} = \{X_{j_1 j_2 \dots j_k} : j_1 = 1, \dots, p_1, \dots, j_k = 1, \dots, p_k\}$ be a k-way random array of dimension $p_1 \times \cdots \times p_k$. Parallel to Definition 1.1, we define the following:

Definition 1.3. *If there are $p_i \times d_i$ matrices α_i ($d_i \leq p_i$) for $i = 1, \dots, k$ such that $Y \perp\!\!\!\perp E(Y|\mathbf{X}) | (\alpha_k \otimes \cdots \otimes \alpha_1)^T vec(\mathbf{X})$, then the column space of α_i is called the ith mean dimension folding space. Hence, the column space of $\alpha_k \otimes \cdots \otimes \alpha_1$ is called a dimension folding space for the conditional mean of $Y|\mathbf{X}$, or a mean dimension folding space of $Y|\mathbf{X}$.*

Similar to Definition 1.2, the intersection of all such spaces, if itself is a mean dimension folding space, is called the Central Mean Folding Space for Y on the array-valued $\mathbf{X}$. We write this space as $\mathcal{S}_{E(Y|\mathbf{X}^{\circ k})}$. Thus, similar theory on $\mathcal{S}_{E(Y|\mathbf{X}^{\circ k})}$ and methods on the estimation of $\mathcal{S}_{E(Y|\mathbf{X}^{\circ k})}$ can be established straightforwardly from previous sections.

1.7 Discussion

In this chapter, we establish a theory of sufficient dimension folding for the regression mean function with matrix-/array-valued predictors. We illustrate two algorithms for reducing the dimension for matrix-valued predictor along the row and column direction at the same time by using local weighted least squares and linear approximation. We introduce a modified BIC criterion to estimate the dimensionality of the proposed folding subspaces. We analyze a primary biliary cirrhosis data to show the efficacy of our method, a novel approach for longitudinal data. A bootstrap method is used to evaluate the significance of individual variables. However, variable selection integrating penalty methods such LASSO penalty into the local estimation method is also a possible approach for variable selection. Thus reducing the dimension and variable selection may be carried out simultaneously. Such an idea is currently under investigation.

1.8 Appendix

Proof of Proposition 1.1:

1. (3) ⇒ (1) and (1) ⇒ (2) are immediate.

2. (3) ⇒ (2) is also immediate. Since if $E(Y|\mathbf{X})$ is a function of $A^T\mathbf{X}B$ then, given $A^\intercal\mathbf{X}B$, $E(Y|\mathbf{X})$ is a constant and hence independent of any other random variable.

3. For proof of (2) ⇒ (3). By (2) we have

$$E[YE(Y|\mathbf{X})|A^\intercal\mathbf{X}B] = E[Y|A^\intercal\mathbf{X}B]E[E(Y|\mathbf{X})|A^\intercal\mathbf{X}B].$$

The left hand side is

$$E\{E[YE(Y|\mathbf{X})|\mathbf{X}]|A^\intercal\mathbf{X}B\} = E\{[E(Y|\mathbf{X})]^2|A^\intercal\mathbf{X}B\}.$$

The right hand side is

$$E[E(Y|\mathbf{X})|A^{\intercal}\mathbf{X}B]E[E(Y|\mathbf{X})|A^{\intercal}\mathbf{X}B] = \{E[E(Y|\mathbf{X})|A^{\intercal}\mathbf{X}B]\}^2.$$

Thus, $Var[E(Y|\mathbf{X})|A^{\intercal}\mathbf{X}B] = 0$ which means given $A^{\intercal}\mathbf{X}B$, $E(Y|\mathbf{X})$ is a constant. $E(Y|\mathbf{X})$ is a function of $A^{\intercal}\mathbf{X}B$.

Proof of Proposition 1.2: Let A and B be basis matrices of $\mathcal{S}_{E(Y|\circ\mathbf{X})}$ and $\mathcal{S}_{E(Y|\mathbf{X}\circ)}$, respectively. Since $\mathbf{Z}$ and $\mathbf{X}$ are one to one correspondent, the following equivalences are valid:

$$\begin{aligned} Y \perp\!\!\!\perp E(Y|\mathbf{X})|A^{\intercal}\mathbf{X}B &\Leftrightarrow Y \perp\!\!\!\perp E(Y|\mathbf{X})|A^{\intercal}A_0^{-\intercal}A_0^{\intercal}\mathbf{X}B_0B_0^{-1}B \\ &\Leftrightarrow Y \perp\!\!\!\perp E(Y|\mathbf{Z})|(A_0^{-1}A)^{\intercal}\mathbf{Z}B_0^{-1}B \end{aligned}$$

Therefore, $\mathcal{S}(A_0^{-1}A) = A_0^{-1}\mathcal{S}_{E(Y|\circ\mathbf{X})}$ is a left mean dimension folding space for $E(Y|\mathbf{Z})$ and $\mathcal{S}(B_0^{-1}B) = B_0^{-1}\mathcal{S}_{E(Y|\mathbf{X}\circ)}$ is a right mean dimension folding space for $E(Y|\mathbf{Z})$. Consequently,

$$\mathcal{S}_{E(Y|\circ\mathbf{Z})} \subseteq A_0^{-1}\mathcal{S}_{E(Y|\circ\mathbf{X})}, \quad \mathcal{S}_{E(Y|\mathbf{Z}\circ)} \subseteq B_0^{-1}\mathcal{S}_{E(Y|\mathbf{X}\circ)},$$

which means $\mathcal{S}_{E(Y|\circ\mathbf{Z}\circ)} \subseteq (B_0^{-1} \otimes A_0^{-1})\mathcal{S}_{E(Y|\circ\mathbf{X}\circ)}$. By the same argument, $\mathcal{S}_{E(Y|\circ\mathbf{X})} \subseteq A_0\mathcal{S}_{E(Y|\circ\mathbf{Z})}$ and $\mathcal{S}_{E(Y|\mathbf{X}\circ)} \subseteq B_0\mathcal{S}_{E(Y|\mathbf{Z}\circ)}$ so that $\mathcal{S}_{E(Y|\circ\mathbf{X}\circ)} \subseteq (B_0 \otimes A_0)\mathcal{S}_{E(Y|\circ\mathbf{Z}\circ)}$. Thus $\mathcal{S}_{E(Y|\circ\mathbf{X}\circ)} = (B_0 \otimes A_0)\mathcal{S}_{E(Y|\circ\mathbf{Z}\circ)}$.

Proof of Proposition 1.3: We use the same assumption as Xia, et al. (2002) but on vec($\mathbf{X}$). In addition, we assume that $\lim_{n\to\infty}\frac{C_n}{nh^{dr}_{(d,r)}}=0$. The proof below follows from Wang and Yin (2008). For simplicity, we assume that vec($\mathbf{X}$) has a compact support over which its density is positive. For any (d_*, r_*), suppose that the semi-orthogonal $p\times d_*$ and $q\times r_*$ matrices of A_{d*} and B_{r*} minimizes $E\left[Y-E(Y|A^{\intercal}_{d*}\mathbf{X}B_{r*})\right]^2$. That is, $Y_i=E(Y|A^{\intercal}_{d*}\mathbf{X}_iB_{r*})+\epsilon_{d*r*i}$, where $E(\epsilon_{d*r*i}|A^{\intercal}_{d*}\mathbf{X}_iB_{r*})=0$. Suppose the semi-orthogonal $p\times d_*$ matrix $\hat{A}_{d*}$ and semi-orthogonal $q\times r_*$ matrix $\hat{B}_{r*}$ are the sample estimators of A_{d*} and B_{r*}, respectively. And suppose $\hat{\sigma}^2_{\hat{A}_{d*},\hat{B}_{r*}}(\hat{A}^{\intercal}_{d*}\mathbf{X}_j\hat{B}_{r*})=\sum_{i=1}^n[y_i-\hat{c}_j-\hat{a}_j^{\intercal}\hat{A}^{\intercal}\mathbf{x}_{ij}\hat{B}\hat{b}_j]^2\omega_{ij}$. Using the notation in Section 2 and 3,

$$
\begin{aligned}
&\frac{1}{n}RSS_{(d*,r*)}-E\left[Y-E(Y|A^{\intercal}_{d*}\mathbf{X}B_{r*})\right]^2\\
=&\frac{1}{n}\sum_{j=1}^{n}\hat{\sigma}^2_{\hat{A}_{d*},\hat{B}_{r*}}(\hat{A}^{\intercal}_{d*}\mathbf{X}_j\hat{B}_{r*})-\frac{1}{n}\sum_{j=1}^{n}\sigma^2_{A_{d*},B_{r*}}(A^{\intercal}_{d*}\mathbf{X}_jB_{r*})+\frac{1}{n}\sum_{j=1}^{n}\sigma^2_{A_{d*},B_{r*}}(A^{\intercal}_{d*}\mathbf{X}_jB_{r*})\\
&-E\left[Y-E(Y|A^{\intercal}_{d*}\mathbf{X}B_{r*})\right]^2\\
=&\frac{1}{n}\sum_{j=1}^{n}\left[\hat{\sigma}^2_{\hat{A}_{d*},\hat{B}_{r*}}(\hat{A}^{\intercal}_{d*}\mathbf{X}_j\hat{B}_{r*})-\frac{1}{n}\sigma^2_{A_{d*},B_{r*}}(A^{\intercal}_{d*}\mathbf{X}_jB_{r*})\right]+o_p(1)\\
=&\frac{1}{n}\sum_{j=1}^{n}\left[\sum_{i}^{n}\epsilon^2_{d*r*i}\omega_{ij}-\sigma^2_{A_{d*},B_{r*}}(A^{\intercal}_{d*}\mathbf{X}_jB_{r*})\right]+o_p(1)\\
=&o_p(1).
\end{aligned}
$$

The second and the last equations hold by the law of large numbers, and the third equation is based on Lemma 1 of Xia, et al. (2002, the long version) and the law of large numbers. Note that $E(Y|\mathbf{X}) = E(Y|A_d^{\intercal}\mathbf{X}B_r)$.

$$
\begin{aligned}
\frac{1}{n}[RSS_{(d*,r*)} - RSS_{(d,r)}] =& E\left[Y - E(Y|A_{d*}^{\intercal}\mathbf{X}B_{r*})\right]^2 - E\left[Y - E(Y|A_d^{\intercal}\mathbf{X}B_r)\right]^2 + o_p(1) \\
=& E\left[Y - E(Y|A_d^{\intercal}\mathbf{X}B_r) + E(Y|A_d^{\intercal}\mathbf{X}B_r) - E(Y|A_{d*}^{\intercal}\mathbf{X}B_{r*})\right]^2 \\
& - E\left[Y - E(Y|A_d^{\intercal}\mathbf{X}B_r)\right]^2 + o_p(1) \\
=& 2E\left\{\left[Y - E(Y|A_d^{\intercal}\mathbf{X}B_r)\right]\left[E(Y|A_d^{\intercal}\mathbf{X}B_r) - E(Y|A_{d*}^{\intercal}\mathbf{X}B_{r*})\right]\right\} \\
& + E\left[E(Y|A_d^{\intercal}\mathbf{X}B_r) - E(Y|A_{d*}^{\intercal}\mathbf{X}B_{r*})\right]^2 + o_p(1).
\end{aligned}
$$

By the chain rule of conditional expectation, the cross-product is zero. We have

$$
\frac{1}{n}[RSS_{(d*,r*)} - RSS_{(d,r)}] = E\left[E(Y|A_d^{\intercal}\mathbf{X}B_r) - E(Y|A_{d*}^{\intercal}\mathbf{X}B_{r*})\right]^2 + o_p(1).
$$

Let us consider all possible combinations of (d_*, r_*): 1, $d_* > d$ and $r_* > r$; 2, $d_* > d$ and $r_* = r$; 3, $d_* > d$ and $r_* < r$; 4, $d_* = d$ and $r_* > r$; 5, $d_* = d$ and $r_* = r$; 6, $d_* = d$ and $r_* < r$; 7, $d_* < d$ and $r_* > r$;8, $d_* < d$ and $r_* = r$; 9, $d_* < d$ and $r_* < r$.

For cases, 1, 2, 4 and 5, we have

$$
\lim_{n\to\infty} \frac{1}{n}(RSS_{(d*,r*)} - RSS_{(d,r)}) \geq 0.
$$

This is true because in such cases, $E(Y|A_d^\intercal \mathbf{X} B_r) - E(Y|A_{d*}^\intercal \mathbf{X} B_{r*}) = 0$ could happen. For cases, 3, 6, 7, 8 and 9, we have

$$\lim_{n\to\infty} \frac{1}{n}(RSS_{(d*,r*)} - RSS_{(d,r)}) > 0.$$

This is true because in such cases, $E(Y|A_d^\intercal \mathbf{X} B_r) - E(Y|A_{d*}^\intercal \mathbf{X} B_{r*}) = 0$ never happens; otherwise, it contradicts to the existence of *central mean folding subspace*. Then,

$$BIC_{(d*,r*)} - BIC_{(d,r)} = \log \frac{RSS_{(d*,r*)}}{RSS_{(d,r)}} + \frac{C_n}{nh_{(d,r)}^{dr}}[d_* r_* n^{\frac{4(d_* r_* - dr)}{(4+d_* r_*)(4+dr)}} - dr]. \qquad (1.8.1)$$

If $\lim_{n\to\infty} \frac{C_n}{nh_{(d,r)}^{dr}} = 0$, then

A: For cases 1, 2, 4 and 5, because $d_* r_* - dr \geq 0$, the second term in equation (1.8.1) is greater than or equal to 0, while the first term is also greater than or equal to 0. We have $BIC_{(d*,r*)} \geq BIC_{(d,r)}$, when n is large enough.

B: For cases 3, 6, 7, 8 and 9, because if $d_* r_* - dr \geq 0$, the second term in equation (1.8.1) is greater than or equal to 0; and if $d_* r_* - dr < 0$, the second term in equation (1.8.1) tends to 0; while the first term is always greater than 0. We have $BIC_{(d*,r*)} > BIC_{(d,r)}$, when n is large enough.

Hence, we complete the proof.

Proof of Theorem 1.1: Here we explain how we prove the asymptotic result in Theorem 1.1, assuming that the dimensions of the central right and left mean folding subspace are d and r. Part of the proofs can be referred in Wang and Xia (2008). In particular, we follow the proofs in Yin and Li (2011). In order to be consistent with the symbols we used in the chapter and for notation conciseness, without confusion, we define $\mathbf{B} = B \otimes A$, where A and B are the basis matrices for the central left- and right– mean folding subspace respectively. Then $\mathbf{B}$ is the orthogonal basis for the *central mean folding subspace.* And define $\mathbf{B}_* = B_* \otimes A_*$, $d_j = b_j \otimes a_j$, $X = \text{vec}(\mathbf{X})$, $X_i = \text{vec}(\mathbf{x}_i)$ and $\mathbf{x}$ is a given value of X, where A_* and B_* are any $p \times d_*$ and $q \times r_*$ matrices. We work on the proofs in the circumstance of $B \otimes A$ and vec($\mathbf{X}$). Because, once we have an estimate $\hat{\mathbf{B}}$ of $\mathbf{B}$ and $\hat{\mathbf{B}}$ is Kronecker product structural between $\hat{B}$ and $\hat{A}$, then we can iteratively estimate $\hat{A}$ and $\hat{B}$ by the matrix decomposition method introduced in Section 3.1. We can prove the asymptotic result following Yin and Li (2011) because

In the *folded-MAVE* method, we solve the quadratic minimization:

$$\sum_{j=1}^{n}\sum_{i=1}^{n}\rho_j[y_i - c_j - a_j^\intercal A^\intercal(\mathbf{x}_i - \mathbf{x}_j)Bb_j]^2 w_{ij}, \tag{1.8.2}$$

over $(c_j, a_j, b_j, A, B) \in \mathbb{R}^1 \times \mathbb{R}^d \times \mathbb{R}^r \times \mathbb{R}^p \times \mathbb{R}^q$, subject to $b_j^\intercal b_j = 1$, $A^\intercal A = I_d$, $B^\intercal B = I_r$, for $i = 1, \ldots, n$ and $j = 1, \ldots, n$. Note that (1.8.2) is equivalent to

$$\sum_{j=1}^{n}\sum_{i=1}^{n}\rho_j[y_i - c_j - (b_j \otimes a_j)^\intercal (B \otimes A)^\intercal \text{vec}(\mathbf{x}_i - \mathbf{x}_j)]^2 w_{ij}. \tag{1.8.3}$$

Under our definitions that $\mathbf{B} = B \otimes A$, $d_j = b_j \otimes a_j$, $X = \mathrm{vec}(\mathbf{X})$, $X_i = \mathrm{vec}(\mathbf{x}_i)$ and $X_{ij} = \mathrm{vec}(\mathbf{x}_i - \mathbf{x}_j)$, (1.8.3) transforms to

$$\sum_{j=1}^{n}\sum_{i=1}^{n} \rho_j [y_i - c_j - d_j^{\intercal} \mathbf{B}^{\intercal} X_{ij}]^2 w_{ij}. \tag{1.8.4}$$

Note that the formulation (1.8.4) is exactly what Yin and Li (2011) had for the objective function, if one sets $f(y_i) = y_i$ and $m = 1$. In such a case, when the central mean subspace of vec($\mathbf{X}$) is the same as that of the folding subspace, estimating $\mathbf{B}$ directly by the weighted least squares method will suffice. Hence, the asymptotic results follow immediately. However, when the central mean subspace of vec($\mathbf{X}$) is not the same as that of the folding subspace, estimating $\mathbf{B}$ directly by the weighted least squares method will return a basis matrix for other subspaces rather than the *central mean folding subspace*, usually a subspace of folding space if dimensions are assumed correctly. To prevent the latter case, we notice that the Kronecker product structure of $\mathbf{B}$ can be recovered by using the local iterative methods we have proposed. Thus, the final estimate, $\hat{\mathbf{B}} = \hat{B} \otimes \hat{A}$, is then the weighted least squares solution of (1.8.4) with such a structure. Therefore, we can prove the asymptotic result exactly following Yin and Li's (2011) procedure.

Bibliography

[1] P. S. Albert and J. H. Shih. An approach for jointly modeling multivariate longitudianl measurements and discrete time-to-event data. *The Annals of Applied Statistics* **4**, 1517-1532, 2010.

[2] P. J. Brown, M. G. Kenward and E. E. Bassett. Bayesian discrimination with longitudianl data. *Biostatistics* **2**, 417–432, 2001.

[3] R. D. Cook. On the interpretation of regression plots. *Journal of the American Statistical Association* **89**, 177–190, 1994.

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

[5] R. D. Cook. Regression Graphics. *Wiley, New York.* 1998.

[6] R. D. Cook. and B. Li. Dimension reduction for conditional mean in regression. *The Annals of Statistics* **32**, 455-474, 2002.

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

[8] P. Dutilleul. The MLE algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation* **64**, 105-123, 1999.

[9] J. Fan and I. Gijbels. Local Polynomial Modelling and its Applications. *Chapman & Hall, London.* 1996.

[10] J. Fan and Q. Yao. Nonlinear Time Series: Nonparametric and Parametric Methods. *Springer-Verlag, New York.* 2003.

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

[12] W. R. Kim, R. H.Wiesner, J. J. Poterucha, T. M. Therneau, J. T. Benson, R. A. F. Krom, and E. R. Dickson. Adaptation of the Mayo primary biliary cirrhosis natural history model for application in liver transplant candidates. *Liver Transplantation* **6**, No 4 (July), 489-494, 2000.

[13] C. H. Lee, P. Dutilleul and A. Roy. Comment on “ Model with a Kronecker product covariance structure: estimation and testing” by M. S. Srivastava, T. von Rosen, and D. von Rosen, Mathematical Methods of Statistics, 17 (2008), pp. 357-370. *Mathematical Methods of Statistics* **19**, 88-90, 2010.

[14] 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.

[15] B. Li and S. Wang. On Directional Regression for Dimension Reduction. *Journal of the American Statistical Association* **102**, 997-1008, 2007.

[16] 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.

[17] B. Li, S. Wen and L. Zhu. On a projective resampling method for dimension reduction with multivariate responses. *Journal of the American Statistical Association* **103**, 1177-1186, 2008.

[18] B. Li, H. Zha and F. Chiaromonte. Contour Regression: a general approach to dimension reduction. *The Annals of Statistics* **33**, 1580-1616, 2005.

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

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

[21] N. Lu and D. L. Zimmerman. The likelohood ratio test for a separable covariance matrix. *Statistics & Probability Letters* **73**, 449-457, 2005.

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

[23] K. V. Mardia and C. Goodall. Spatial-temporal analysis of multivariate enviromental monitoring data. *G.P. Patil, C.R. Rao. Multivariate Enviromental Statistics. vol 6. North Holland, New York*, 347-385, 1993.

[24] H. G. Müller. Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics* **32**, 223-240, 2005.

[25] 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.

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

[27] A. Roy and R. Khattree. On implementation of a test for Kronecker product covariance structure for multivariate repeated measures data. *Statistical Methodology* **2**, 297-306, 2005.

[28] R. J. Schott. Matrix analysis for statistics. *Wiley, New York*, 1997.

[29] J. M. Shapiro, H. Smith and F. Schaffner. Serum bilirubin: a prognostic factor in primary biliary cirrhosis. *Gut* **20**, 137-140, 1979.

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

[31] M. S. Srivastava, T. von Rosen and D. von Rosen. Models with a Kronecker product covariance structure: estimation and testing. *Mathematical Methods of Statistics* **17**, 357-370, 2008.

[32] M. S. Srivastava, T. von Rosen and D. von Rosen. Estimation and testing in general multivariate linear models with Kronecker product covariance structure. *The Indian Journal of Statistics* **71-A**, part 2, 137-163, 2009.

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

[34] H. Wang and Y. Xia. Sliced regression for dimension reduction. *Journal of the American Statistical Association* **103**, 811-821, 2008.

[35] Q. Wang and X. Yin. A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse MAVE. *Computational Statistics and Data Analysis* **52**, 4512-4520, 2008.

[36] Y. Xia. A constructive approach to the estimation of dimension reduction directions. *The Annals of Statistics* **35**, 2654-2690, 2007.

[37] 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.

[38] 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**, 1773-1757, 2008.

[39] X. Yin and B. Li. Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *The Annals of Statistics* **39**, 3392-3416, 2011.

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

Chapter 2

Dimension folding for a functional of conditional distribution of matrix- or array-valued objects

2.1 Introduction

Sufficient dimension folding (Li, Kim and Altman, 2010; Chapter 1) is a technology for reducing the dimensions of matrix-/array-valued objects as much as possible while preserving the underlying structure interpretation of the predictor. Sufficient dimension folding subspace (Li, Kim and Altman, 2010) focuses on reducing the dimensions of the matrix-/array-valued predictor ($\mathbf{X}$) as it appears in the conditional distribution of a univariate response Y given $\mathbf{X}$. Li, Kim and Altman (2010) proposed the concept of Kronecker envelope based on which they suggested folded-SIR, folded-SAVE and folded-DR three estimation methods for the CFS. Sufficient mean dimension folding subspace (Chapter 1) is only interested in the conditional mean of Y given $\mathbf{X}$ instead of the full conditional distribution itself. To estimate the CMFS, Xue and Yin (2012) constructed two local estimation methods: *folded-OPG* and *folded-MAVE*. However, if the assumption of homoscedasticity is violated or outliers present, we need to consider the robust regression relationship between a univariate response Y and a matrix-/array-valued predictor $\mathbf{X}$.

For a vector-valued predictor, there is a huge literature of research on robustness. For instance, trimmed mean estimator, M-estimator and Huber estimator (Huber, 1964). Quantile regression is an important robust statistics (Koenker, 2005; Wu, Yu and Yu, 2010; Yu and Jones, 1998; Zou and Yuan, 2008; Kai, Li and Zou, 2010). However, to our knowledge, a robust approach of dimension folding for matrix-/array-valued predictor has not been studied by any researcher yet. In this chapter, we reformulate dimension folding and characterize specific aspects of the conditional distribution of Y given a matrix-/array-valued predictor by adopting a general functional of the conditional distribution. By performing dimension folding in reference to a functional, the relation between Y and the predictor reflected in that functional is preserved. For a general functional T, we propose a concept of central T dimension folding subspace (CTFS). The CFS and CMFS can be synthesized as two special cases of the CTFS. Besides, we introduce several other special cases of the CTFS, such as the central variance dimension folding subspace (CVFS), central $k-$th moment dimension folding subspace (CKMFS) and central quantile dimension folding subspace (CQFS). Further, we establish a class of estimation methods for dimension folding on robust estimators. We focus on estimating the CQFS as a special example to illustrate the proposed methods. A special case when the predictor is vector-valued can be found in Yin and Li (2012).

The rest of this chapter is organized as follows. In Section 2.2, we define the central folding space for a general functional of conditional distribution of matrix objects and investigate its properties as well as several special cases. In Section 2.3, we introduce a general approach to estimate the CTFS. We consider estimating the CQFS as an illustration on how to apply the proposed algorithms to estimate a certain aspect of the functional in Section 2.4. Simulations on estimating CQFS and application are included in Section 2.5 and Section 2.6 respectively, followed by a brief extension to array-valued predictors in Section 2.7. A short discussion is provided in Section 2.8.

2.2 Central folding space for a functional of conditional distribution of matrix object

Without confusion, we will let $\mathcal{S}(A)$ denote a space spanned by the columns of a matrix A. We use $\mathcal{S}_{(\cdot)}$ to denote a dimension-folding subspace, where the subscript indicates what response and predictors are involved, and whether the whole conditional distribution or only the conditional mean, variance or quantiles is of interest. And we use

P_A to denote the orthogonal projection onto $\mathcal{S}(A)$, that is, $P_A = A(A^{\intercal}A)^{\dagger}A$, where "$\dagger$" denotes the Moore-Penrose inversion.

Throughout the chapter, we assume that $\mathbf{X}$ is a $p \times q$-dimensional random matrix and Y is a random variable and their sample points are $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, where n is the sample size. We assume that the support Ω of $(\mathbf{X}, Y)$ is the Cartesian product $\Omega_{\mathbf{X}} \times \Omega_Y$, where $\Omega_{\mathbf{X}} \subseteq \mathbb{R}^{p\times q}$ is the support of $\mathbf{X}$ and $\Omega_Y \subseteq \mathbb{R}^{\mathbf{1}}$ is the support of Y. Let F be the joint distribution of $(\mathbf{X}, Y)$, and $F_{Y|\mathbf{X}}$ be the conditional distribution of $Y|\mathbf{X}$. For each fixed $\mathbf{x} \in \Omega_{\mathbf{X}}$, $F_{Y|\mathbf{X}}(\cdot|\mathbf{x})$ is a probability measure on Ω_Y and for each measurable set $A_Y \subseteq \Omega_Y$, $F_{Y|\mathbf{X}}(A_Y|\mathbf{x})$ is a measurable function of $\mathbf{x}$. Then the set $\{F_{Y|\mathbf{X}}(\cdot|\mathbf{x}) : \mathbf{x} \in \Omega_{\mathbf{X}}\}$ defines a family of probability measures on Ω_Y. Let T be a functional defined on this family:

$$T : \{F_{Y|\mathbf{X}}(\cdot|\mathbf{x}) : \mathbf{x} \in \Omega_{\mathbf{X}}\} \to \mathbb{R}^{\mathbf{1}}.$$

The functional T can be viewed as a parameter associated with the conditional distribution $F_{Y|\mathbf{X}}$. From here on, we will abbreviate this functional $T(F_{Y|\mathbf{X}}(\cdot|\mathbf{x}))$ as $T(\mathbf{x})$. The functional T, when considered as a mapping $F_{Y|\mathbf{X}}(\cdot|\mathbf{x}) \to T(F_{Y|\mathbf{X}}(\cdot|\mathbf{x}))$, is a functional of conditional distribution, when considered as a mapping $\mathbf{x} \to T(F_{Y|\mathbf{X}}(\cdot|\mathbf{x}))$, is simply a function of $\mathbf{x}$.

In the sufficient dimension folding, one considers finding matrices $A \in \mathbb{R}^{p\times d}$ and $B \in \mathbb{R}^{q\times r}$ such that d and r are as small as possible and $F_{Y|\mathbf{X}}$ depends on $\mathbf{X}$ only through $A^{\intercal}\mathbf{X}B$, while preserving the matrix structure of $\mathbf{X}$ (Li, Kim and Altman, 2010). Spaces spanned by the columns of A and B are called the left- and right- dimension folding subspace, respectively. Let $\mathcal{S}_{Y|\circ\mathbf{X}}$ or $\mathcal{S}_{Y|\mathbf{X}\circ}$ be the intersection of all left- or right- dimension folding subspaces for $Y|\mathbf{X}$. The subspace $\mathcal{S}_{Y|\mathbf{X}\circ} \otimes \mathcal{S}_{Y|\circ\mathbf{X}}$ is defined as the central dimension folding subspace (CFS) denoted by $\mathcal{S}_{Y|\circ\mathbf{X}\circ}$, where "$\otimes$" is the Kronecker product. In Chapter 1, we introduced the central mean dimension folding subspace (CMFS), which particularly focuses on the conditional mean function of $F_{Y|\mathbf{X}}$. We considered finding matrices $A \in \mathbb{R}^{p\times d}$ and $B \in \mathbb{R}^{q\times r}$, $d \leq p$ and $r \leq q$, such that $Y \perp\!\!\!\perp E(Y|\mathbf{X})|A^{\intercal}\mathbf{X}B$, where "$\perp\!\!\!\perp$" stands for conditional independent. The subspaces $\mathcal{S}(A)$ and $\mathcal{S}(B)$ are called a left- or right- mean dimension folding subspace, respectively. If the intersection of all left- or right- mean folding subspace is itself a left- or right- mean folding subspace of $E(Y|\mathbf{X})$, we denoted the intersection subspace by $\mathcal{S}_{E(Y|\circ\mathbf{X})}$ or $\mathcal{S}_{E(Y|\mathbf{X}\circ)}$. We also defined $\mathcal{S}_{E(Y|\circ\mathbf{X})}$ and $\mathcal{S}_{E(Y|\mathbf{X}\circ)}$ as the central left- and right- mean dimension folding subspace, respectively. Then $\mathcal{S}_{E(Y|\mathbf{X}\circ)} \otimes \mathcal{S}_{E(Y|\circ\mathbf{X})}$ is the sufficient central mean dimension folding subspace.

CFS and CMFS are interested in a certain parameter, as described by the above functional T, associate with $F_{Y|\mathbf{X}}$ and others not related to this functional are analogous to the nuisance parameter in a classical setting. Targeting on different dimension folding subspaces, we only need to fold the dimensions of predictors with respect to some specific parameter in the associated functional T (Li, Kim and Altman, 2010; Xue and Yin, 2012).

In this chapter, our goal is to derive a comprehensive method to estimate the dimension folding subspace associated with a general functional T. We define the dimension folding spaces with respect to a functional T of the conditional distribution when $\mathbf{X}$ is a random matrix as:

Definition 2.1. *If there are matrices $A \in \mathbb{R}^{p\times d}$ and $B \in \mathbb{R}^{q\times r}$ $(d \leq p,\ r \leq q)$ such that $T(\mathbf{x})$ depends on $\mathbf{x}$ only through $A^{\intercal}\mathbf{x}B$, that is $T(\mathbf{x}) = T(\mathbf{x}^{*})$ whenever $A^{\intercal}\mathbf{x}B = A^{\intercal}\mathbf{x}^{*}B$, then the column space of A or B are called a left- or right- dimension folding space for the functional T, or T-left- or right- dimension folding space. Let $S_{Y|\circ\mathbf{X}}(T)$ or $S_{Y|\mathbf{X}\circ}(T)$ be the intersection of all T-left- or right- dimension folding space and itself is a left- or right- T dimension folding space. Then $S_{Y|\circ\mathbf{X}}(T)$ and $S_{Y|\mathbf{X}\circ}(T)$ are defined as the central left- and right- T dimension folding space. The space $\mathcal{S}_{Y|\mathbf{X}\circ}(T) \otimes \mathcal{S}_{Y|\circ\mathbf{X}}(T)$ is called the central dimension folding space for functional T, or the central T dimension folding subspace (CTFS) and is written as $\mathcal{S}_{Y|\circ\mathbf{X}\circ}(T)$.*

That the intersection of two T-dimension folding spaces is again a T-dimension folding space can be established under very mild conditions using the same argument recently made by Yin, Li and Cook (2008); see also Proposition 6.4 of Cook (1998) and Theorems 1 and 2 of Li, Kim and Altman (2010). We state it here without proof. The following corollary 2.1 states the condition when the CTFS exists and is the unique minimal T-dimension folding space. We assume that this condition holds throughout the article and denote the dimensionality of $S_{Y|\circ\mathbf{X}}(T)$ or $S_{Y|\mathbf{X}\circ}(T)$ by d or r.

Corollary 2.1. *Suppose that A_i, B_i, $i = 1, 2$, are $p \times d_i$, $q \times r_i$ with $d_i \leq p$ and $r_i \leq q$ such that $\mathcal{S}(B_1 \otimes A_1)$ and $\mathcal{S}(B_2 \otimes A_2)$ are both T-dimension folding subspaces. Let A_3, B_3 be $p \times d_3$ and $q \times r_3$ dimensional matrices whose columns span the intersection of*

$\mathcal{S}(B_1 \otimes A_1)$ and $\mathcal{S}(B_2 \otimes A_2)$. Suppose, in addition, that for each $\mu \in \mathbb{R}^{d_3 \times r_3}$ such that $A_3^{\intercal}\mathbf{x}B_3 = \mu$ for some $\mathbf{x} \in \Omega_{\mathbf{X}}$, the set

$$\{(A_1^{\intercal}\mathbf{x}B_1, A_2^{\intercal}\mathbf{x}B_2) : A_3^{\intercal}\mathbf{x}B_3 = \mu, \mathbf{x} \in \Omega_{\mathbf{X}}\} \tag{2.2.1}$$

is an M-set in $\mathbb{R}^{d_1 \times r_1} \times \mathbb{R}^{d_2 \times r_2}$ as defined by Yin, Li and Cook (2008). Then $\mathcal{S}(B_1 \otimes A_1) \cap \mathcal{S}(B_2 \otimes A_2)$ is also a T dimension folding subspace.

Lemma 2.1 below whose proof is detailed in the Appendix indicates that it is the space spanned by the columns of A or B that we should care about rather than the specific value of the matrices. The definition of central T folding subspace involves using the point 3, but based on the equivalences of points 1 and 2, it is the subspace that uniquely determines the sufficient dimensions.

Lemma 2.1. *Suppose matrices $A \in \mathbb{R}^{p \times d}$ and $B \in \mathbb{R}^{q \times r}$ ($d \leq p$, $r \leq q$) and the column vectors of each matrix are linearly independent, and let $f(t)$ be a function of $t \in \mathbb{R}^{p \times q}$. For any $\Sigma_A \in \mathbb{R}_+^{p \times p}$ and $\Sigma_B \in \mathbb{R}_+^{q \times q}$, let $P(\Sigma_A) = A(A^{\intercal}\Sigma_A A)^{-1}A^{\intercal}\Sigma_A$ and $P(\Sigma_B) = B(B^{\intercal}\Sigma_B B)^{-1}B^{\intercal}\Sigma_B$ be the projection onto $\mathcal{S}(A)$ and $\mathcal{S}(B)$ respectively with respect to inner product $\langle a, b\rangle = a^{\intercal}\Sigma b$. Then the following statements are equivalent:*

1. *there are $\Sigma_A \in \mathbb{R}_+^{p \times p}$ and $\Sigma_B \in \mathbb{R}_+^{q \times q}$ such that $f(t) = f(P^{\intercal}(\Sigma_A)tP(\Sigma_B))$ for all t;*

2. *for every $\Sigma_A \in \mathbb{R}_+^{p \times p}$ and $\Sigma_B \in \mathbb{R}_+^{q \times q}$, $f(t) = f(P^{\intercal}(\Sigma_A)tP(\Sigma_B))$ for all t;*

3. *$f(t)$ depends on t only through $A^{\intercal}tB$; that is, whenever $A^{\intercal}t_1B = A^{\intercal}t_2B$, we have $f(t_1) = f(t_2)$.*

Like the CMFS, under affine linear transformation, the CTFS is invariant. We delay the proof of the following proposition to the Appendix as well.

Proposition 2.1. *Let $\mathbf{Z} = A^{\intercal}\mathbf{X}B$ where A and B are full rank, $p \times p$ and $q \times q$ matrices respectively. Then $\mathcal{S}_{Y|\circ\mathbf{Z}\circ}(T) = (B^{-1} \otimes A^{-1})\mathcal{S}_{Y|\circ\mathbf{X}\circ}(T)$.*

The formulation of CTFS accommodates CFS and CMFS as its special cases, which aims at estimating a certain aspect of the conditional distribution. We can link CFS and CMFS to CTFS as in the following Example 2.1 and 2.2. Also, the formulation of CTFS suggests some other examples as follows.

Example 2.1. *For an $a \in \Omega_Y$, let T_a be defined as the evaluation of $F_{Y|\mathbf{X}}(\cdot|\mathbf{x})$ at $y = a$. That is, T_a assigns to each function $F_{Y|\mathbf{X}}(\cdot|\mathbf{x})$ the number $F_{Y|\mathbf{X}}(a|\mathbf{x})$. Then the union of all $\mathcal{S}_{Y|\circ\mathbf{X}\circ}(T_a)$ reduces to the Central Dimension Folding Space (Li, Kim and Altman, 2010).*

Example 2.2. *If $T(\mathbf{x}) = \int y dF_{Y|\mathbf{X}}(y|\mathbf{x})$, then $T(\mathbf{x})$ is the conditional mean $E(Y|\mathbf{X} = \mathbf{x})$. Thus the CTFS reduces to the Central Mean Dimension Folding Space (Chapter 1).*

Example 2.3. *If $T(\mathbf{x}) = \int y^2 dF_{Y|\mathbf{X}}(y|\mathbf{x}) - (\int y dF_{Y|\mathbf{X}}(y|\mathbf{x}))^2$, then $T(\mathbf{x})$ is the conditional variance, $\mathrm{Var}(Y|\mathbf{X} = \mathbf{x})$, given a matrix predictor $\mathbf{X}$. We define this folding subspace as the central variance dimension folding subspace (CVFS) denoted by $\mathcal{S}_{V(Y|\circ\mathbf{X}\circ)}$.*

Example 2.4. *If $T(\mathbf{x}) = [M^{(1)}(Y|\mathbf{X}), M^{(2)}(Y|\mathbf{X}), \ldots, M^{(k)}(Y|\mathbf{X})]$, where $M^{(k)}(Y|\mathbf{X}) = \int [y - \int y dF_{Y|\mathbf{X}}(y|\mathbf{x})]^k dF_{Y|\mathbf{X}}(y|\mathbf{x})$ for $k \geq 2$ and $M^{(1)}(Y|\mathbf{X}) = \int y dF_{Y|\mathbf{X}}(y|\mathbf{x})$, then we define this special example of the CTFS as the central k-th moment dimension folding subspace (CKMFS).*

The concept of linear functionals, the intrinsically linear functionals and the minimization functionals discussed in Yin and Li (2012) can be easily extended to the domain of dimension folding by replacing the vector-valued predictor X by the matrix-/array-valued predictor $\mathbf{X}$. Their properties also remain the same as for vector predictors. Based on the definition of minimization functionals for the matrix predictor $\mathbf{X}$, we illustrate the *central quantile folding subspace* as follows.

Example 2.5. *Let $\rho(\theta, y)$ be a function of θ and y. Suppose, for each $\mathbf{x} \in \Omega_{\mathbf{X}}$, the following function of θ*

$$\int \rho(\theta, y) dF_{Y|\mathbf{X}}(y|\mathbf{x}) \tag{2.2.2}$$

has a unique minimum. Then the functional T defined as the minimizer of the above function is a minimization functional, or an M-functional. Suppose that $0 < \tau < 1$ and $I(\cdot)$ is the indicator function and let

$$\begin{aligned}\rho_\tau(\theta, y) &= \tau I(y - \theta > 0)(y - \theta) + (\tau - 1) I(y - \theta \leq 0)(y - \theta) \\ &= (y - \theta)[\tau - I(y - \theta \leq 0)].\end{aligned}$$

Then the CTFS of this type is the central quantile folding space (CQFS) denoted by $\mathcal{S}_{Q_\tau}(Y|\mathbf{X})$. If $\tau = 0.5$ or $\rho(\theta, y) = |y - \theta|$, $\mathcal{S}_{Q_{0.5}}(Y|\mathbf{X})$ is the central median folding subspace, denoted by $\mathcal{S}_{M(Y|\circ\mathbf{X}\circ)}$.

Next, we propose a class of methods that can be used to estimate the central T dimension folding subspace for matrix predictors. We also establish a new theory for local central folding space for a functional of conditional distribution, from which we provide a connection between local theory of sufficient dimension folding and sufficient dimension folding. In the population level, local method can be used as a class of estimation method for sufficient dimension folding. As a special case, we use the proposed method to estimate the CQFS for the consideration of robust regression.

2.3 General approach to estimate the CTFS

2.3.1 Measurement of accuracy

To evaluate the accuracy of estimates we use the matrix norm

$$\Delta(B \otimes A, \hat{B} \otimes \hat{A}) = ||P_{B \otimes A} - P_{\hat{B} \otimes \hat{A}}||$$

(Li, Zha and Chiaromonte, 2005). Here, we use the Frobenius norm and denote it as Δ_f. A small value of Δ_f means the two spaces are closed to each other. To understand how accurate the estimates are, we adopt the benchmark distance in Li, Wen and Zhu (2008). Benchmark distance measures the discrepancy between two spaces that are not related at all. Li, Wen and Zhu (2008) defined the benchmark distance as follows. Let $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ be $s \times t$ random matrices whose entries are independent standard normal distributed random variables and satisfy $\boldsymbol{\alpha} \perp\!\!\!\perp \boldsymbol{\beta}$. Let $P_{\boldsymbol{\alpha}}$ and $P_{\boldsymbol{\beta}}$ be the projections onto the column space spanned by $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ respectively. Then $E(||P_{\boldsymbol{\alpha}} - P_{\boldsymbol{\beta}}||)$ is the benchmark distance determined only by the value of s and t. We estimate the benchmark distance by running 10,000 simulations and denote it by Δ_B.

2.3.2 Local central folding space for a functional of conditional distribution

Here, we define a local central folding space for a functional of conditional distribution. We replace Ω by a local support Δ, where $\Delta \subseteq \Omega$. For simplicity, $\Delta = \Delta_{\mathbf{x}} \times \Delta_y$, where $\Delta_{\mathbf{x}} \subseteq \Omega_{\mathbf{X}}$ and $\Delta_y \subseteq \Omega_Y$. For convenience, we denote a local CTFS as $\mathcal{S}_{Y|\circ\mathbf{X}^\circ}(T_\Delta)$, where T_Δ is the same as T but with $(\mathbf{X}, Y) \in \Delta$. Thus if $\Delta = \Omega$, then $\mathcal{S}_{Y|\circ\mathbf{X}^\circ}(T_\Delta) = \mathcal{S}_{Y|\circ\mathbf{X}^\circ}(T)$. We use $\text{vec}(\cdot)$ as the operator that stacks a matrix into a vector, column by column. We also use $\text{mat}_t(\cdot)$ to define the inverse operation of $\text{vec}(\cdot)$, which transforms a vector to a matrix with t rows. For example, for any matrix $\mathbf{M} \in \mathbb{R}^{t\times s}$ and $\mathbf{M} = (\mathbf{m}_1, \ldots, \mathbf{m}_s)$, then $\text{vec}(\mathbf{M}) = (\mathbf{m}_1^\intercal, \ldots, \mathbf{m}_s^\intercal)^\intercal$. Suppose $\mathbf{m} = \text{vec}(\mathbf{M})$ be a vector in $\mathbb{R}^{ts}$, where $ts = t \times s$, then $\text{mat}_t(\mathbf{m}) = \mathbf{M}$. In short, $\text{mat}_t[\text{vec}(\mathbf{M})] = \mathbf{M}$, and $\text{vec}[\text{mat}_t(\mathbf{m})] = \mathbf{m}$.

Let A and B be the $p \times d$ and $q \times r$ basis matrices for $\mathcal{S}_{Y|\circ\mathbf{X}}(T_\Delta)$ and $\mathcal{S}_{Y|\mathbf{X}^\circ}(T_\Delta)$, respectively. Then $B \otimes A$ is a basis matrix of $\mathcal{S}_{Y|\circ\mathbf{X}^\circ}(T_\Delta)$. Suppose $\mathbf{s} = A^\intercal\mathbf{X}B$, then the key fact underlying the subsequent development is

$$\frac{\partial T_\Delta(\mathbf{x})}{\partial \text{vec}(\mathbf{x})} = \frac{\partial \text{vec}(\mathbf{s})^\intercal}{\partial \text{vec}(\mathbf{x})} \cdot \frac{\partial T_\Delta(\mathbf{s})}{\partial \text{vec}(\mathbf{s})} = (B \otimes A) \cdot \frac{\partial T_\Delta(\mathbf{s})}{\partial \text{vec}(\mathbf{s})}, \tag{2.3.1}$$

or

$$\mathrm{mat}_p[\partial T_\Delta(\mathbf{x})/\partial \mathrm{vec}(\mathbf{x})] = A\mathrm{mat}_d[\partial T_\Delta(\mathbf{s})/\partial \mathrm{vec}(s)]B^\intercal \tag{2.3.2}$$

(Schott 1997).

Equation (2.3.1) implies $\partial T_\Delta(\mathbf{x})/\partial \mathrm{vec}(\mathbf{x}) \in \mathcal{S}_{Y|\circ\mathbf{X}^\circ}(T_\Delta)$. We have the following two propositions and their proofs are delayed to the Appendix as well.

Proposition 2.2. $\mathcal{S}\{\partial T_\Delta(\mathbf{x})/\partial vec(\mathbf{x}) : (\mathbf{x}, y) \in \Delta\} = \mathcal{S}_{Y|\circ\mathbf{X}^\circ}(T_\Delta)$

Proposition 2.2 implies that the collection of $\partial T_\Delta(\mathbf{x})/\partial \mathrm{vec}(\mathbf{x})$ for all $(\mathbf{x}, y) \in \Delta$ spans the local CTFS, regardless of the form of the functional T_Δ. If $\Delta = \Omega$, then it characterizes the CTFS. One can develop estimation methods based on Proposition 2.2. For instance, let $w(\mathbf{x}, y) > 0$ be a weight function for all $(\mathbf{x}, y) \in \Delta$, and define $\boldsymbol{\eta}_\Delta = E\left[w(\mathbf{x}, y)\frac{\partial T_\Delta(\mathbf{x})}{\partial \mathrm{vec}(\mathbf{x})}\frac{\partial T_\Delta(\mathbf{x})}{\partial \mathrm{vec}(\mathbf{x})^\intercal}\right]$. Then $\mathcal{S}(\boldsymbol{\eta}_\Delta) = \mathcal{S}_{Y|\circ\mathbf{X}^\circ}(T_\Delta)$. One possible choice is $w(\mathbf{x}, y) \equiv 1$.

The next proposition indicates that it is enough to estimate the specific CTFS based on a local support of $\mathbf{X}$ only. Based on the following proposition, we propose an approach to estimate the CTFS through the M-functionals and the respective local CTFS.

Proposition 2.3. *For any $\Delta \subseteq \Omega$, it is always true that $\mathcal{S}_{Y|\circ\mathbf{X}^\circ}(T_\Delta) \subseteq \mathcal{S}_{Y|\circ\mathbf{X}^\circ}(T)$ and further, $\mathcal{S}\{\mathcal{S}_{Y|\circ\mathbf{X}^\circ}(T_\Delta), \Delta = \Delta_\mathbf{x} \times \Omega_Y\ for\ all\ \mathbf{x} \in \Omega_\mathbf{X}\} = \mathcal{S}_{Y|\circ\mathbf{X}^\circ}(T)$.*

2.3.3 Estimation methods of CTFS

An initial estimator

In this section, we assume d and r, the dimensions of $S_{Y|\circ\mathbf{X}}(T)$ and $S_{Y|\mathbf{X}\circ}(T)$, are known. By using the local linear smoothing (Fan and Gijbels, 1996), for each $j = 1, \ldots, n$, at sample point $\mathbf{x}_j$ we minimize the following objective function over $a_j \in \mathbb{R}^p$, $b_j \in \mathbb{R}^q$ and $c_j \in \mathbb{R}^{\mathbf{1}}$ subject to $b_j^{-1} b_j = 1$. The general objective function is

$$\frac{1}{n} \sum_{i=1}^{n} \rho(y_i, c_j + a_j^{\intercal} \mathbf{x}_{ij} b_j) w_{ij}, \tag{2.3.3}$$

where $\mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j$ and $w_{ij} \geq 0$ is the kernel weight centered at $\mathbf{x}_{ij}$ with $\sum_{i=1}^{n} w_{ij} = 1$. We use the usual kernel weight

$$w_{ij}(h) = K_h(\text{vec}(\mathbf{x}_{ij})) / \sum_{j=1}^{n} K_h(\text{vec}(\mathbf{x}_{ij})),$$

where $K_h(\cdot)$ is the chosen kernel function with the bandwidth $h > 0$. We set $h = n^{-1/(p \times q + 4)} \times \sqrt{p \times q}$ (Silverman, 1986).

In order to estimate a_j, b_j and c_j, we propose the following iterative algorithm:

1. Generate the initial values of $a_j \in \mathbb{R}^p$ from a sample of $N(0, 1)$ random variables.
2. For each $j = 1, \ldots, n$, given a fixed $a_j \in \mathbb{R}^p$, minimize (2.3.3) over $c_j \in \mathbb{R}^{\mathbf{1}}$, $b_j \in \mathbb{R}^q$ subject to $b_j^{-1} b_j = 1$. Suppose the minimizers are $\hat{c}_j$ and $\hat{b}_j$.

3. For each $j = 1, \dots, n$, given the estimate $\hat{c}_j$ and $\hat{b}_j$ from Step 1, minimize (2.3.3) over $a_j \in \mathbb{R}^p$. And suppose the estimate is $\hat{a}_j$.

4. Let $\hat{A}_{(k)}$, $\hat{B}_{(k)}$ be the first d or r leading eigenvectors of

$$\hat{\Sigma}_1 = \frac{1}{n}\sum_{j=1}^{n} \hat{a}_j \hat{a}_j^{\mathsf{T}} \quad \textit{or} \quad \hat{\Sigma}_2 = \frac{1}{n}\sum_{j=1}^{n} \hat{b}_j \hat{b}_j^{\mathsf{T}}$$

   obtained in the k-th iteration. Let $\boldsymbol{\eta}_{(k)} = \hat{B}_{(k)} \otimes \hat{A}_{0(k)}$, $\boldsymbol{\eta}_{0(k-1)} = \hat{B}_{(k-1)} \otimes \hat{A}_{(k-1)}$, where $\hat{A}_{(k-1)}$ and $\hat{B}_{(k-1)}$ are the estimators of A and B in the $(k-1)$-th iteration. If $||\boldsymbol{\eta}_{(k)}\boldsymbol{\eta}_{(k)}^{\mathsf{T}} - \boldsymbol{\eta}_{(k-1)}\boldsymbol{\eta}_{(k-1)}^{\mathsf{T}}||$ is smaller than some pre-specified tolerance value, such as 10^{-6}, stop the iteration and set $A = \hat{A}_{(k)}$, $B = \hat{B}_{(k)}$ as the estimates of $S_{Y|\circ\mathbf{X}}(T)$ and $S_{Y|\mathbf{X}\circ}(T)$ respectively; otherwise, set $k := k+1$ and go to Step 2.

The refined estimator

In order to attain more accurate estimates of A and B, we can further refine the A_{opg} and B_{opg}, the basis matrix for $S_{Y|\circ\mathbf{X}}(T)$ and $S_{Y|\mathbf{X}\circ}(T)$ obtained by the previous procedure. We suggest the following algorithm: given $A_{(k)}$ and $B_{(k)}$, the estimates of A and B in the k-th iteration, we can iteratively estimate $A_{(k+1)}$ and $B_{(k+1)}$ in the next iteration by minimizing the global loss function:

$$n^{-2}\sum_{j=1}^{n}\sum_{i=1}^{n} \rho(y_i, c_j + a_j^{\mathsf{T}} A^{\mathsf{T}} \mathbf{x}_{ij} B b_j) \mathbf{K}_{h_{(d,r)}}(\mathrm{vec}(A_{(k)}^{\mathsf{T}} \mathbf{x}_{ij} B_{(k)})), \tag{2.3.4}$$

subject to $c_j \in \mathbb{R}^{\mathbf{1}}$, $a_j \in \mathbb{R}^d$, $b_j \in \mathbb{R}^r$ with $b_j^\intercal b_j = 1$, $A \in \mathbb{R}^{p\times d}$ with $A^\intercal A = I_d$ and $B \in \mathbb{R}^{q\times r}$ with $B^\intercal B = I_r$. We will use the refined kernel weights and summarize the algorithm as:

1. Generate the initial values of $a_j \in \mathbb{R}^d$ from a sample of $N(0, 1)$ random variables. Set the initial estimates of A and B as A_{opg} and B_{opg}.

2. In the k-th iteration, for a fixed $a_{j(k-1)}$, $A_{(k-1)}$, and $B_{(k-1)}$, calculate $\hat{b}_{j(k)}$ and $\hat{c}_{j(k)}$, the estimates of b_j and c_j, as the minimizers of

$$\sum_{i=1}^{n} \rho(y_i, c_j + a_{j(k-1)}^\intercal A_{(k-1)}^\intercal \mathbf{x}_{ij} B_{(k-1)} b_j)\mathbf{K}_{h_{(d,r)}}(\mathrm{vec}(A_{(k-1)}^\intercal \mathbf{x}_{ij} B_{(k-1)})).$$

3. In the k-th iteration, given the fixed $c_{j(k)}$, $b_{j(k)}$, $A_{(k-1)}$, and $B_{(k-1)}$, $\hat{a}_{j(k)}$, the estimate of $a_{j(k)}$, is the minimizer of

$$\sum_{i=1}^{n} \rho(y_i, c_{j(k)} + a_j^\intercal A_{(k-1)}^\intercal \mathbf{x}_{ij} B_{(k-1)} b_{j(k)})\mathbf{K}_{h_{(d,r)}}(\mathrm{vec}(A_{(k-1)}^\intercal \mathbf{x}_{ij} B_{(k-1)})).$$

4. Given the fixed $a_{j(k)}$, $b_{j(k)}$, $c_{j(k)}$ and $B_{(k-1)}$, minimizing the objective function below over A with $A^\intercal A = I_d$, in the k-th iteration.

$$n^{-2} \sum_{j=1}^{n} \sum_{i=1}^{n} \rho(y_i, c_{j(k)} + a_{j(k)}^\intercal A^\intercal \mathbf{x}_{ij} B_{(k-1)} b_{j(k)})\mathbf{K}_{h_{(d,r)}}(\mathrm{vec}(A_{(k-1)}^\intercal \mathbf{x}_{ij} B_{(k-1)})),$$

to obtain $\hat{A}_{(k)}$

5. Given the $a_{j(k)}$, $b_{j(k)}$, $c_{j(k)}$ and $A_{(k)}$, minimize the objective function:

$$n^{-2}\sum_{j=1}^{n}\sum_{i=1}^{n}\rho(y_i, c_{j(k)} + a_{j(k)}^{\intercal}A_{(k)}^{\intercal}\mathbf{x}_{ij}Bb_{j(k)})\mathbf{K}_{h_{(d,r)}}(\text{vec}(A_{(k)}^{\intercal}\mathbf{x}_{ij}B_{(k-1)})),$$

   over B subject to $B^{\intercal}B = I_r$, to obtain $\hat{B}_{(k)}$.

6. Check the convergence. Let $\boldsymbol{\eta}_{(k)} = B_{(k)} \otimes A_{(k)}$, $\boldsymbol{\eta}_{(k-1)} = B_{(k-1)} \otimes A_{(k-1)}$. If $||\boldsymbol{\eta}_{(k)}\boldsymbol{\eta}_{(k)}^{\intercal} - \boldsymbol{\eta}_{(k-1)}\boldsymbol{\eta}_{(k-1)}^{\intercal}||$ is smaller than some pre-specified tolerance value, such as 10^{-6}, stop the iteration and set $\hat{A} = A_{(k)}$, $\hat{B} = B_{(k)}$; Otherwise, set $k := k + 1$ and go to Step 2.

2.3.4 Estimation of the CTFS dimension

We can adopt the modified BIC criterion introduced in Chapter 1 to estimate d and r. However, $RSS_{(d*,r*)}$, the residual sum of squares when $d = d_*$ and $r = r_*$ changes to

$$\sum_{j=1}^{n}\sum_{i=1}^{n}\rho(y_i, \hat{c}_j + \hat{a}_j^{\intercal}\hat{A}^{\intercal}\mathbf{x}_{ij}\hat{B}\hat{b}_j)\mathbf{K}_{h_{(d_*,r_*)}}(\text{vec}(\hat{A}^{\intercal}\mathbf{x}_{ij}\hat{B})). \qquad (2.3.5)$$

In this chapter, we also set $C_n = (.5\log(n) + .1n^{1/3})/2$ (Zhu, Miao and Peng, 2006). We do not consider the asymptotic properties for the quantile dimension folding because there exist no closed form formulas with explicit solutions for the quantile regression. Kong and Xia (2012) and Wu, Yu and Yu (2010) investigated the asymptotic properties of the quantile estimators for single index model. However, for multiple index model, investigating the asymptotic properties poses a challenge.

2.4 Estimation of the CQFS

In what follows, we work on estimation of the CQFS, a special case of the CTFS. From the aspect of dimension folding, the absolute loss function $\rho(\cdot)$ in the objective functions (2.3.3) and (2.3.4) will be $u[\tau - I(u < 0)]$, where u is the prediction error and $0 < \tau < 1$ is the τ-th quantile we are interested in. The selection of the bandwidth for estimating the CQFS can be done as in Yu and Jones (1998). Yu and Jones (1998) suggested the relationship between the optimal bandwidth for conditional mean regression and that for conditional quantile as:

$$h_\tau = h_{mean}[\tau(1-\tau)/\phi(\Phi^{-1}(\tau))]^{1/5},$$

where functions $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function and the cumulative distribution function of the standard normal distribution respectively. The bandwidth h_τ is the bandwidth for single-index quantile regression and h_{mean} is the optimal bandwidth for local linear smoothing estimator in single-index mean regression. For h_{mean}, we can choose $h_{mean} = g(d,r)n^{-1/(dr+4)}$ (Silverman, 1986), where $g(d,r) = \frac{4}{dr+2}^{1/(4+dr)}$ and dr is the dimension of the refined kernel function. The bandwidth h_τ is a function of τ, d and r. In our simulations, we consider $\tau = 0.5$ or $\tau = 0.75$ thus the coefficient $[\tau(1-\tau)/\phi(\Phi^{-1}(\tau))]^{1/5}$ equals to 0.9107643 and 0.8998619 respectively. And $g(d,r)$ equals to 1.059224, 1 or 0.9505789. Since the $g(d,r)[\tau(1-\tau)/\phi(\Phi^{-1}(\tau))]^{1/5}$ ranges from 0.8553906 to 0.9647034, we can simply set $h_{d,r} = n^{-1/(d\times r+4)}$.

2.5 Numerical Study

In the following models, we assume the error ϵ is independent of the predictor matrix $\mathbf{X}$. For $\mathbf{X}$, $\text{vec}(\mathbf{X}) \sim N_{pq}(0, \Sigma_{\text{vec}(\mathbf{X})})$ and $\Sigma_{\text{vec}(\mathbf{X})}$ is a $pq \times pq$ positive definite matrix with $(j_1, j_2)-$th entry $0.5^{|j_1 - j_2|}$. And suppose $e_1 = (1, 0, 0, 0)^\intercal$, $e_2 = (0, 1, 0, 0)^\intercal$ and $e_3 = (0, 0, 1, 0)^\intercal$. Each simulation is run 100 replications. We report accuracy of estimating CQFS as: mean ± standard deviation, and accuracy of dimensions estimation in percentages. And we use the same standardization approach for $\mathbf{X}$.

Example 2.6. *We consider the following nonlinear model with $p = q = 5$, $\beta_1 = (1, 1, 0, 0, 0)^\intercal$ and $\beta_2 = (0, 1, 0, 0, 0)^\intercal$.*

$$Y = (\beta_1^\intercal \mathbf{X} \beta_2)^{-1} + .2\epsilon.$$

This model will produce extreme values around the origin. And $\mathcal{S}_{M(Y|\circ\mathbf{X}\circ)} = \mathcal{S}_{Q_\tau(Y|\circ\mathbf{X}\circ)} = \mathcal{S}(\beta_2) \otimes \mathcal{S}(\beta_1)$ with $d = 1, r = 1$. We consider three distributions for ϵ : standard normal, and the heavy tailed distribution, t_1, t-distribution with one-degree of freedom. Since the expectation of t_1 and the variance of t_2 are not defined, we also consider t_3, t-distribution with three-degree of freedom for the error term ϵ. The respective results for $\epsilon \sim t_1$ and $\epsilon \sim t_3$ are reported in the parenthesis and square bracket. We apply our method with $\tau = 0.5$ and $\tau = 0.75$ and sample size of 200, 300, 400 and 600. The benchmark distance is 1.384.

From Table 2.1, we can see the accuracy increases as sample size increases and the accuracy decreases when error distribution switches from normal distribution to t_3 *and from* t_3 *to* t_1 *for all models. Since relatively* $\tau = 0.75$ *calculates extreme tails, the accuracy decreases from* $\tau = 0.5$ *to* $\tau = 0.75$*. The estimated percentage for correctly identifying dimensionality seems reasonable, which also depends on the accuracy of the estimation of the respective CTFS.*

Table 2.1: Example 2.6: Accuracy of estimates

Methods	n	Δ_f *OPG*	Δ_f *MAVE*	$(\hat{d},\hat{r})=(d,r)$	*Distribution of* ϵ
$\tau=0.5$	*200*	0.8253 ± 0.1754	0.1530 ± 0.1501	84%	z
		(0.9650 ± 0.1878)	(0.4084 ± 0.3435)	(82%)	t_1
		$[0.8302 \pm 0.1804]$	$[0.1603 \pm 0.0908]$	[82%]	t_3
	300	0.6902 ± 0.1615	0.0781 ± 0.0330	86%	z
		(0.7729 ± 0.1780)	(0.1497 ± 0.1213)	(85%)	t_1
		$[0.6995 \pm 0.1760]$	$[0.0858 \pm 0.0395]$	[86%]	t_3
	400	0.6265 ± 0.1463	0.0587 ± 0.0243	90%	z
		(0.7390 ± 0.1875)	(0.1019 ± 0.0627)	(88%)	t_1
		$[0.6635 \pm 0.1605]$	$[0.0858 \pm 0.0395]$	[89%]	t_3
	600	0.5020 ± 0.1111	0.0432 ± 0.0169	94%	z
		(0.5843 ± 0.1569)	(0.0676 ± 0.0338)	(91%)	t_1
		$[0.5492 \pm 0.1581]$	$[0.0449 \pm 0.0153]$	[93%]	t_3
$\tau=0.75$	*200*	1.0609 ± 0.1953	0.4723 ± 0.3883	80%	z
		(1.1861 ± 0.1498)	(0.9894 ± 0.3494)	(79%)	t_1
		$[1.0659 \pm 0.2074]$	$[0.5787 \pm 0.4246]$	[78%]	t_3
	300	0.9088 ± 0.2019	0.2046 ± 0.1992	83%	z
		(1.0506 ± 0.2031)	(0.5844 ± 0.4161)	(78%)	t_1
		$[0.9177 \pm 0.2256]$	$[0.2354 \pm 0.2852]$	[81%]	t_3
	400	0.8058 ± 0.1758	0.1051 ± 0.0552	86%	z
		(0.9652 ± 0.2156)	(0.4164 ± 0.3324)	(82%)	t_1
		$[0.8349 \pm 0.1923]$	$[0.1624 \pm 0.1106]$	[85%]	t_3
	600	0.6876 ± 0.1730	0.0680 ± 0.0278	90%	z
		(0.8069 ± 0.2018)	(0.1551 ± 0.1003)	(85%)	t_1
		$[0.7118 \pm 0.1758]$	$[0.0903 \pm 0.0389]$	[88%]	t_3

Example 2.7. *This example shows that for different quantiles, the bases of CQFS may be different. Let $p = q = 4$,*

$$Y = 1 + X_{11} + (1 + 1.5X_{21})\epsilon,$$

where the distribution of ϵ is standard normal or student's t distribution with degrees of freedom 1 or 3. For this example, the quantile folding subspaces are different from the central folding subspace. Define $\beta(\tau) = (1, 1.5F^{-1}(\tau), 0, 0)^{\intercal}$, for $0 < \tau < 1$, where F is the CDF for standard normal, t_1 or t_3 random variable. Note that, if ϵ is a t_1 random variable, the expected value of ϵ is undefined and thus the corresponding cental mean folding subspace is undefined as well. If ϵ is a standard normal or t_3 random variable, $\mathcal{S}_{Y|\circ\mathbf{X}\circ} = \mathcal{S}(e_1) \otimes \mathcal{S}(\beta(.5), e_2)$ with $d = 2, r = 1$. In contrast, $\mathcal{S}_{E(Y|\circ\mathbf{X}\circ)} = \mathcal{S}_{M(Y|\circ\mathbf{X}\circ)} = \mathcal{S}(e_1) \otimes \mathcal{S}(\beta(.5))$ and $\mathcal{S}_{Q_\tau(Y|\circ\mathbf{X}\circ)} = \mathcal{S}(e_1) \otimes \mathcal{S}(\beta(\tau))$ with $d = 1, r = 1$. If ϵ is a t_1 random variable, then $\mathcal{S}_{Y|\circ\mathbf{X}\circ} = \mathcal{S}(e_1) \otimes \mathcal{S}(\beta(.5), e_2)$ with $d = 2$ and $r = 1$. However, $\mathcal{S}_{M(Y|\circ\mathbf{X}\circ)} = \mathcal{S}(e_1) \otimes \mathcal{S}(\beta(.5))$ and $\mathcal{S}_{Q_\tau(Y|\circ\mathbf{X}\circ)} = \mathcal{S}(e_1) \otimes \mathcal{S}(\beta(\tau))$ with $d = 1$ and $r = 1$. The benchmark distance for $\mathcal{S}_{M(Y|\circ\mathbf{X}\circ)}$ or $\mathcal{S}_{Q_\tau(Y|\circ\mathbf{X}\circ)}$ is 1.367.

The accuracy in Table 2.2 increases as sample size increases and the accuracy decreases when error distribution switches from normal distribution to t_3 and from t_3 to t_1 or switches from $\tau = 0.5$ to $\tau = 0.75$ same as in Example 2.1. The estimated percentage for correctly identifying dimensions is in the same change pattern as in the CQFS estimation accuracy for all the models.

Table 2.2: Example 2.7: Accuracy of estimates

Methods	n	Δ_f *OPG*	Δ_f *MAVE*	$(\hat{d},\hat{r})=(d,r)$	*Distribution of* ϵ
$\tau = 0.5$	*200*	0.3229 ± 0.1147	0.3211 ± 0.1159	96%	z
		(0.4636 ± 0.1534)	(0.4623 ± 0.1849)	(88%)	t_1
		$[0.3577 \pm 0.1426]$	$[0.3623 \pm 0.1402]$	[95%]	t_3
	300	0.2428 ± 0.0975	0.2388 ± 0.0976	98%	z
		(0.3539 ± 0.1341)	(0.3503 ± 0.1439)	(89%)	t_1
		$[0.2770 \pm 0.1286]$	$[0.2813 \pm 0.1323]$	[94%]	t_3
	400	0.2172 ± 0.0884	0.2155 ± 0.0904	99%	z
		(0.2190 ± 0.0795)	(0.2177 ± 0.0802)	(91%)	t_1
		$[0.2102 \pm 0.0787]$	$[0.2074 \pm 0.0811]$	[98%]	t_3
	600	0.1657 ± 0.0711	0.1636 ± 0.0702	100%	z
		(0.2080 ± 0.0790)	(0.2071 ± 0.0724)	(92%)	t_1
		$[0.1827 \pm 0.0766]$	$[0.1825 \pm 0.0749]$	[98%]	t_3
$\tau = 0.75$	*200*	0.4427 ± 0.1627	0.4398 ± 0.1785	84%	z
		(0.7245 ± 0.2637)	(0.7183 ± 0.2540)	(73%)	t_1
		$[0.6019 \pm 0.2270]$	$[0.5968 \pm 0.2201]$	[84%]	t_3
	300	0.3591 ± 0.1303	0.3512 ± 0.1331	85%	z
		(0.5750 ± 0.2097)	(0.5740 ± 0.2228)	(75%)	t_1
		$[0.5843 \pm 0.2094]$	$[0.5575 \pm 0.2116]$	[84%]	t_3
	400	0.3131 ± 0.1209	0.3003 ± 0.1144	88%	z
		(0.5665 ± 0.2044)	(0.5590 ± 0.2070)	(80%)	t_1
		$[0.5191 \pm 0.1856]$	$[0.5053 \pm 0.1777]$	[86%]	t_3
	600	0.2934 ± 0.1115	0.2742 ± 0.1135	90%	z
		(0.4881 ± 0.1922)	(0.4757 ± 0.2004)	(84%)	t_1
		$[0.4670 \pm 0.1645]$	$[0.4473 \pm 0.1659]$	[89%]	t_3

Example 2.8. . *We also consider the Example 2 in Xue and Yin (2012),*

$$Y = X_{11}/(0.5 + (X_{21} + 1.5)^2) + 0.2 \times \epsilon, \tag{2.5.1}$$

whose corresponding vector-valued model can be found in Li (1991) and Xia et. al (2002). We set $p = q = 5$. *Here,* $\mathcal{S}_{Y|vec(\mathbf{X})} = \mathcal{S}_{E(Y|vec(\mathbf{X}))} = \mathcal{S}_{Y|\circ\mathbf{X}\circ} = \mathcal{S}_{E(Y|\circ\mathbf{X}\circ)} = \mathcal{S}_{M(Y|\circ\mathbf{X}\circ)} = \mathcal{S}_{Q_\tau(Y|\circ\mathbf{X}\circ)} = \mathcal{S}(e_1) \otimes \mathcal{S}(e_1, e_2)$. *Thus,* $d = 2$ *and* $r = 1$ *and the benchmark distance is 1.916.*

As shown in Table 2.3, for Example 2.8, the accuracy of estimating CQFS and percentage for correctly estimating the dimensions follow the same pattern as in the two examples before. As sample size increase, the accuracy and correct percentage increase as we expect.

Table 2.3: Example 2.8: Accuracy of estimates

Methods	n	Δ_f *OPG*	Δ_f *MAVE*	$(\hat{d}, \hat{r}) = (d, r)$	*Distribution of* ϵ
$\tau = 0.5$	*200*	0.8293 ± 0.3068	0.4353 ± 0.3139	74%	z
		(1.0489 ± 0.2712)	(0.8380 ± 0.3633)	(13%)	t_1
		$[0.9168 \pm 0.3116]$	$[0.4829 \pm 0.3192]$	[62%]	t_3
	300	0.6466 ± 0.2808	0.2413 ± 0.1986	90%	z
		(0.8573 ± 0.3322)	(0.5828 ± 0.3860)	(15%)	t_1
		$[0.7142 \pm 0.2772]$	$[0.3239 \pm 0.2695]$	[89%]	t_3
	400	0.5964 ± 0.2211	0.1844 ± 0.0658	92%	z
		(0.7756 ± 0.2819)	(0.4390 ± 0.3401)	(24%)	t_1
		$[0.6532 \pm 0.2203]$	$[0.2421 \pm 0.1596]$	[91%]	t_3
	600	0.4907 ± 0.1891	0.1271 ± 0.0396	100%	z
		(0.5931 ± 0.2303)	(0.2831 ± 0.1768)	(33%)	t_1
		$[0.5087 \pm 0.2203]$	$[0.1598 \pm 0.0621]$	[100%]	t_3
	800	0.3991 ± 0.1281	0.1107 ± 0.0335	100%	z
		(0.5377 ± 0.2152)	(0.2060 ± 0.0805)	(54%)	t_1
		$[0.4271 \pm 0.1537]$	$[0.1426 \pm 0.0677]$	[100%]	t_3
$\tau = 0.75$	*200*	0.8602 ± 0.3271	0.4971 ± 0.3130	79%	z
		(1.1154 ± 0.2764)	(1.0203 ± 0.3197)	(12%)	t_1
		$[0.8600 \pm 0.3541]$	$[0.6095 \pm 0.3768]$	[56%]	t_3
	300	0.6462 ± 0.2752	0.2960 ± 0.1977	94%	z
		(0.9678 ± 0.3072)	(0.8156 ± 0.3603)	(13%)	t_1
		$[0.7640 \pm 0.3043]$	$[0.4277 \pm 0.2545]$	[68%]	t_3
	400	0.5505 ± 0.2305	0.2160 ± 0.0682	100%	z
		(0.8690 ± 0.3009)	(0.6728 ± 0.3461)	(14%)	t_1
		$[0.6604 \pm 0.2692]$	$[0.3374 \pm 0.2155]$	[75%]	t_3
	600	0.4333 ± 0.1601	0.1619 ± 0.0528	100%	z
		(0.7177 ± 0.2840)	(0.4724 ± 0.2878)	(27%)	t_1
		$[0.5269 \pm 0.2323]$	$[0.2338 \pm 0.1672]$	[90%]	t_3
	800	0.3528 ± 0.1200	0.1370 ± 0.0330	100%	z
		(0.6759 ± 0.2727)	(0.3303 ± 0.1035)	(48%)	t_1
		$[0.4509 \pm 0.1872]$	$[0.1870 \pm 0.1175]$	[100%]	t_3

2.6 Application

In this section, we apply the estimation methods to the primary biliary cirrhosis (PBC) data set used in Chapter 1. We consider forging the central median dimension folding space of the repeatedly measured longitudinal data. Concentrating on the median arises that the median often has higher efficiency than the mean for skewed data and always has an easy interpretation. We adopt the primary biliary cirrhosis (PBC) data set used in Xue and Yin (2012) to illustrate our method. Same as in Xue and Yin (2012), we have a univariate response Y and a 3×4 matrix-valued predictor $\mathbf{X}$. The measurements of biliribin, albumin level and prothrombin time are recorded at time point 6 month, 1 year, 2 year and 3 year. And the sample size n is 187.

Literature that comments on quantile regression for longitudinal data includes Koenker (2004), Geraci and Bottai (2007) and Fu and Wang (2012), Koenker (2004) proposed a weighted composite quantile regression (Zou and Yuan, 2008 and Kai, Li and Zou, 2010) model with fixed effects and by combining L_1 shrinkage (Tibshirani, 1996). Koenker (2004) considered the penalized quantile regression model for longitudinal data. However, we consider the quantile regression for repeated measured longitudinal data from the aspect of dimension folding.

We firstly standardize X and apply the modified BIC criterion to estimate d and r of the central median folding space. We have that $\hat{d} = \hat{r} = 1$ same as for the central mean folding space. Applying the refined estimation method and transforming the estimated bases in $\mathbf{Z}$-scale back to the $\mathbf{X}$-scale, the estimated bases in X-scale are

$A = (\alpha_1, \alpha_2, \alpha_3)^\intercal = (0.12369641, -0.9905826, -0.05869675)^\intercal$ and $B = (\beta_1, \beta_2, \beta_3, \beta_4)^\intercal = (-0.2740056, -0.1680333, -0.2602606, -0.910467)^\intercal$. The estimated bases are quite close to those of the *central mean folding subspace*. The Frobenius norm between the two spaces is 0.1594707 comparing to the benchmark distance 1.349.

In order to test the significance of each coefficient in A and B, we compute their 95% bootstrap confidence intervals respectively with 1000 bootstrap samples. The confidence intervals in Table 2.4 indicate albumin level and time point year 3 are significant at 0.05 level. The positive relationship between albumin level and the transplant-free or survival time is consistent to the medical outcome (Shapiro, Smith and Schaffner 1979). In the contrast, for the central mean dimension folding subspace, Xue and Yin (2012) found biliribin, albumin level and time year 3 are significant variables. Because the albumin levels can suggest liver disease and albumin testing is used in a variety of settings to help diagnose liver disease, to monitor disease progression, and time point year 3 reflects the progressive nature of the disease, the result for the central median dimension folding subspace is plausible. A summary plot of the response versus the reduced predictor and predicted model based on 50% quantile smoothing spline (Koenker, Ng and Portnoy, 1994) is shown in Figure 2.1. To fit the median quantile prediction model, we implement the “R” function “qsreg” from package “fields” (Oh et al., 2004; Lee and Cox, 2010). The fitted smoothing spline for the central median folding space looks pretty good and is quite similar to the spline plot in Xue and Yin (2012).

Table 2.4: 95% bootstrap confidential interval

	lower bound	upper bound
α_1	-0.5319895	0.55129965
α_2	-0.9986458	-0.06310098
α_3	-0.9469642	0.96603348
β_1	-0.8437331	0.90092703
β_2	-0.8148785	0.89193525
β_3	-0.8299503	0.89907690
β_4	-0.8509049	-0.01474479

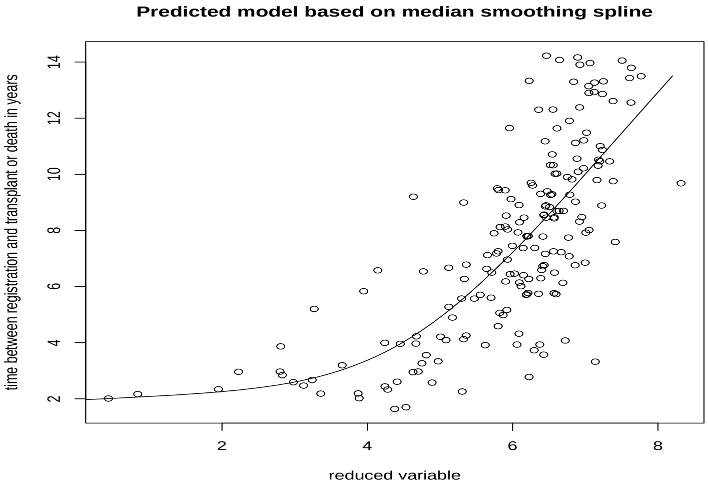

Figure 2.1: Summary plot and predicted model

2.7 Generalization to array-valued predictors

In this section, we briefly discuss how to extend the theory of dimension folding for a functional of conditional distribution of matrix-valued objects to array-valued predictors. Let $\mathbf{X} = \{X_{j_1 j_2 \cdots j_k} : j_1 = 1, \ldots, p_1, \ldots, j_k = 1, \ldots, p_k\}$ be a k-way random array of dimension $p_1 \times, \cdots, \times p_k$. Parallel the definition of central T dimension folding for matrix-valued predictor we generalize the definition to the array-valued predictor as follow:

Definition 2.2. *If there are matrices $\boldsymbol{\alpha}_i \in \mathbb{R}^{p_i \times d_i}$ $(d_i \leq p_i)$ for $i = 1, \ldots, k$ such that $T(\mathbf{x})$ depends on $\mathbf{x}$ only through $(\boldsymbol{\alpha}_k \otimes \cdots \otimes \boldsymbol{\alpha}_1)^{\intercal} vec(\mathbf{x})$, that is $T(\mathbf{x}) = T(\mathbf{x}^*)$ whenever $(\boldsymbol{\alpha}_k \otimes \cdots \otimes \boldsymbol{\alpha}_1)^{\intercal} vec(\mathbf{x}) = (\boldsymbol{\alpha}_k \otimes \cdots \otimes \boldsymbol{\alpha}_1)^{\intercal} vec(\mathbf{x}^*)$, then the column space of $\boldsymbol{\alpha}_k \otimes \cdots \otimes \boldsymbol{\alpha}_1$ is called a k-way dimension folding space for the functional T, or a k-way T dimension folding space.*

Definition 2.3. *If the intersection of all the k-way T dimension folding spaces as defined in Definition 2.2 itself is a k-way T dimension folding space then it is called the k-way central T dimension folding space for Y on the array-valued predictor $\mathbf{X}$. We write this space as $\mathcal{S}_{Y|\circ \mathbf{X}^{\circ k}}(T)$.*

When $k = 1$, $\mathcal{S}_{Y|\circ \mathbf{X}^{\circ 1}}(T)$ usually written as $\mathcal{S}_{Y|\circ \mathbf{X}^{\circ}}(T)$ for conciseness is the central T dimension folding space defined in Definition 2.1. Similar theory on $\mathcal{S}_{Y|\circ \mathbf{X}^{\circ k}}(T)$ and estimation methods on $\mathcal{S}_{Y|\circ \mathbf{X}^{\circ k}}(T)$ can be established straightforwardly from previous sections.

2.8 Conclusion

In this chapter, we establish the theory of sufficient dimension folding for a general functional, T, of the conditional distribution of matrix-/array-valued objects. We generalize the sufficient dimension folding and sufficient mean dimension folding spaces into the frame of sufficient functional T dimension folding space and propose the central variance and central k-th moment dimension folding spaces. We construct the relationship between local central T folding space and the central T folding space based on which we propose a class of local estimation methods. We also introduce a modified BIC criterion to estimate the dimensions of the proposed central T folding space. Besides, we consider the dimension folding for the quantile regression for the repeated measured longitudinal data. We illustrate our method by analyzing the primary biliary cirrhosis data used in Xue and Yin (2012). To evaluate the significance of individual variable, we adopt the bootstrap method to calculate the 95% confidential interval for each coefficient in the directions of the central T folding space. However, there is a challenge to investigate the asymptotic properties of the quantile dimension folding and we propose it as our further research.

2.9 Appendix

Proof of Lemma 2.1:

That $2 \Rightarrow 1$ is obviously. Only need to show $1 \Rightarrow 3$ and $3 \Rightarrow 2$.

$1 \Rightarrow 3$. Let t_1 and t_2 be any two matrices in $\mathbb{R}^{p\times q}$ such that $A^\intercal t_1 B = A^\intercal t_2 B$ and $\Sigma_A \in \mathbb{R}_+^{p\times p}$, $\Sigma_B \in \mathbb{R}_+^{q\times q}$ such that 1 holds. Then we have

$$\begin{aligned}
f(t_1) &= f[P^\intercal(\Sigma_A)t_1P(\Sigma_B)] \\
&= f[\Sigma_A A(A^\intercal\Sigma_A A)^{-1}(A^\intercal t_1 B)(B^\intercal\Sigma_B B)^{-1}B^\intercal\Sigma_B] \\
&= f[\Sigma_A A(A^\intercal\Sigma_A A)^{-1}(A^\intercal t_2 B)(B^\intercal\Sigma_B B)^{-1}B^\intercal\Sigma_B] \\
&= f[P^\intercal(\Sigma_A)t_2P(\Sigma_B)]
\end{aligned}$$

which tells us that f depends on t only through $A^\intercal tB$.

$3 \Rightarrow 2$. Let Σ_A be any element $\in \mathbb{R}_+^{p\times p}$ and Σ_B be any element $\in \mathbb{R}_+^{q\times q}$. Take $t_1 = t$ and $t_2 = P^\intercal(\Sigma_A)tP(\Sigma_B)$. Since $P(\Sigma_A)$ and $P(\Sigma_B)$ are projections onto $\mathcal{S}(A)$ and

$\mathcal{S}(B)$ accordingly, then $A = P(\Sigma_A)A$ and $B = P(\Sigma_B)B$. And further

$$\begin{aligned} A^{\intercal}t_1B = A^{\intercal}tB &= [P(\Sigma_A)A]^{\intercal}tP(\Sigma_B)B \\ &= A^{\intercal}P^{\intercal}(\Sigma_A)tP(\Sigma_B)B \\ &= A^{\intercal}t_2B \end{aligned}$$

Since f is a function of $A^{\intercal}tB$, $f(t) = f(P^{\intercal}(\Sigma_A)tP(\Sigma_B))$.

Proof of Proposition 2.1: Suppose $\boldsymbol{\alpha}_L$ is a $p \times d$ matrix whose columns form a basis of $S_{Y|\circ\mathbf{Z}}(T)$ and $\boldsymbol{\beta}_R$ is a $q \times r$ matrix whose columns form a basis of $S_{Y|\mathbf{Z}\circ}(T)$.

Since $\mathbf{Z} = A^{\intercal}\mathbf{X}B$ and A and B are full rank, by lemma 2.1

$$\begin{aligned} T_{\mathbf{X}}(\mathbf{X}) = T_{\mathbf{Z}}(\mathbf{Z}) &= T_{\mathbf{Z}}(P^{\intercal}(\Sigma_{\boldsymbol{\alpha}_L})\mathbf{Z}P(\Sigma_{\boldsymbol{\beta}_R})) \\ &= T_{\mathbf{X}}(P^{\intercal}(\Sigma_{\boldsymbol{\alpha}_L})A^{\intercal}\mathbf{X}BP(\Sigma_{\boldsymbol{\beta}_R})) \\ &= T_{\mathbf{X}}([AP(\Sigma_{\boldsymbol{\alpha}_L})]^{\intercal}\mathbf{X}[BP(\Sigma_{\boldsymbol{\beta}_R})]) \end{aligned}$$

$\Rightarrow$ $AP(\Sigma_{\boldsymbol{\alpha}_L})$ is a T-left dimension folding space for $Y|\mathbf{X}$ and $BP(\Sigma_{\boldsymbol{\beta}_R})$ is a T-right dimension folding space for $Y|\mathbf{X}$.

$\Rightarrow \mathcal{S}(A_0) \subseteq \mathcal{S}(AP(\Sigma_{\boldsymbol{\alpha}_L})) = A\mathcal{S}(P(\Sigma_{\boldsymbol{\alpha}_L}))$ and $\mathcal{S}(B_0) \subseteq \mathcal{S}(BP(\Sigma_{\boldsymbol{\beta}_R})) = B\mathcal{S}(P(\Sigma_{\boldsymbol{\beta}_R}))$

$\Rightarrow \mathcal{S}_{Y|\circ\mathbf{X}\circ}(T) \subseteq (B \otimes A)\mathcal{S}_{Y|\circ\mathbf{Z}\circ}(T)$.

And,

$$\begin{aligned} T_{\mathbf{Z}}(\mathbf{Z}) = T_{\mathbf{X}}(\mathbf{X}) &= T_{\mathbf{Z}}(P^{\intercal}(\Sigma_{A_0})\mathbf{X}P(\Sigma_{B_0})) \\ &= T_{\mathbf{Z}}(P^{\intercal}(\Sigma_{A_0})A^{\text{-}\intercal}A^{\intercal}\mathbf{X}BB^{\text{-}1}P(\Sigma_{B_0})) \\ &= T_{\mathbf{Z}}([A^{\text{-}1}P(\Sigma_{A_0})]^{\intercal}\mathbf{Z}[B^{\text{-}1}P(\Sigma_{B_0})]) \end{aligned}$$

$\Rightarrow A^{\text{-}1}P(\Sigma_{A_0})$ and $B^{\text{-}1}P(\Sigma_{B_0})$ are a T-left- or right– dimension folding space for $Y|\mathbf{Z}$.

$\Rightarrow \mathcal{S}(\boldsymbol{\alpha}_L) \subseteq \mathcal{S}(A^{\text{-}1}P(\Sigma_{A_0})) = A^{\text{-}1}\mathcal{S}(P(\Sigma_{A_0}))$ and $\mathcal{S}(\boldsymbol{\beta}) \subseteq \mathcal{S}(B^{\text{-}1}P(\Sigma_{B_0})) = B^{\text{-}1}\mathcal{S}(P(\Sigma_{B_0}))$

$\Rightarrow \mathcal{S}_{Y|\circ\mathbf{X}\circ}(T) \subseteq (B \otimes A)\mathcal{S}_{Y|\circ\mathbf{Z}\circ}(T)$

$\Rightarrow \mathcal{S}_{Y|\circ\mathbf{Z}\circ}(T) = (B^{\text{-}1} \otimes A^{\text{-}1})\mathcal{S}_{Y|\circ\mathbf{X}\circ}(T)$.

Proof of Proposition 2.2: We follow the proof of Proposition 2 in Yin and Li (2012). However, we work on a matrix predictor instead of a vector. We write a general form of $T_\Delta(\mathbf{x}) = T_\Delta(y|\mathbf{x})$ and $T_\Delta(\mathbf{x})$ may or may not depend on Y. For example, the local density $p_\Delta(y|\mathbf{x})$ depends on y but the local mean function $E_\Delta(y|\mathbf{x})$ does not depend on y. We use y in the form of T_Δ without confusion.

Define $\boldsymbol{\beta} = B \otimes A$ and it is sufficient to show for any $\boldsymbol{\alpha} \in \mathbb{R}^{pq\times 1}$, $\boldsymbol{\alpha}^{\intercal}\boldsymbol{\beta} = \mathbf{0}$ iff $\boldsymbol{\alpha}^{\intercal}\frac{\partial}{\partial \text{vec}(\mathbf{x})}T_\Delta(y|\mathbf{x}) = 0$ for all $(\mathbf{x}, y) \in \Delta$. By the chain rule of differentiation, $\frac{\partial}{\partial \text{vec}(\mathbf{x})}T_\Delta(y|\mathbf{x}) = \boldsymbol{\beta}\frac{\partial}{\partial \text{vec}(\mathbf{s})}T_\Delta(y|\mathbf{s})$. Thus, $\boldsymbol{\alpha}^{\intercal}\boldsymbol{\beta} = 0$ implies that $\boldsymbol{\alpha}^{\intercal}\frac{\partial}{\partial \text{vec}(\mathbf{x})}T_\Delta(y|\mathbf{x}) = 0$ for all $(\mathbf{x}, y) \in \Delta$.

The other way can be showed by contradiction. Assume that there exists $\boldsymbol{\alpha}_0 \in \mathbb{R}^{pq}$ such that $\boldsymbol{\alpha}_0^{\intercal} \frac{\partial}{\partial \text{vec}(\mathbf{x})} T_\Delta(y|\mathbf{x}) = 0$ for all $(\mathbf{x}, y) \in \Delta$, but $\boldsymbol{\alpha}_0^{\intercal}\boldsymbol{\beta} \neq 0$. Then $\xi_1 = \boldsymbol{\beta}^{\intercal}\boldsymbol{\alpha}_0/||\boldsymbol{\beta}^{\intercal}\boldsymbol{\alpha}_0||$ is a nonzero $dr \times 1$ vector. Therefore, $\boldsymbol{\alpha}_0^{\intercal} \frac{\partial}{\partial \text{vec}(\mathbf{x})} T_\Delta(y|\mathbf{x}) = \boldsymbol{\beta} \frac{\partial}{\partial \text{vec}(\mathbf{s})} T_\Delta(y|\mathbf{s}) = 0$ implies that $\xi_1^{\intercal} \frac{\partial}{\partial \text{vec}(\mathbf{s})} T_\Delta(y|\mathbf{s}) = 0$. Then the directional derivative of T_Δ as a function of vec(s) along ξ_1 is always 0. Thus, $T_\Delta(y|\mathbf{s}) = T_\Delta(y;\mathbf{s})$ is a constant along ξ_1, which means $T_\Delta(y;\mathbf{s}+t\xi) = T_\Delta(y;\mathbf{s})$ for all $t \in \mathbb{R}$.

We can expand ξ_1 to form an orthogonal basis for $\mathbb{R}^{dr}$, say, $A_\xi = (\xi_1, \dots, \xi_{dr})$. Define $v = A_\xi^{\intercal} vec(\mathbf{s}) = (v_1, \dots v_{dr})^{\intercal}$, then $T_\Delta(y;\mathbf{s}) = T_\Delta(y; A_\xi v)$, and $\frac{\partial}{\partial v_1} T_\Delta(y; A_\xi v) = \xi_1^{\intercal} \frac{\partial}{\partial \text{vec}(\mathbf{s})} T_\Delta(y|\mathbf{s}) = 0$. Therefore, $T_\Delta(y; A_\xi v)$ does not depend on v_1. And thus $T_\Delta(y|\mathbf{s}) = T_\Delta(y;\mathbf{s}) = T_\Delta(y; A_\xi v)$ can be written as a function $g(y; v_2, \dots, v_{dr}) = g(y; \xi_2^{\intercal}\boldsymbol{\beta}^{\intercal}\mathbf{x}, \dots \xi_{dr}^{\intercal}\boldsymbol{\beta}^{\intercal}\mathbf{x})$. Then $\boldsymbol{\beta}\xi_2, \dots, \boldsymbol{\beta}\xi_{dr}$ is a local CTFS with structural dimension $dr-1$, which is contradicts to the local CTFS has dimension dr. The proof is completed.

Proof of Proposition 2.3: The proof can be easily extended from the proof of Proposition 3 in Yin and Li (2012). Here, we work on a matrix predictor instead of a vector predictor.

For $(\mathbf{x}, y) \in \Delta$, we have

$$f_\Delta(\mathbf{x}, y) := f((\mathbf{x}, y)|(\mathbf{X}, Y) \in \Delta) = \frac{f(\mathbf{x}, y)}{\mathbb{P}_\Delta}, \tag{2.9.1}$$

where $\mathbb{P}_\Delta$ is the probability $(\mathbf{X}, Y) \in \Delta = \int\int_\Delta f(\mathbf{x}, y) d\mathbf{x} dy$. Let $f_{\Delta_y}(\mathbf{x}) = \int_{\Delta_y} f(\mathbf{x}, y) dy$, then

$$f_\Delta(\mathbf{x}) := f(\mathbf{x}|(\mathbf{X}, Y) \in \Delta) = \frac{f_{\Delta_y}(\mathbf{x}, y)}{\mathbb{P}_\Delta}. \tag{2.9.2}$$

Define $w = w(y) = 1$ if $y \in \Delta_y$, and 0, otherwise. We then have

$$f(w = 1|\mathbf{x}) := E_{(Y|\mathbf{X}=\mathbf{x})} w(Y) = \mathbb{P}(w(y) = 1|\mathbf{x}),$$

which means that

$$f(w = 1|\mathbf{x}) = \int_{\Omega_Y} w(y) f(y|\mathbf{x}) dy = \int_{\Delta_y} f(y|\mathbf{x}) dy = f_{\Delta_y}(\mathbf{x})/f(\mathbf{x}). \tag{2.9.3}$$

Combining (2.9.1),(2.9.2) and (2.9.3), we have

$$f_\Delta(y|\mathbf{x}) = \frac{f_\Delta(\mathbf{x}, y)}{f_\Delta(\mathbf{x})} = \frac{f(\mathbf{x}, y)}{f_{\Delta_y}(\mathbf{x})} = \frac{f(y|\mathbf{x})}{f(w = 1|\mathbf{x})},$$

or,

$$f(y|\mathbf{x}) = f_\Delta(y|\mathbf{x}) f(w = 1|\mathbf{x}). \tag{2.9.4}$$

Suppose A_0 and B_0 is a basis matrix of $\mathcal{S}_{Y|\circ\mathbf{X}}$ and $\mathcal{S}_{Y|\mathbf{X}\circ}$, respectively. Then $F_{Y|\mathbf{X}}(\cdot|\mathbf{x}) = F_{Y|\mathbf{X}}(A_0^\intercal \mathbf{x} B_0)$ by the definition of CFS. Thus, $f(y|\mathbf{x}) = f(y|A_0^\intercal \mathbf{x} B_0)$, and $f(w = 1|\mathbf{x}) = f(w = 1|A_0^\intercal \mathbf{x} B_0)$, since $w = 1$ is a function of y. By (2.9.4), we have for any $(\mathbf{x}, y) \in \Delta$,

$f_\Delta(y|\mathbf{x}) = f_\Delta(y|A_0^\intercal \mathbf{x} B_0)$. Therefore, $F_{\Delta(Y|\mathbf{X})}(\cdot|\mathbf{x}) = F_{\Delta(Y|\mathbf{X})}(\cdot|A_0^\intercal \mathbf{x} B_0)$. By the definition of $T_\Delta(\mathbf{x})$, the subspace spanned by the columns of A_0 forms a central T-left dimension folding subspace, and B_0 forms a central T-right dimension folding subspace. And by the definition of local CTFS, we have that $\mathcal{S}_{Y|\mathbf{X}}(T_\Delta) \subseteq \mathcal{S}_{Y|\mathbf{X}}$.

If $\Delta_y = \Omega_y$, then $w = 1$ for all y. For any $\mathbf{x} \in \Omega_\mathbf{x}$, there exist $\Delta_\mathbf{x}$ such that $\mathbf{x} \in \Delta_\mathbf{x} \subset \Omega_\mathbf{x}$. By (2.9.4), we have $f_\Delta(y|\mathbf{x}) = f(y|\mathbf{x})$, and we have $T_\Delta(\mathbf{x}) = T(\mathbf{x})$. From Proposition 2.2, we have:

$$\mathcal{S}\{\mathcal{S}_{Y|\circ\mathbf{X}\circ}(T_\Delta), \Delta = \Delta_\mathbf{x} \times \Omega_Y \ \textit{for all}\ \mathbf{x} \in \Omega_\mathbf{X}\} = \mathcal{S}_{Y|\circ\mathbf{X}\circ}(T).$$

Bibliography

[1] Čížek and Härdle. Robust estimation of dimension reduction space. *Computational Statistics and Data Analysis* **51**, 545–555, 1994.

[2] R. D. Cook. Regression Graphics. *Wiley, New York.* 1998.

[3] J. Fan and I. Gijbels. Local Polynomial Modelling and its Applications. *Chapman & Hall, London.* 1996.

[4] L. Fu and Y. G. Wang. Quantile regression for longitudinal data with a working correlation model. *Computational Statistics and Data Analysis* **56**, 2526–2538, 2012.

[5] M. Geraci and M. Bottai. Quantile regression for longitudinal data using the asymmertic Laplace distribution. *Biostatistics* **8**, 140–154, 2007.

[6] P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Ststistics* **35**, 73–101, 1964.

[7] J. S. Lee and D. D. Cox. Robust smoothing: smoothing parameter selection and applications to fluorescence spectroscopy. *Computational Statistics and Data Analysis***54**, 3131-3143, 2010.

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

[10] B. Li, H. Zha and F. Chiaromonte. Contour Regression: a general approach to dimension reduction. *The Annals of Statistics* **33**, 1580–1616, 2005.

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

[12] B. Kai, Z. Li and H. Zou. Local composite quantile regression smoothing: an efficient and safe alternative to local polynomial regression. *Journal of Royal Statistical Society, Series B* **72**, 49–69, 2010.

[13] R. Koenker. Quantile regression for longitudinal data. *Journal of Multivariate Analysis* **91**, 74–89, 2004.

[14] R. Koenker. Quantile regression. *Cambridge University Press, New York*, 2005.

[15] R. Koenker, P. Ng and S. Portnoy. Quantile smoothing splines. *Biometrika* **81**, 673–680, 1994.

[16] E. Kong and Y. Xia. A single-index quantile regression model and its estimation. *Econometric Theory* **28**, 730–768, 2012.

[17] N. Lu and D. L. Zimmerman. The likelohood ratio test for a separable covariance matrix. *Statistics & Probability Letters* **73**, 449-457, 2005.

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

[19] A. Roy and R. Khattree. On implementation of a test for Kronecker product covariance structure for multivariate repeated measures data. *Statistical Methodology* **2**, 297-306, 2005.

[20] H. Oh, D. Nychka, T. Brown and P. Charbonneau. Period analysis of variable stars by robust smoothing. *Journal of the Royal Statistical Society, Series C*, **53**, 1530, 2004

[21] R. J. Schott. Matrix analysis for statistics. *Wiley, New York*, 1997.

[22] J. M. Shapiro, H. Smith and F. Schaffner. Serum bilirubin: a prognostic factor in primary biliary cirrhosis. *Gut* **20**, 137-140, 1979.

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

[24] M. S. Srivastava, T. von Rosen and D. von Rosen. Models with a Kronecker product covariance structure: estimation and testing. *Mathematical Methods of Statistics* **17**, 357-370, 2008.

[25] M. S. Srivastava, T. von Rosen and D. von Rosen. Estimation and testing in general multivariate linear models with Kronecker product covariance structure. *The Indian Journal of Statistics* **71-A**, part 2, 137-163, 2009.

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

[27] Q. Wang and X. Yin. A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse MAVE. *Computational Statistics and Data Analysis* **52**, 4512–4520, 2008.

[28] Z. Wu, K. Yu and Y. Yu. Single-index quantile regression. *Journal of Multivariate Analysis* **101**, 1607-1621, 2010.

[29] 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.

[30] Y. Xue and X. Yin. Sufficient dimension folding for regression mean function. *manuscript*, 2012.

[31] X. Yin and B. Li. Dimension reduction for a functional of conditional distribution. *manuscript*, 2012.

[32] 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**, 1773–1757, 2008.

[33] K. Yu, M. C. Jones. Local linear quantile regression. *Journal of the American Statistical Association* **93**, 228–237, 1998.

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

[35] H. Zou and M. Yuan. Composite quantile regression and the oracle model selection theory. *The Annals of Statistics* **36**, 1108–1126, 2008.

Chapter 3

Sufficient dimension folding based on an ensemble of minimum average variance estimators

3.1 Introduction

To remedy the defect that *folded-MAVE* method can not recover directions outside the CMFS, we adopt the idea of ensemble (Yin and Li, 2011) to cover the CFS by the unison of several central mean dimension folding subspaces. We consider a general family $\mathfrak{F}$ of functions of Y and, for each $f \in \mathfrak{F}$, $\mathcal{S}_{E[f(Y)|\circ\mathbf{X}\circ]}$ is the CMFS for the conditional mean $E[f(Y)|\mathbf{X}]$. If the subspace spanned by the collection of subspaces $\{\mathcal{S}_{E[f(Y)|\circ\mathbf{X}\circ]} : f \in \mathfrak{F}\}$ equals to the CFS, $\mathfrak{F}$ is said to be characterizing the CFS. The functions $f_1, \ldots, f_m$ can be randomly sampled from $\mathfrak{F}$ according to a probability measure on $\mathfrak{F}$ and the assembly of the CMFS $\mathcal{S}_{E[f_l(Y)|\circ X \circ]}, l = 1, \ldots, m$ recovers the CFS.

In this chapter, we construct iterative algorithms named as folded-OPG ensemble and *folded-MAVE* ensemble and show the ensemble estimators can exhaustively recover the CFS. The properties of folded ensemble methods can be investigated based on the result in Yin and Li (2011) since $Y \perp\!\!\!\perp \mathbf{X}|A^{\intercal}\mathbf{X}B$ or $Y \perp\!\!\!\perp E(Y|\mathbf{X})|A^{\intercal}\mathbf{X}B$ are equivalent to $Y \perp\!\!\!\perp \text{vec}(\mathbf{X})|(B \otimes A)^{\intercal}\text{vec}(\mathbf{X})$ or $Y \perp\!\!\!\perp E[Y|\text{vec}(\mathbf{X}))|(B \otimes A)^{\intercal}\text{vec}(\mathbf{X})$, where $\text{vec}(\cdot)$ is an operation that stacks a matrix into a vector column by column. Nevertheless, to exhaustively estimate the CFS, an iterative procedure is indispensable in the folded ensemble methods. We mainly focus on the ensemble algorithms on matrix-valued predictor and a brief extension of the ensemble idea to the array-valued predictor is followed.

In Section 3.2, we extend the theory on characterizing central subspace to the central folding subspace. In Section 3.3, we introduce the algorithms of *folded-MAVE* ensemble and its variation, folded-OPG ensemble, to estimate the central folding subspace. In Section 3.4, we introduce a cross validation criterion for estimating the number of directions in CFS. We then summarize our future work.

3.2 Estimating the CFS through CMFS

Characterizing the central folding subspace is carried out by combining the dimension folding subspaces for $E[f(Y)|\mathbf{X}]$ in unity, as long as the samplings of function f are sufficiently dense. Suppose $\mathbf{X} \in \mathbb{R}^{p\times q}$ is a $p \times q$ random matrix with its support $\Omega_{\mathbf{X}}$ and Y is a s-dimensional random vector with its support Ω_Y. A transformation f can project a vector-valued Y to the scalar field $\mathbb{F}$. Let $\mathfrak{F}$ be a family of functions $f : \Omega_Y \to \mathbb{F}$, where $\mathbb{F}$ can be the set of complex numbers $\mathbb{C}$ or real numbers $\mathbb{R}$. And let $P_{\mathcal{S}}$ be the projection onto a space $\mathcal{S}$. Then for the $\mathcal{S}_{E[f(Y)|\circ X \circ]}$, the CMFS of conditional mean $E[f(Y)|\circ X\circ]$, we have

$$E[f(Y)| \circ \mathbf{X}\circ] = E[f(Y)|P_{\mathcal{S}_{E[f(Y)|\circ\mathbf{X}\circ]}}\text{vec}(\mathbf{X})]. \tag{3.2.1}$$

And for the CFS, $\mathcal{S}_{Y|\circ\mathbf{X}\circ}$, we have

$$Y \perp\!\!\!\perp \mathbf{X}|P_{\mathcal{S}_{Y|\circ\mathbf{X}\circ}}\text{vec}(\mathbf{X}). \tag{3.2.2}$$

Definition 3.1. *Let $\mathfrak{F}$ be a family of measurable $\mathbb{F}$-valued functions defined on Ω_Y. If*

$$\{\mathcal{S}_{E[f(Y)|\circ\mathbf{X}\circ]} : f \in \mathfrak{F}\} = \mathcal{S}_{Y|\circ\mathbf{X}\circ}, \tag{3.2.3}$$

the family $\mathfrak{F}$ is said to characterize the central folding subspace.

Let F_Y denote the distribution of Y, and let $L_1(F_Y)$ be the class of functions $f(Y)$ such that $E|f(Y)| < \infty$, together with the norm $E|f(Y)|$. If $f \in L_1(F_Y)$, $E[f(Y)|\mathbf{X}]$ is finite. Let $L_2(F_Y)$ be the class of functions $f(Y)$ with finite variances, together with the inner product $\langle f_1, f_2\rangle = E[f_1(Y)f_2(Y)]$. We denote the subspace on the left hand side of (3.2.3) by $\mathcal{S}(\mathfrak{F})$. We have the following lemma which can be proved similarly to that of Yin and Li (2011).

Lemma 3.1. *Suppose that $\mathfrak{F} \subseteq L_1(F_Y)$, then,*

1. $\mathcal{S}(\mathfrak{F}) \subseteq \mathcal{S}_{Y|\circ\mathbf{X}\circ}$.
2. *If (3.2.1) being satisfied for all $f \in \mathfrak{F}$ implies (3.2.2), then $\mathcal{S}_{Y|\circ\mathbf{X}\circ} \subseteq \mathcal{S}(\mathfrak{F})$.*

Let $\mathfrak{B}$ be the family of measurable indicator functions of Y. That is, $\mathfrak{B} = \{I_B : B \text{ is a Borel set in } \Omega_Y\}$. Then $\mathfrak{B} \subseteq L_2(F_Y)$. The following theorem grantees that several families can be adopt to characterize the CFS.

Theorem 3.1. *If $\mathfrak{F}$ is a subset of $L_2(F_Y)$ that is dense in $\mathfrak{B}$, then $\mathfrak{F}$ characterizes the central dimension folding subspace.*

Example 3.1. (**Polynomials**) Let $\mathfrak{F} = \{Y^t : t = 1, 2, \ldots\}$. Then $\mathcal{S}_{E[f(Y)|\circ\mathbf{X}\circ]} = \mathcal{S}_{E(Y^t|\circ\mathbf{X}\circ)}$. If the conditional moment generating function $E(e^{tY}|\mathbf{X})$ is finite in an open interval that contains 0, then $\mathfrak{F}$ is dense in $L_2(F_Y)$, and hence characterizes $\mathcal{S}_{Y|\circ\mathbf{X}\circ}$.

Example 3.2. (**Kernel density**) Let $b > 0$ and H be a symmetric probability density function defined on $\mathbb{R}$. Let $\mathfrak{F} = \{b^{-1}H[(y-t)/b] : t \in \mathbb{R}, b \in \mathbb{R}^+\}$, which is dense in $L_2(F_Y)$ when H is the Normal density and thus $\mathcal{S}_{Y|\circ\mathbf{X}\circ}$ can be recovered by estimating $\mathcal{S}_{E[f(Y)|\circ\mathbf{X}\circ]}$ for $f \in \mathfrak{F}$.

Example 3.3. (**Slices**) Let $\mathfrak{F} = \{I_{(-\infty,t)}(y) : t \in \mathbb{R}\}$. Then $\mathfrak{F}$ is dense in $\mathfrak{B}$ and can be used to characterize $\mathcal{S}_{E[f(Y)|\circ\mathbf{X}\circ]}$.

Example 3.4. (**Box-Cox transformations** (Box and Cox, 1964)) Let Y be a nonnegative random variable, and consider the family of transformations

$$f_t(y) = \begin{cases} \frac{y^t-1}{t} & t \neq 0 \\ \log(y) & t = 0 \end{cases}, \tag{3.2.4}$$

which contains the family in Example 3.1 and thus it characterizes the CFS.

Example 3.5. (**Characteristic function**) Let $\mathfrak{F} = \{e^{\iota t y} : t \in \mathbb{R}\}$, where $\iota = \sqrt{-1}$. Note that $E(e^{\iota t Y}|\mathbf{X})$ is simply the conditional characteristic function of $Y|\mathbf{X}$ and this family is dense in $L_2(F_Y)$.

Theorem 3.2 below whose proof follow the proof of Theorem 2.2 in Yin and Li (2011), demonstrates that, with probability 1, the CFS can be characterized by a finite number of functions in a characterizing family.

Let $B = (\beta_1, ..., \beta_{dr})$ be an orthogonal basis for the central folding subspace, $\mathcal{S}_{Y|\circ\mathbf{X}\circ}$, whose dimensions are d and r. We will randomly sample $T_1, \ldots, T_m$ from Ω_T and assume that these random elements are defined on a measurable space $(\Omega, \mathcal{A})$. Then Ω_T is interpreted as the range of the mapping $T_i : \Omega \to \Omega_T$ and we denote a generic member of Ω by ω.

Theorem 3.2. *Suppose that $\mathfrak{F}$ characterizes the central folding subspace, $T_1, T_2, \ldots$ is an i.i.d. sequence of random variables supported on Ω_T and, for each integer m, $B_0(T_1, ..., T_m)$ is an orthogonal basis matrix of $\{\mathcal{S}_{E[f_{T_i}(Y)|\circ\mathbf{X}\circ]} : i = 1, \ldots, m\}$. Then the* *following event has probability 1:*

$$\{\omega \in \Omega : \textit{there is an integer } m_0(\omega) \textit{ such that},$$

$$\textit{for all } m \geq m_0(\omega), \quad \mathrm{Span}(B_0(T_1(\omega), \ldots, T_m(\omega)) = \mathrm{Span}(B)\}.$$

3.3 Estimate the CFS through *folded-MAVE*: Algorithm

The *folded-MAVE* can exhaustively estimate directions in the CMFS only. To remedy this deficiency, we can adopt the idea of sliced regression (SR, Wang and Xia, 2008) to enable the *folded-MAVE* to recover directions in the CFS. Let $-\infty = s_0 < s_1, \ldots, < s_H = +\infty$ be the pre-selected grid point to slice the response, and let $v_k = I(s_{(k-1)} < Y < s_k)$, where $I(\cdot)$ is the indicator function, then the CMFS of $(v_1, \ldots, v_H)^\intercal$ coincides with the CFS of Y.

We can synthesize the *folded-MAVE* ensemble which extends the estimation of CMFS to the estimation of CFS in more general conditions. We call it *folded-MAVE* ensemble. folded-OPG ensemble as a variation of *folded-MAVE* ensemble is also considered.

3.3.1 Folded MAVE ensemble

Suppose $\mathbf{X} \in \mathbb{R}^{p\times q}$ defined on $\Omega_{\mathbf{X}}$, $Y \in \mathbb{R}^{s}$ defined on Ω_Y. We only consider the parametric characterizing family $\mathfrak{F}$ in this chapter. That is $\mathfrak{F}$ is of the form $\{f_t : t \in \Omega_T\}$ where Ω_T is a subset of a Euclidean space. Let $\mathfrak{F} = \{I_{(-\infty,t)}(Y) : t \in \mathbb{R}\}$, then for $f \in \mathfrak{F}$, the unison of estimations of $\mathcal{S}_{E(f(Y)|\circ\mathbf{X}\circ)}$ is $\mathcal{S}_{Y|\circ\mathbf{X}\circ}$. Thus, our previous modification of *folded-MAVE* based on SR is indeed a special case of the *folded-MAVE* ensemble.

Let $f_t(y) = f_t(y,1)+if_t(y,2)$, where $f_t(y,1)$ and $f_t(y,2)$ are the real and imaginary parts of $f_t(y)$, and let T be a random vector defined on Ω_T with distribution F_T. The population level objective function can be constructed by applying the *folded-MAVE* procedure to the transformed response $f_t(Y)$ and integrating with respect to the distribution F_T. It is formularized as:

$$\sum_{l=1}^{2}\int_{\Omega_T\times\Omega_{\mathbf{X}}} E\{[f_t(Y,l)-c_l(\mathbf{x})-a_l^{\intercal}(\mathbf{x})A^{\intercal}(\mathbf{X}-\mathbf{x})Bb_l(\mathbf{x})]^2K_h[(B\otimes A)^{\intercal}\text{vec}(\mathbf{X}-\mathbf{x})]\}dF_{\mathbf{X}}(\mathbf{x})dF_T(t). \tag{3.3.1}$$

We minimize the function (3.3.1) over all $c_l(\cdot) \in \mathbb{R}$, all $a_l(\cdot) \in \mathbb{R}^d$, $b_l(\cdot) \in \mathbb{R}^r$, all $p \times d$ constant matrices A and all $q \times r$ constant matrices B, $l = 1, 2$.

At the sample level, let $T_1, \ldots, T_m$ be an independent sample from F_T, we minimize the objective function

$$\sum_{l=1}^{2}\sum_{k=1}^{m}\sum_{j=1}^{n}\sum_{i=1}^{n}\rho_j w_{ij}(h)[f_{T_k}(Y_i,l) - c_{jk}(l) - a_{jk}(l)^{\intercal}A^{\intercal}(\mathbf{x}_i - \mathbf{x}_j)Bb_{jk}(l)]^2 \tag{3.3.2}$$

over scalars $\{c_{jk}(l) : j = 1, \ldots, n, k = 1, \ldots, m, l = 1, 2\}$, $\{a_{jk}(l) \in \mathbb{R}^d : j = 1, \ldots, n, k = 1, \ldots, m, l = 1, 2\}$, $\{b_{jk}(l) \in \mathbb{R}^r : j = 1, \ldots, n, k = 1, \ldots, m, l = 1, 2\}$, $p \times d$ matrices A and $q \times r$ matrices B. The coefficient ρ_j is the same as defined in (1.3.5).

For fixed $a_{jk}(l) \in \mathbb{R}^d$, $A \in \mathbb{R}^{p\times d}$, $B \in \mathbb{R}^{q\times r}$, minimizing (3.3.2) jointly over $c_{jk}(l), b_{jk}(l)$ for $j = 1, \ldots, n$, $k = 1, \ldots, m$, $l = 1, 2$, is equivalent to minimizing

$$\sum_{i=1}^{n} \rho_j w_{ij}(h)[f_{T_k}(Y_i, l) - c_{jk}(l) - a_{jk}(l)^\intercal A^\intercal(\mathbf{x}_i - \mathbf{x}_j)Bb_{jk}(l)]^2 \tag{3.3.3}$$

individually. More specifically, we suggest the following *folded-MAVE* ensemble procedures.

1. Generate $a_{jk}(l) \in \mathbb{R}^d$ from $N(0, 1)$. $A \in \mathbb{R}^{p\times d}$, $B \in \mathbb{R}^{q\times r}$ can be obtained from the folded OPG ensemble introduced later or generated from $N(0, 1)$.

2. For fixed $a_{jk}(l) \in \mathbb{R}^d$, $A \in \mathbb{R}^{p\times d}$, $B \in \mathbb{R}^{q\times r}$, the least-squares solution of (3.3.3), for each triplet (j, k, l), is

$$\begin{pmatrix} \hat{c}_{jk}(l) \\ \hat{b}_{jk}(l) \end{pmatrix} = \left[\sum_{i=1}^{n} \rho_j w_{ij}(h)\Delta_{ij}(a_{jk}(l), A, B)\Delta_{ij}^\intercal(a_{jk}(l), A, B)\right]^{\text{-1}} \times \left[\sum_{i=1}^{n} \rho_j w_{ij}(h)\Delta_{ij}(a_{jk}(l), A, B) f_{T_k}(Y_i, l)\right],$$

where $\Delta_{ij}(a_{jk}(l), A, B) = [1, (vec(\mathbf{x}_i - \mathbf{x}_j))^\intercal(B \otimes A)(I_r \otimes a_{jk}(l))]^\intercal$.

3. For fixed $c_{jk}(l)$, $b_{jk}(l) \in \mathbb{R}^r$, $A \in \mathbb{R}^{p\times d}$, $B \in \mathbb{R}^{q\times r}$, the least-squares solution of (3.3.3), for each triplet (j, k, l), is then

$$\hat{a}_{jk}(l) = \left[\sum_{i=1}^{n} \rho_j w_{ij}(h)\Delta_{ij}(b_{jk}(l), A, B)\Delta_{ij}^{\intercal}(b_{jk}(l), A, B)\right]^{\text{-1}} \\ \times \left[\sum_{i=1}^{n} \rho_j w_{ij}(h)\Delta_{ij}(b_{jk}(l), A, B)(f_{T_k}(Y_i, l) - c_{jk}(l))\right],$$

where $\Delta_{ij}(b_{jk}(l), A, B) = [(vec(\mathbf{x}_i - \mathbf{x}_j))^{\intercal}(B \otimes A)(b_{jk}(l) \otimes I_d)]^{\intercal}$.

4. For fixed $c_{jk}(l)$, $a_{jk}(l)$, $b_{jk}(l)$ and A, $j = 1, \ldots, n$, $k = 1, \ldots, m$, $l = 1, 2$, the minimization of (3.3.2) is also a least-squares problem. The solution is

$$vec(\hat{B}) = \left[\sum \rho_j w_{ij}(h)\Delta_{ij}(a_{jk}(l), b_{jk}(l), A)\Delta_{ij}^{\intercal}(a_{jk}(l), b_{jk}(l), A)\right]^{\text{-1}} \\ \times \left[\sum \rho_j w_{ij}(h)\Delta_{ij}(a_{jk}(l), b_{jk}(l), A)(f_{T_k l}(y_i, l) - c_{jk}(l))\right],$$

where $\Delta_{ij}(a_{jk}(l), b_{jk}(l), A) = [I_r \otimes ((\mathbf{x}_i - \mathbf{x}_j)^{\intercal}A)](b_{jk}(l) \otimes a_{jk}(l))$ and the summation is over

$$(i, j, k, l) \in \{1, \ldots, n\} \times \{1, \ldots, n\} \times \{1, \ldots, m\} \times \{1, 2\}.$$

5. For fixed $c_{jk}(l)$, $a_{jk}(l)$, $b_{jk}(l)$ and A, $j = 1, \ldots, n$, $k = 1, \ldots, m$, $l = 1, 2$,

$$vec(\hat{A}^{\intercal}) = \left[\sum \rho_j w_{ij}(h)\Delta_{ij}(a_{jk}(l), b_{jk}(l), B)\Delta_{ij}^{\intercal}(a_{jk}(l), b_{jk}(l), B)\right]^{\text{-1}} \\ \times \left[\sum \rho_j w_{ij}(h)\Delta_{ij}(a_{jk}(l), b_{jk}(l), B)(f_{T_k l}(y_i, l) - c_{jk}(l))\right],$$

where $\Delta_{ij}(a_{jk}(l), b_{jk}(l), B) = [((\mathbf{x}_i - \mathbf{x}_j)B) \otimes I_d](b_{jk}(l) \otimes a_{jk}(l))$ and the summation is also over

$$(i, j, k, l) \in \{1, \dots, n\} \times \{1, \dots, n\} \times \{1, \dots, m\} \times \{1, 2\}.$$

We iteratively estimate the parameters between the above five steps until convergence.

3.3.2 Folded OPG Ensemble

Let $\mathfrak{F}$, F_T, T_1, $\dots$, T_m and $w_{ij}(h)$ be as defined in the previous section. For folded OPG ensemble, we minimize the objective function for each j, k, l,

$$\sum_{i=1}^{n} w_{ij}(h)[f_{T_k}(y_i, l) - c - a^{\intercal}(\mathbf{x}_i - \mathbf{x}_j)b]^2 \tag{3.3.4}$$

over $(c, a, b) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^q$ for each $j = 1, \dots, n$, $k = 1, \dots, m$ and $l = 1, 2$.

The folded-OPG ensemble is summarized as

1. Generate $a_{jk}(l) \in \mathbb{R}^p$ from $N(0, 1)$.

2. For fixed $a_{jk}(l)$, $\hat{c}_{jk}(l)$ and $\hat{b}_{jk}(l)$ that minimize (3.3.4) are

$$\begin{pmatrix} \hat{c}_{jk}(l) \\ \hat{b}_{jk}(l) \end{pmatrix} = \left[\sum_{i=1}^{n} w_{ij}(h)\Delta_{ij}(a_{jk}(l))\Delta_{ij}^{\intercal}(a_{jk}(l))\right]^{\text{-1}} \left[\sum_{i=1}^{n} w_{ij}(h)\Delta_{ij}(a_{jk}(l)) f_{T_k}(y_i, l)\right],$$

   where $\Delta_{ij}(a_{jk}(l)) = (1, (vec(\mathbf{x}_i - \mathbf{x}_j))^{\intercal}(I_q \otimes a_{jk}(l)))^{\intercal}$.

3. For fixed $c_{jk}(l)$ and $b_{jk}(l)$, $\hat{a}_{jk}(l)$ that minimizes (3.3.4) is

$$\hat{a}_{jk}(l) = \left[\sum_{i=1}^{n} w_{ij}(h)\Delta_{ij}(b_{jk}(l))\Delta_{ij}^{\intercal}(b_{jk}(l))\right]^{\text{-1}} \left[\sum_{i=1}^{n} w_{ij}(h)\Delta_{ij}(b_{jk}(l))(f_{T_k}(Y_i, l) - c_{jk}(l))\right],$$

   where $\Delta_{ij}(b_{jk}(l)) = ((vec(\mathbf{x}_i - \mathbf{x}_j))^{\intercal}(b_{jk}(l) \otimes I_p))^{\intercal}$.

4. Compute the OPG matrices as:

$$\sum_{l=1}^{2}\sum_{k=1}^{m}\sum_{j=1}^{n} \rho_j \hat{a}_{jk}(l)\hat{a}_{jk}^{\intercal}(l), \tag{3.3.5}$$

   and

$$\sum_{l=1}^{2}\sum_{k=1}^{m}\sum_{j=1}^{n} \rho_j \hat{b}_{jk}(l)\hat{b}_{jk}^{\intercal}(l). \tag{3.3.6}$$

   The folded-OPG ensemble also involves iteratively estimating the $c_{jk}(l)$, $a_{jk}(l)$ and $b_{jk}(l)$. After convergence, A and B are the first d and r eigenvectors according to the d and r largest eigenvalues of (3.3.5) or (3.3.6) respectively.

3.4 Estimations of d, r and choices of $\mathfrak{F}$

In describing the foregoing algorithms we have assumed d and r, the dimension of the $\mathcal{S}_{Y|\circ\mathbf{X}}$ and $\mathcal{S}_{Y|\mathbf{X}\circ}$ respectively to be known. In practice those dimensions must also be estimated. We now propose a cross validation method to estimate d and r. Let $\hat{A}$ and $\hat{B}$ be the estimated bases of $\mathcal{S}_{Y|\circ\mathbf{X}}$ and $\mathcal{S}_{Y|\mathbf{X}\circ}$ respectively for a fixed working dimension d_0

and r_0. Then the leave-one-out fitted value of $f_{T_k}(Y_j, \ell)$, for $j = 1, \ldots, n$, $k = 1, \ldots, m$, and $\ell = 1, 2$, is

$$\hat{\mu}_{kj}(d_0, r_0, \ell) = \sum_{i \neq j} K_h[(\hat{B} \otimes \hat{A})^{\intercal} \mathrm{vec}(\mathbf{X}_i - \mathbf{X}_j)] f_{T_k}(Y_i, \ell) / \sum_{i \neq j} K_h[(\hat{B} \otimes \hat{A})^{\intercal} \mathrm{vec}(\mathbf{X}_i - \mathbf{X}_j)].$$

The corresponding cross validation value is

$$\mathrm{CV}(d_0, r_0) = \frac{1}{2mn} \sum_{\ell=1}^{2} \sum_{k=1}^{m} \sum_{j=1}^{n} [f_{T_k}(Y_j, \ell) - \hat{\mu}_{kj}(d_0, r_0, \ell)]^2.$$

To include the trivial case of $d_0 = 0$ or $r_0 = 0$, we define $\hat{\mu}_{kj}(0, r_0, \ell)$, $\hat{\mu}_{kj}(d_0, 0, \ell)$ and $\hat{\mu}_{kj}(0, 0, \ell)$ to be $(n-1)^{-1} \sum_{i \neq j} f_{T_k}(Y_j, \ell)$, so that $\mathrm{CV}(d_0, r_0)$ is defined for all $d_0 = 0, \ldots, p$ and $r_0 = 0, \ldots, q$. The structural dimensions d and r are estimated by

$$(\hat{d}, \hat{r}) = \mathrm{argmin}\{\mathrm{CV}(d_0, r_0) : d_0 = 0, \ldots, p;\ r_0 = 0, \ldots, q\}.$$

In this chapter we pay special attention to the family determined by the characteristic function, as discussed in Example 3.5. That is,

$$\mathfrak{F}_C = \{e^{\iota t^{\intercal} y} : t \in \mathbb{R}\}.$$

And we named the folded-OPG and *folded-MAVE* ensemble based on $\mathfrak{F}_C$ as folded-OPG-$\mathfrak{F}_C$ and folded-MAVE-$\mathfrak{F}_C$. An advantage of the family $\mathfrak{F}_C$ is that its members are bounded functions, and as such are relatively robust against the outliers in Y. Moreover, it requires virtually no condition on the distribution of Y. Also note that when t ranges over $\mathbb{R}^s$, the function $e^{\iota t^\intercal y}$ fully recovers the joint information of the random vector Y.

Having prepared the basics above, we will investigate asymptotic properties and run simulations for the performance of the proposed methods, and apply them to some data sets in the future.

Bibliography

[1] G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B* **26**, 211–252, 1964.

[2] R. D. Cook. Regression Graphics. *Wiley, New York.* 1998.

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

[4] 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.

[5] B. Li and S. Wang. On Directional Regression for Dimension Reduction. *Journal of the American Statistical Association* **102**, 997-1008, 2007.

[6] B. Li, H. Zha and F. Chiaromonte. Contour Regression: a general approach to dimension reduction. *The Annals of Statistics* **33**, 1580-1616, 2005.

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

[8] H. Wang and Y. Xia. Sliced regression for dimension reduction. *Journal of the American Statistical Association* **103**, 811-821, 2008.

[9] Y. Xue and X. Yin. Sufficient dimension folding for regression mean function. *manuscripts*, 2012.

[10] X. Yin and B. Li. Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *The Annals of Statistics* **39**, 3392-3416, 2011.

[11] 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**, 1773-1757, 2008.