Statistica Sinica **24** (2014), 463-492
doi:http://dx.doi.org/10.5705/ss.2012.138

DIMENSION FOLDING PCA AND PFC FOR MATRIX-VALUED PREDICTORS

Shanshan Ding and R. Dennis Cook

University of Minnesota

Abstract: Conventional dimension reduction methods deal mainly with simple data structure and are inappropriate for data with matrix-valued predictors. Li, Kim, and Altman (2010) proposed dimension folding methods that effectively improve major moment-based dimension reduction techniques for the more complex data structure. Their methods, however, are moment-based and rely on slicing the responses to gain information about the conditional distribution of $X|Y$. This can be inadequate when the number of slices is not chosen properly. We propose model-based dimension folding methods that can be treated as extensions of conventional principal components analysis (PCA) and principal fitted components (PFC). We refer to them as dimension folding PCA and dimension folding PFC. The proposed methods can simultaneously reduce a predictor's multiple dimensions and inherit asymptotic properties from maximum likelihood estimation. Dimension folding PFC gains further efficiency by effective use of the response information. Both methods can provide robust estimation and are computationally efficient. We demonstrated their advantages by both simulation and data analysis.

Key words and phrases: Central dimension folding subspace, central subspace, inverse regression, matrix normal distribution, sufficient dimension reduction.

1. Introduction

Dimension reduction methods are among the main techniques for studying high dimensional data. Typical dimension reduction analyses explore the dependence between a response $Y \in \mathbb{R}^1$ and a predictor vector $X \in \mathbb{R}^p$. Cook (1994, 1998) introduced sufficient dimension reduction (SDR), whose basic idea is to reduce the dimension of the predictor vector X by replacing it with its projection $\mathrm{P}_{\mathcal{S}}X$ onto a subspace $\mathcal{S}$ of the predictor space without loss of information on the conditional distribution of $Y|X$. This requirement can be stated as $Y \perp\!\!\!\perp X \mid \mathrm{P}_{\mathcal{S}}X$, where '$\perp\!\!\!\perp$' indicates independence. Under mild conditions, the intersection of all such dimension reduction subspaces $\mathcal{S} \subseteq \mathbb{R}^p$ is also a dimension reduction subspace and is called the central subspace, $\mathcal{S}_{Y|X}$.

Numerous dimension reduction methods have been developed that can be incorporated into this rationale under certain conditions. Sliced inverse regression (SIR; Li (1991)) and sliced average variance estimation (SAVE;

Cook and Weisberg (1991)) are two early techniques for dimension reduction. Since then, bootstrap dimension reduction (Ye and Weiss (2003)), inverse regression estimation (IRE; Cook and Ni (2005)), directional regression (DR; Li and Wang (2007)) and many other methods were developed to improve the estimation of $\mathcal{S}_{Y|X}$. Most of the proposed methods use the first two moments of $X|Y$ to perform estimation, so called moment-based methods. In contrast, Cook (2007) and Cook and Forzani (2008) presented model-based SDR techniques, including principal fitted components (PFC), that give the maximum likelihood estimators (MLE) of the central subspace based on normal inverse models of X on Y.

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

In face recognition and image analysis, certain unsupervised dimension reduction techniques were developed to deal with such data, based only on the marginal distribution of X. These methods include 2DPCA (Yang et al. (2004)), $(2D)^2$PCA (Zhang and Zhou (2005)), GLRAM (Ye (2005)), Unified PCA (Shan et al. (2008)), probabilistic higher-order PCA (Yu, Bi, and Ye (2011)), etc. Li, Kim, and Altman (2010) proposed supervised and moment-based dimension folding approaches that extend SIR, SAVE, and DR to data with matrix-valued predictors, in order to reduce the predictor's row and column dimensions simultaneously without loss of information on $Y|X$. The idea of dimension folding can be expressed as the condition: $Y \perp\!\!\!\perp X \mid \Gamma_2^T X \Gamma_1$ or, equivalently, $Y \perp\!\!\!\perp \mathrm{vec}(X) \mid (\Gamma_1 \otimes \Gamma_2)^T \mathrm{vec}(X)$, where $\Gamma_1 \in \mathbb{R}^{p_R \times d_R}$ and $\Gamma_2 \in \mathbb{R}^{p_L \times d_L}$ have the smallest column dimensions d_R and d_L ($d_R \le p_R$, $d_L \le p_L$), and '$\otimes$' stands for the Kronecker product. The subspace $\mathrm{Span}(\Gamma_1 \otimes \Gamma_2)$ or, equivalently, $\mathrm{Span}(\Gamma_1) \otimes \mathrm{Span}(\Gamma_2)$ is called the central dimension folding (CDF) subspace for $Y|X$, and denoted as $\mathcal{S}_{Y|\circ X\circ}$.

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

The remainder of this paper is organized as follows. In Section 2 we introduce dimension folding PCA and its estimation. Section 3 is devoted to the development of dimension folding PFC. Section 4 provides robustness results. Prediction methods are discussed in Section 5. Section 6 and 7 contain illustrations of the performance of our methods with simulation studies and data analysis. Discussion is given in Section 8.

2. Dimension Folding PCA

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

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

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

In their case, $X \in \mathbb{R}^p$ is the predictor vector, $\mu \in \mathbb{R}^p$ is the overall mean of X, $\Gamma \in \mathbb{R}^{p\times d}(d \le p)$ is a coefficient matrix with rank d, $\nu \in \mathbb{R}^d$ is a latent random

vector, and $\varepsilon \in \mathbb{R}^p$ is the random error. Additionally, ν and ε are assumed to be independent and both have standard multivariate normal distributions with zero means and identity covariance matrices. A random error with this structure is called an isotropic error. The identity covariance assumption for ν is not a restriction, since one can always combine a non-identity covariance matrix with Γ. Thus, the parameter Γ itself is not identified but Span(Γ) is identified.

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

2.1. Formulation of dimension folding PCA

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

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

where $X \in \mathbb{R}^{p_L \times p_R}$, $\Gamma_1 \in \mathbb{R}^{p_R \times d_R}$ $(d_R \leq p_R)$ and $\Gamma_2 \in \mathbb{R}^{p_L \times d_L}$ $(d_L \leq p_L)$ are semi-orthogonal matrices that reduce the column and row dimensions of X, $\mu \in \mathbb{R}^{p_L \times p_R}$ is the overall mean of X, and $\nu \in \mathbb{R}^{d_L \times d_R}$ is a latent matrix with mean zero. The random error ε is assumed to be independent of ν and have a matrix normal distribution. The matrix normal distribution is briefly reviewed in the appendix. As dimension folding PCA is a starting model, we simplify the error to be isotropic, so ε is $\mathbf{N}_{p_L \times p_R}(0_{p_L \times p_R}, I_{p_R}, I_{p_L})$. More general error structures will be discussed in the dimension folding PFC section. In (2.2), neither Γ_1 nor Γ_2 is identified: if Γ_1, Γ_2 and ν are replaced by $\Gamma_2 A_2$, $\Gamma_1 A_1$ and $A_2^{-1}\nu(A_1^T)^{-1}$, equation (2.2) remains the same, where A_1 and A_2 are any nonsingular matrices. Thus, the dimension folding PCA model depends on Γ_1 and Γ_2 only through their column spaces. Under (2.2), ν contains the coordinates of the centered conditional mean $\mathrm{E}(X|\nu) - \mu$ relative to Γ_1 and Γ_2, and the relationship $\mathrm{E}(X|\nu) - \mu = P_{\Gamma_2}[\mathrm{E}(X|\nu) - \mu]P_{\Gamma_1}$ holds. Therefore, the predictor's important row and column signals are preserved by Span(Γ_1) and Span(Γ_2).

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

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

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

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

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

Proposition 1. (a) *Under* (2.2)*, the distribution of* $\nu|X$ *is the same as the distribution of* $\nu|\Gamma_2^T X\Gamma_1$ *over all values of* X*;* (b) *under* (2.3)*, the distribution of* $\nu|vec(X)$ *is the same as the distribution of* $\nu|(\Gamma_1 \otimes \Gamma_2)^T vec(X)$ *for all values of* X*.*

Based on Proposition 1, $R(X) = \Gamma_2^T X\Gamma_1$ is a sufficient reduction (folding) satisfying $X \perp\!\!\!\perp \nu \mid \Gamma_2^T X\Gamma_1$. Since both Γ_1 and Γ_2 have the minimum column dimensions, $\mathrm{Span}(\Gamma_1 \otimes \Gamma_2)$ forms the central dimension folding subspace $\mathcal{S}_{\nu|\circ X\circ}$.

2.2. Estimation of dimension folding PCA

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

we can apply a fast and stable algorithm that uses three eigen-based iterations and provides connections to the conventional PCA model.

For an independent sample $\{X_i\}$, according to (S1.1) in the supplement file, the full log likelihood of (2.2) can be written as

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

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

Proposition 2. *Suppose that $X_i \in \mathbb{R}^{p_L\times p_R}$, $i = 1, \ldots, n$, are observed matrices. Let $(\hat{\Gamma}_1, \hat{\Gamma}_2, \hat{\nu}_1, \ldots, \hat{\nu}_n)$ be minimizers of*

$$\sum_{i=1}^{n}\operatorname{tr}[(X_i - G_2\omega_iG_1^T)^T(X_i - G_2\omega_iG_1^T)] \tag{2.5}$$

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

(i) *For fixed G_1, the columns of the minimizer $\hat{\Gamma}_2$ are given by the d_L eigenvectors of the matrix $\hat{\Sigma}_L = \sum_{i=1}^n X_i\mathrm{P}_1X_i^T/n$ corresponding to its d_L largest nonzero eigenvalues, where $\mathrm{P}_1 = G_1G_1^T$.*

(ii) *For fixed G_2, the columns of the minimizer $\hat{\Gamma}_1$ consist of the d_R eigenvectors of the matrix $\hat{\Sigma}_R = \sum_{i=1}^n X_i^T\mathrm{P}_2X_i/n$ corresponding to its d_R largest nonzero eigenvalues, where $\mathrm{P}_2 = G_2G_2^T$.*

(iii) *For fixed G_1 and G_2, the minimizer $\hat{\nu}_i = G_2^TX_iG_1$, $i = 1, \ldots, n$.*

Based on Proposition 2, for fixed G_1 and G_2, if ω_i is replaced by $\hat{\nu}_i = G_2^TX_iG_1$, the objective function (2.5) is

$$L_1 = \operatorname{tr}(\sum_{i=1}^{n}X_i^TX_i) - \operatorname{tr}[\sum_{i=1}^{n}(X_i^T\mathrm{P}_2X_i)\mathrm{P}_1].$$

Then for fixed P_2, L_1 is minimized by choosing the columns of G_1 to be the first d_R eigenvectors of $\sum_{i=1}^n X_i^T\mathrm{P}_2X_i$. So we need to choose P_2 to minimize $L_{12} = \sum_{k=1}^{d_R}\lambda_k(\sum_{i=1}^n X_i^T\mathrm{P}_2X_i)$, where $\lambda_k(A)$ indicates the kth eigenvalue of A. This can be treated as an optimization problem over a Grassmann manifold but it is hard to solve because eigenvalues are involved in the objective function. Instead, we apply an iterative algorithm that can solve the problem efficiently. We assume that the predictors are centered.

1. Generate an initial value of $\Gamma_{10} \in \mathbb{R}^{p_L \times d_L}$ and let $\hat{\Gamma}_1 = \Gamma_{10}$.
2. For given $\hat{\Gamma}_1$, compute the matrix $\hat{\Sigma}_L = \sum_{i=1}^n X_i \hat{\Gamma}_1 \hat{\Gamma}_1^T X_i^T / n$ and find its first d_L eigenvectors, denoted as $\hat{v}_1, \hat{v}_2, ..., \hat{v}_{d_L}$. Estimate Γ_2 as $\hat{\Gamma}_2 = [\hat{v}_1, \hat{v}_2, ..., \hat{v}_{d_L}]$.
3. For given $\hat{\Gamma}_2$, compute $\hat{\Sigma}_R = \sum_{i=1}^n X_i^T \hat{\Gamma}_2 \hat{\Gamma}_2^T X_i / n$; find the first d_R eigenvectors of $\hat{\Sigma}_R$, denoted as $\hat{l}_1, \hat{l}_2, ..., \hat{l}_{d_R}$, which form the columns of $\hat{\Gamma}_1$ as $\hat{\Gamma}_1 = [\hat{l}_1, \hat{l}_2, ..., \hat{l}_{d_R}]$.
4. For given $\hat{\Gamma}_1$ and $\hat{\Gamma}_2$, compute $\hat{\nu}_i = \hat{\Gamma}_2^T X_i \hat{\Gamma}_1$, $i = 1, \ldots, n$.
5. Repeat Step 2 to 4 and iterate each time using the updated $\hat{\Gamma}_1$ and $\hat{\Gamma}_2$ until $\sum_{i=1}^n \mathrm{tr}[(X_i - \Gamma_2 \nu_i \Gamma_1^T)^T (X_i - \Gamma_2 \nu_i \Gamma_1^T)]$ converges.

The MLE of the central dimension folding subspace $\mathcal{S}_{\nu|\circ X\circ}$ is then equal to $\mathrm{Span}(\hat{\Gamma}_1) \otimes \mathrm{Span}(\hat{\Gamma}_2)$. Consequently, $\hat{\sigma}^2$ is equal to $[1/(np_Lp_R)] \sum_{i=1}^n \mathrm{tr}[(X_i - \hat{\Gamma}_2 \hat{\nu}_i \hat{\Gamma}_1^T)^T (X_i - \hat{\Gamma}_2 \hat{\nu}_i \hat{\Gamma}_1^T)]$. The estimators obtained from the dimension folding model inherit the asymptotic properties of likelihood estimation under normality.

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

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

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

Conventional PCA targeting vectorized X costs $O(p_L^2 p_R^2 n)$ operations, which is more expensive under the mild condition that $\max(d_L, d_R)m < \min(p_L, p_R)^2$.

2.3. Relationship with tensor PCA

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

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

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

Tensor PCA is a well-established data-analytic method but is not associated with any probabilistic model. Dimension folding PCA can be treated as a model-based tensor PCA. It gains properties from maximum likelihood estimation when the predictors are approximately normally distributed. The normality assumption, however, is not essential in our model and can be relaxed to a general distribution. In this case, dimension folding PCA is equivalent to tensor PCA. The robustness of the dimension folding model regarding its normality assumption will be further discussed in Section 4.2.

3. Dimension Folding PFC

Although dimension folding PCA can reduce the predictor's row and column dimensions simultaneously, it performs dimension folding marginally and

the relationship between the predictor and the response is omitted. Instead of regressing X on a latent matrix ν, dimension folding PFC models the inverse regression of $X|Y$ and provides more informative estimation of the central dimension folding subspace $\mathcal{S}_{Y|\circ X\circ}$.

3.1. Formulation of dimension folding PFC

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

$$X = \mu + \Gamma_2\beta_2 f(Y)\beta_1^T\Gamma_1^T + \varepsilon \tag{3.1}$$

or, equivalently,

$$\mathrm{vec}(X) = \mathrm{vec}(\mu) + (\Gamma_1 \otimes \Gamma_2)(\beta_1 \otimes \beta_2)\mathrm{vec}(f(Y)) + \mathrm{vec}(\varepsilon), \tag{3.2}$$

where $f(Y) \in \mathbb{R}^{r_L\times r_R}$ contains elements formalized as functions of Y, $\beta_1 \in \mathbb{R}^{d_R\times r_R}(d_R \le r_R)$ and $\beta_2 \in \mathbb{R}^{d_L\times r_L}(d_L \le r_L)$ are the coefficient matrices of rank d_R and d_L, and ε is the random error independent of Y. It can be isotropic following the matrix normal distribution $\sigma\mathbf{N}_{p_L\times p_R}(0_{p_L\times p_R}, I_{p_R}, I_{p_L})$ or more general with $\mathbf{N}_{p_L\times p_R}(0_{p_L\times p_R}, \Omega, M)$ error. In Section 4.2, we show that the normality assumption is not necessary in order to obtain consistent estimation. The other terms in (3.1) are defined as in Section 2.1. Based on (3.1), each coordinate X_{ij} of X is a linear function of the elements in $f(Y)$ plus a random error. In addition,

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

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

The function $f(Y)$ is determinable in some cases, for instance when inverse response plots (Cook (1998, Chap.10)) of X_{ij} versus Y are informative about $f(Y)$, or when the response Y is categorical. In other cases, one can approximate $f(Y)$ by a series of basis functions or piecewise basis functions. Usually $f(Y)$ can be chosen as a diagonal matrix with dimension $r_L = r_R = r$. We use this matrix form in the rest of this paper. When using polynomial approximations, $f(Y)$ is then a diagonal matrix with diagonal elements of $Y, Y^2,\dots,Y^r$. Correspondingly, the conditional expectation $[\Gamma_2^T\mathrm{E}(X|Y)\Gamma_1]_{ij}$ is $(\Gamma_2^T\mu\Gamma_1)_{ij} + \sum_{k=1}^{r}\beta_{ik}^{(2)}\beta_{kj}^{(1)}Y^k$, which often captures the main regression shape of X on Y when r is relative large. In fact, in Section 4.1 we show that in order to receive a consistent estimator for the central dimension folding subspace, the selected fitting function does

not need to be very close to the true function, it is only required to be correlated to it. This indicates that an approximation with a finite dimension for $f(Y)$ is generally adequate.

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

Compared with slicing-based methods, dimension folding PFC provides the flexibility to formulate the relationship between X and Y. It can more effectively use the response information by choosing an appropriate fitting function to perform dimension folding. Slicing function can be considered as one special choice for fitting $f(Y)$ but it is generally less accurate when Y is continuous. A proposition identifies the central dimension folding subspace for the dimension folding model (3.1).

Proposition 3. *Under* (3.1)*, when the random error ε is isotropic the central dimension folding subspace $\mathcal{S}_{Y|\circ X\circ} = Span(\Gamma_1) \otimes Span(\Gamma_2)$; when ε has a general matrix normal distribution $\mathbf{N}_{p_L\times p_R}(0_{p_L\times p_R}, \Omega, M)$, the central dimension folding subspace $\mathcal{S}_{Y|\circ X\circ} = Span(\Omega^{-1}\Gamma_1) \otimes Span(M^{-1}\Gamma_2)$.*

Other ways to formulate the dimension folding PFC model are discussed in Section 8. We focus on estimating model (3.1) with both isotropic error and general error in the next section. Without loss of generality, the predictor X and the fitting function $f(Y)$ are assumed to be centered.

3.2. Estimation of dimension folding PFC

3.2.1. Isotropic error

When ε is isotropic with distribution $\sigma\mathbf{N}_{p_L\times p_R}(0_{p_L\times p_R}, I_{p_R}, I_{p_L})$, the central dimension folding subspace $\mathcal{S}_{Y|\circ X\circ}$ is equal to $\text{Span}(\Gamma_1)\otimes\text{Span}(\Gamma_2)$. For a random sample of size n from (Y, X), the MLE of $\mathcal{S}_{Y|\circ X\circ}$ is obtained based on the log likelihood function of (3.1):

$$\begin{aligned} &l(\mu, \mathcal{S}_{\Gamma_1}, \mathcal{S}_{\Gamma_2}, \sigma^2, \beta_1, \beta_2) \\ &\quad = -\frac{np_Lp_R}{2}\log(2\pi) - \frac{np_Lp_R}{2}\log\sigma^2 \\ &\qquad -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\text{tr}((X_i - \mu - \Gamma_2\beta_2 f(Y_i)\beta_1^T\Gamma_1^T)^T(X_i - \mu - \Gamma_2\beta_2 f(Y_i)\beta_1^T\Gamma_1^T). \quad (3.3) \end{aligned}$$

It is easy to see that the MLE $\hat{\mu} = \bar{X}$. Thus for any arbitrary σ^2, maximizing (3.3) is equivalent to minimizing the empirical expectation

$$\mathrm{E}_n\{\mathrm{tr}[(X - \Gamma_2\beta_2 f(Y)\beta_1^T\Gamma_1^T)^T(X - \Gamma_2\beta_2 f(Y)\beta_1^T\Gamma_1^T)]\} \tag{3.4}$$

over X and Y.

Proposition 4. *Suppose that $X \in \mathbb{R}^{p_L \times p_R}$ is a random matrix and $Y \in \mathbb{R}^1$ is a random variable. Let $(\hat{\Gamma}_1, \hat{\Gamma}_2, \hat{\beta}_1, \hat{\beta}_2)$ be minimizers of*

$$\mathrm{E}_n\{\mathrm{tr}[(X - G_2 b_2 f(Y) b_1^T G_1^T)^T(X - G_2 b_2 f(Y) b_1^T G_1^T)]\}. \tag{3.5}$$

over all $G_1 \in \mathbb{R}^{p_R \times d_R}$, $G_2 \in \mathbb{R}^{p_L \times d_L}$, $b_1 \in \mathbb{R}^{d_R \times r_R}$, and $b_2 \in \mathbb{R}^{d_L \times r_L}$. Then

(i) *For fixed G_1 and b_1, the columns of the minimizer $\hat{\Gamma}_2$ over G_2 are given by the d_L eigenvectors of the matrix*

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

corresponding to its d_L largest nonzero eigenvalues, where $f^* = f(Y)b_1^T$. The minimizer $\hat{\beta}_2 = \hat{\Gamma}_2^T \mathrm{E}_n(XG_1 f^{*^T})[\mathrm{E}_n(f^* f^{*^T})]^{-1}$.

(ii) *For fixed G_2 and b_2, the columns of the minimizer $\hat{\Gamma}_1$ over G_1 consist of the d_R eigenvectors of the matrix*

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

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

Similar to Proposition 2, after replacing G_2 and b_2 with their optimum solutions $\hat{\Gamma}_2$ and $\hat{\beta}_2$ obtained from Proposition 4(i), the problem becomes an optimization over a Grassmann manifold, but it is complicated to solve. Instead, we choose a simple iterative algorithm to estimate the likelihood function (3.3) as follows.

1. Generate initial values of Γ_{10} and β_{10} and let $\hat{\Gamma}_1 = \Gamma_{10}$ and $\hat{\beta}_1 = \beta_{10}$.
2. For given $\hat{\Gamma}_1$ and $\hat{\beta}_1$, compute the matrix $\hat{\Sigma}_{\mathrm{fit}_L} = \mathbb{X}_L^T \mathrm{P}_{\mathbb{F}_L}\mathbb{X}_L/n$, where $\mathbb{X}_L = (X_1\hat{\Gamma}_1, \ldots, X_n\hat{\Gamma}_1)^T$, $\mathbb{F}_L = (f_1^*, \ldots, f_n^*)^T$ with $f_i^* = f(Y_i)\hat{\beta}_1^T$. Then the term $\mathrm{P}_{\mathbb{F}_L}\mathbb{X}_L$ represents the fitted values from the multivariate regression of $X\hat{\Gamma}_1$ on $f(Y)\hat{\beta}_1^T$. Therefore, $\hat{\Sigma}_{\mathrm{fit}_L}$ is the sample column covariance matrix of the fitted values of $X\hat{\Gamma}_1$. Then the columns of $\hat{\Gamma}_2$ are estimated by the first d_L eigenvectors of $\hat{\Sigma}_{\mathrm{fit}_L}$ and $\hat{\beta}_2 = \hat{\Gamma}_2^T \mathbb{X}_L^T \mathbb{F}_L(\mathbb{F}_L^T\mathbb{F}_L)^{-1}$.

3. For given $\hat{\Gamma}_2$ and $\hat{\beta}_2$, compute the matrix $\hat{\Sigma}_{\text{fit}_R} = \mathbb{X}_R^T \mathrm{P}_{\mathbb{F}_R} \mathbb{X}_R / n$, where $\mathbb{X}_R = (X_1^T \hat{\Gamma}_2, \ldots, X_n^T \hat{\Gamma}_2)^T$, $\mathbb{F}_R = ({f_1^*}^T, \ldots, {f_n^*}^T)^T$ with $f_i^* = \hat{\beta}_2 f(Y_i)$. The term $\mathbb{X}_R^T \mathrm{P}_{\mathbb{F}_R}$ represents the fitted values from the multivariate regression of $\hat{\Gamma}_2^T X$ on $\hat{\beta}_2 f(Y)$. Then $\hat{\Sigma}_{\text{fit}_R}$ represents the sample row covariance matrix of the fitted values of $\hat{\Gamma}_2^T X$. The columns of $\hat{\Gamma}_1$ are given by the first d_R eigenvectors of $\hat{\Sigma}_{\text{fit}_R}$ and $\hat{\beta}_1 = \hat{\Gamma}_1^T \mathbb{X}_R^T \mathbb{F}_R (\mathbb{F}_R^T \mathbb{F}_R)^{-1}$.
4. Repeat Steps 2−3 and iterate each time with the updated estimators until the objective function (3.4) converges.

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

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

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

From a computational perspective, the proposed algorithm is more economical than conventional PFC and dimension folding SIR. Its major costs come from the computation of $\hat{\Sigma}_{\text{fit}_L}$ and $\hat{\Sigma}_{\text{fit}_R}$. For $\hat{\Sigma}_{\text{fit}_L}$, computing $\mathbb{X}_L$ and $\mathbb{F}_L$ requires $np_Lp_Rd_R$ and $nr_Lr_Rd_R$ operations, and computing $\mathbb{X}_L^T \mathbb{F}_L$ and $\mathbb{F}_L^T \mathbb{F}_L$ requires $nd_Rp_Lr_L$ and $nd_Rr_L^2$ operations. The inverse of $\mathbb{F}_L^T \mathbb{F}_L$ costs $O(r_L^3)$. Therefore, the total cost of $\hat{\Sigma}_{\text{fit}_L}$ is at most $O(\max(nd_R, r_L)\max(p_L, p_R, r_L, r_R)^2)$. Similarly, the cost of $\hat{\Sigma}_{\text{fit}_R}$ is of order $O(\max(nd_L, r_R)\max(p_L, p_R, r_L, r_R)^2)$. Thus, dimension folding PFC with an isotropic error requires at most $O(\max(nd_L, nd_R, r_L, r_R)\max(p_L, p_R, r_L, r_R)^2 m)$ operations with m iterations. Analogously, it can be shown that the computations of conventional PFC and dimension folding SIR targeting on $\mathrm{vec}(X)$ take at least $O(\max(n, p_Lp_R)\max(p_Lp_R, r)r)$ and $O(p_L^2 p_R^2 \max(p_Lp_R, n)k)$ operations, which are in general more than dimension folding PFC when p_L and p_R are relative large. Here r is the dimension of the fitting function in conventional PFC and k is the iteration number in dimension folding SIR.

3.2.2. General error

In this section, we consider a general error structure for ε with the matrix normal distribution $\mathbf{N}_{p_L \times p_R}(0_{p_L \times p_R}, \Omega, M)$. Based on this covariance structure,

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

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

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

$$
\begin{aligned}
&l(\mu, \mathcal{S}_{\Gamma_1}, \mathcal{S}_{\Gamma_2}, \beta_1, \beta_2, \Omega, M) \\
&= -\frac{np_L p_R}{2}\log(2\pi) - \frac{np_L}{2}\log|\Omega| - \frac{np_R}{2}\log|M| \\
&\quad -\frac{1}{2}\sum_{i=1}^{n} \text{tr}\{\Omega^{-1}(X_i - \mu - \Gamma_2\beta_2 f(Y_i)\beta_1^T\Gamma_1^T)^T M^{-1}(X_i - \mu - \Gamma_2\beta_2 f(Y_i)\beta_1^T\Gamma_1^T)\}. \quad (3.6)
\end{aligned}
$$

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

Proposition 5. *Suppose that $X_i \in \mathbb{R}^{p_L \times p_R}$, $i = 1, \ldots, n$ are observed and centered matrices, and let $(\hat{\Gamma}_1, \hat{\Gamma}_2, \hat{\beta}_1, \hat{\beta}_2, \hat{\Omega}, \hat{M})$ be the minimizers of* (3.6).

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

Table 1. Comparison of computation complexity.

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

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

To estimate the parameters in (3.6), one can begin with initial estimates of Ω, Γ_1, and β_1, then iterate the two steps in Proposition 5 until the log likelihood function (3.6) converges. The computational cost of dimension folding PFC under a general error is in general less expensive than that of conventional PFC and dimension folding SIR. We summarize the results for all models in Table 1.

Remark 1. According to Proposition 5, $\hat{M}$ is invertible when $\hat{M}_{\text{res}}$ is invertible. The existence of $\hat{M}_{\text{res}}^{-1}$ only requires that $p_L \le np_R-1$ and $\text{Rank}(I-\text{P}_{\mathbb{F}_L})=p_L$. The latter condition is generally satisfied since the nonzero eigenvalues of $\text{P}_{\mathbb{F}_L}$ are unlikely to be exactly equal to one and they are unlikely to be all identical. Hence it is usually guaranteed that $\hat{M}^{-1}$ and $\hat{\Omega}^{-1}$ exist if $p_L \le np_R-1$ and $p_R \le np_L-1$ or, equivalently, $n>\max(p_L/p_R,p_R/p_L)-1$.

Remark 2. The maximum matrix dimension required in Proposition 5 is $np_L\times np_L$ or $np_R\times np_R$, from $\text{P}_{\mathbb{F}_L}$ or $\text{P}_{\mathbb{F}_R}$. This dimension could be very large ($>30000\times30000$) in some cases (e.g. the EEG data) and exceed the storage limit in R software. In this case, one can apply an equivalent iteration algorithm that i) chooses moment estimators of Ω and M as initial values of $\hat{\Omega}$ and $\hat{M}$; ii) standardizes the predictors as $Z_i=\hat{M}^{-1/2}X_i\hat{\Omega}^{-1/2}$; iii) applies isotropic dimension folding PFC to the standardized data; iv) updates $\hat{\Omega}$ and $\hat{M}$ according to (S1.4) and (S1.5), the MLEs of matrix normal distribution (Dutilleul (1999)) described in the supplement file; v) repeats ii)-iv) using the updated parameter values until the likelihood function converges.

Remark 3. Although the proposed algorithms are quite efficient for estimating the central dimension folding subspace based on random initial values, using the

conventional PFC model to obtain initial values can guarantee consistency of the estimators when the fitted function $f(Y)$ is misspecified. This is discussed in Section 4 and the supplement file.

Let $\mathcal{S}_d(A)$ denote the span of the d eigenvectors of A corresponding to its d largest eigenvalues, and let $\mathcal{S}_d(A,B) = A^{-1/2}\mathcal{S}_d(A^{-1/2}BA^{-1/2})$. Corollary 1 provides five equivalent forms of the MLE of the central dimension folding subspace. We applied the original form $\mathcal{S}_{d_R}(\hat{\Omega},\hat{\Sigma}_{\text{fit}_R}) \otimes \mathcal{S}_{d_L}(\hat{M},\hat{\Sigma}_{\text{fit}_L})$ in our simulation and data analysis.

Corollary 1. *The MLE of* $\mathcal{S}_{Y|\circ X\circ}$ *under* (3.1) *with an general error is* $\mathcal{S}_{d_R}(\hat{\Omega},\hat{\Sigma}_{\text{fit}_R})\otimes\mathcal{S}_{d_L}(\hat{M},\hat{\Sigma}_{\text{fit}_L})$. *It is equivalent to* $\mathcal{S}_{d_R}(\hat{\Omega}_{\text{res}},\hat{\Sigma}_{\text{fit}_R})\otimes\mathcal{S}_{d_L}(\hat{M}_{\text{res}},\hat{\Sigma}_{\text{fit}_L}) = \mathcal{S}_{d_R}(\tilde{\Omega},\hat{\Sigma}_{\text{fit}_R}) \otimes \mathcal{S}_{d_L}(\tilde{M},\hat{\Sigma}_{\text{fit}_L}) = \mathcal{S}_{d_R}(\hat{\Omega},\tilde{\Omega}) \otimes \mathcal{S}_{d_L}(\hat{M},\tilde{M}) = \mathcal{S}_{d_R}(\hat{\Omega}_{\text{res}},\tilde{\Omega}) \otimes \mathcal{S}_{d_L}(\hat{M}_{\text{res}},\tilde{M})$.

4. Robustness

In this section, we study the robustness of the estimator $\mathcal{S}_{d_R}(\hat{\Omega},\hat{\Sigma}_{\text{fit}_R}) \otimes \mathcal{S}_{d_L}(\hat{M},\hat{\Sigma}_{\text{fit}_L})$ when $f(Y)$ in model (3.1) is misspecified and the normality assumption is violated.

4.1. Misspecification of $f(Y)$

Under (3.1), we now assume that the true fitting function $f(Y)$ is misspecified by using the user-selected function $h(Y)$ in place of $f(Y)$. It can be shown that the estimator of the central dimension folding subspace is still consistent under certain conditions. To simplify the notation, let $g = \beta_2 f(Y)\beta_1^T$ and $l = \kappa_2 h(Y)\kappa_1^T$ be the misspecified fitting components. Note that g and l are both centered. We take $\rho_L = \text{var}_c^{-1/2}(g)\text{cov}_c(g,l)\text{var}_c^{-1/2}(l)$ to be the $d_L \times d_L$ column correlation matrix between the elements of g and l, where $\text{var}_c(g) = \text{E}(gg^T)$ is the column variance of g, $\text{var}_c(l) = \text{E}(ll^T)$ is the column variance of l, and $\text{cov}_c(g,l) = \text{E}(gl^T)$ is the column covariance matrix between g and l; let $\rho_R = \text{var}_r^{-1/2}(g)\text{cov}_r(g,l)\text{var}_r^{-1/2}(l)$ be the $d_R \times d_R$ row correlation matrix between the elements of g and l, where $\text{var}_r(g) = \text{E}(g^Tg)$ and $\text{var}_r(l) = \text{E}(l^Tl)$ are row variance matrices of g and l, and $\text{cov}_r(g,l) = \text{E}(g^Tl)$ is the row covariance matrix between g and l.

Proposition 6. $\mathcal{S}_{d_R}(\hat{\Omega},\hat{\Sigma}_{\text{fit}_R}) \otimes \mathcal{S}_{d_L}(\hat{M},\hat{\Sigma}_{\text{fit}_L})$ *is a* $\sqrt{n}$ *consistent estimator of* $\text{Span}(\Omega^{-1}\Gamma_1)\otimes\text{Span}(M^{-1}\Gamma_2)$ *if and only if* ρ_L *has rank* d_L *and* ρ_R *has rank* d_R.

Thus $\mathcal{S}_{d_R}(\hat{\Omega},\hat{\Sigma}_{\text{fit}_R}) \otimes \mathcal{S}_{d_L}(\hat{M},\hat{\Sigma}_{\text{fit}_L})$ can still be a reasonable estimator when $f(Y)$ is misspecified and the normality assumption is violated, as long as the row and columns correlations between the true fitting function and the selected

fitting function have full ranks. This result is a generalization of Theorem 3.5 in Cook and Forzani (2008), and it is a mild condition. Nevertheless, in applications care should be taken when selecting $f(Y)$ in order to obtain better estimates. Polynomial approximations can be simple and good choices.

4.2 Normality assumption

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

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

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

Proposition 7. *Under (3.1), when the random error is elliptically contoured distributed as* $\mathrm{EC}_{p_Lp_R}(0, \Omega \otimes M, Q)$, $\mathcal{S}_{fSIR} \subseteq \mathcal{S}_{fPFC} \subseteq \mathcal{S}_{Y|\circ X\circ}$, *where* $\mathcal{S}_{fPFC} = \mathrm{Span}(\Omega^{-1}\Gamma_1) \otimes \mathrm{Span}(M^{-1}\Gamma_2)$.

Thus, under the elliptically symmetric condition, the subspace $\mathrm{Span}(\Omega^{-1}\Gamma_1)\otimes \mathrm{Span}(M^{-1}\Gamma_2)$ given by dimension folding PFC is not guaranteed to be the true

central dimension folding subspace but a subspace of it. It contains the dimension folding SIR subspace at the population level and its sample estimate can be more accurate since the fitting function $f(Y)$ is generally more efficient than a slicing function. Therefore, under this minimum condition, dimension folding PFC is still useful. Both algorithms in Section 3.2 provide $\sqrt{n}$ consistent estimators for $\mathcal{S}_{fPFC}$ without normality, because the algorithm in Section 3.2.1 is equivalent to a least square estimation and the consistent estimation of the algorithm in Section 3.2.2 is given by Proposition 6, which does not rely on normality.

Similarly, Proposition 7 holds for dimension folding PCA in terms of $\mathcal{S}_{\nu|\circ X\circ}$ and $\zeta = \mathrm{E}[\mathrm{vec}(X)|\nu]$. Hence dimension folding PCA and PFC are beneficial under the minimum elliptically symmetric condition.

5. Prediction

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

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

The second method was motivated by a nonparametric prediction technique of Adragni and Cook (2009). Let $f(X)$ and $f(X|Y)$ be the density functions of X and $X|Y$. Let $R(X)$ denote a sufficient folding assumed to have a density. Then $\mathrm{E}[Y|X = x] = \mathrm{E}\{Y f[R(x)|Y]\}/\mathrm{E}\{f[R(x)|Y]\}$. This provides the key idea of this nonparametric prediction approach because the estimated prediction function

$\hat{\mathrm{E}}[Y|X = x]$ can be written as $\hat{\mathrm{E}}[Y|X = x] = \sum_{i=1}^{n} \omega_i(x) Y_i$, where $\omega_i(x) = \hat{f}[\hat{R}(x)|Y_i] / \sum_{i=1}^{n} \hat{f}[\hat{R}(x)|Y_i]$.

Once the density function $f(X|Y)$ is estimated, the predicted value $\hat{Y}$ can be easily obtained since it is the weighted average of the observed responses. This method is applicable to our proposed dimension folding models since the conditional distribution of $X|Y$ is known through the model assumptions. According to (S1.1) in the supplement file, when the random error ε is isotropic we have

$$\begin{aligned}\hat{f}[\hat{R}(x)|Y_i] &= \hat{f}[\hat{R}(\mathrm{vec}(x))|Y_i] \\ &\propto \exp\{-(2\hat{\sigma}^2)^{-1} || (\hat{\Gamma}_1 \otimes \hat{\Gamma}_2)^T [\mathrm{vec}(x) - \mathrm{vec}(\hat{X}_i)] ||^2\} \\ &= \exp\{-(2\hat{\sigma}^2)^{-1} || \hat{R}(\mathrm{vec}(x)) - \hat{R}(\mathrm{vec}(\hat{X}_i)) ||^2\}, \end{aligned} \tag{5.1}$$

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

$$\begin{aligned}\hat{f}[\hat{R}(x)|Y_i] = \hat{f}[\hat{R}(\mathrm{vec}(x))|Y_i] \propto \exp\{-\frac{1}{2} \| [(\hat{\Gamma}_1 \otimes \hat{\Gamma}_2)^T (\hat{\Omega} \otimes \hat{M})^{-1} (\hat{\Gamma}_1 \otimes \hat{\Gamma}_2)]^{-1/2} \\ \times [\hat{R}(\mathrm{vec}(x)) - \hat{R}(\mathrm{vec}(\hat{X}_i))] \|^2\}, \end{aligned} \tag{5.2}$$

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

Each method outperforms the other under certain conditions. The inverse regression prediction relies on the density function $f[R(X)|Y]$ but does not make any parametric assumption on modeling $Y|X$, while forward regression prediction usually assumes a parametric model on $Y|X$ or it depends on the estimation of $Y|X$. Thus, the inverse prediction method shows its advantages when the distribution of the random error ε in model (3.1) is known or can be well estimated. The forward prediction is beneficial when the assumption made on $Y|X$ is reasonable.

In addition, the choice of $f(Y)$ can affect the prediction accuracy. Consider the mean squared error $\mathrm{MSE} = \mathrm{E}[Y - \hat{Y}(X)]^2$ for which the minimum prediction error is achieved when $\hat{Y}(X)$ is the conditional mean $\mathrm{E}(Y|X)$. According to Proposition 6, when the row and column correlations of the selected fitting function $\kappa_2 h(Y) \kappa_1$ and the true function both have full ranks, which indicates that the two are correlated, the estimator of the central dimension folding subspace is $\sqrt{n}$ consistent. For the forward prediction method, we have $\hat{Y}(X) = \hat{\mathrm{E}}(Y|\hat{R}(X)) = \hat{\mathrm{E}}(Y|\hat{\Gamma}_2^T \hat{M}^{-1} X \hat{\Omega}^{-1} \hat{\Gamma}_1)$. If one chooses $\hat{\mathrm{E}}(Y|R(X))$ to be a consistent estimator for $\mathrm{E}(Y|R(X))$, such as the Nadaraya-Watson estimator, then under mild regularity conditions, $\hat{Y}(X) \to \mathrm{E}(Y|R(X)) = \mathrm{E}(Y|X)$ when the selected fitting function is correlated to the true function. Thus the prediction

error can reach its minimum asymptotically if the condition in Proposition 6 is satisfied. For the inverse prediction method, we have

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

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

6. Simulation Studies

6.1. Evaluation of estimation accuracy

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

$$\text{PCDF\_Error} = \|P_{\hat{\mathcal{S}}_{Y|\circ X\circ}} - P_{\mathcal{S}_{Y|\circ X\circ}}\|_F^2, \tag{6.1}$$

$$\text{PCS\_Error} = \|P_{\hat{\mathcal{S}}_{Y|\mathrm{vec}(X)}} - P_{\mathcal{S}_{Y|\circ X\circ}}\|_F^2, \tag{6.2}$$

where $||\cdot||_F$ is the Frobenius norm.

To evaluate the performance of the dimension folding PCA model (2.2), the data were generated as follows: Let $d_L = d_R = 2$ and $p_L = p_R = p$, with sample size $n = 100$. The components of Γ_1 and Γ_2 were generated from $N(0,1)$ and the components ν_i before centering were generated from $N(1,2)$, $i = 1,\ldots,n$. The vectorized isotropic error ε was obtained from the multivariate normal with mean zero and covariance matrix $0.8I_{p_Lp_R}$. We chose $p = 5, 10, 15, 20$ and 30, and ran each simulation 1,000 times. The results are summarized in Figure 1. We used "DF-PCA", "DF-PFC" and "DF-SIR" to denote dimension folding PCA, dimension folding PFC, and dimension folding SIR in figures and tables. It can be seen that for all selected dimensions of p, dimension folding PCA was noticeably more accurate than PCA. As the predictor's dimension increases both methods showed ascending estimation distance from the true projection space,

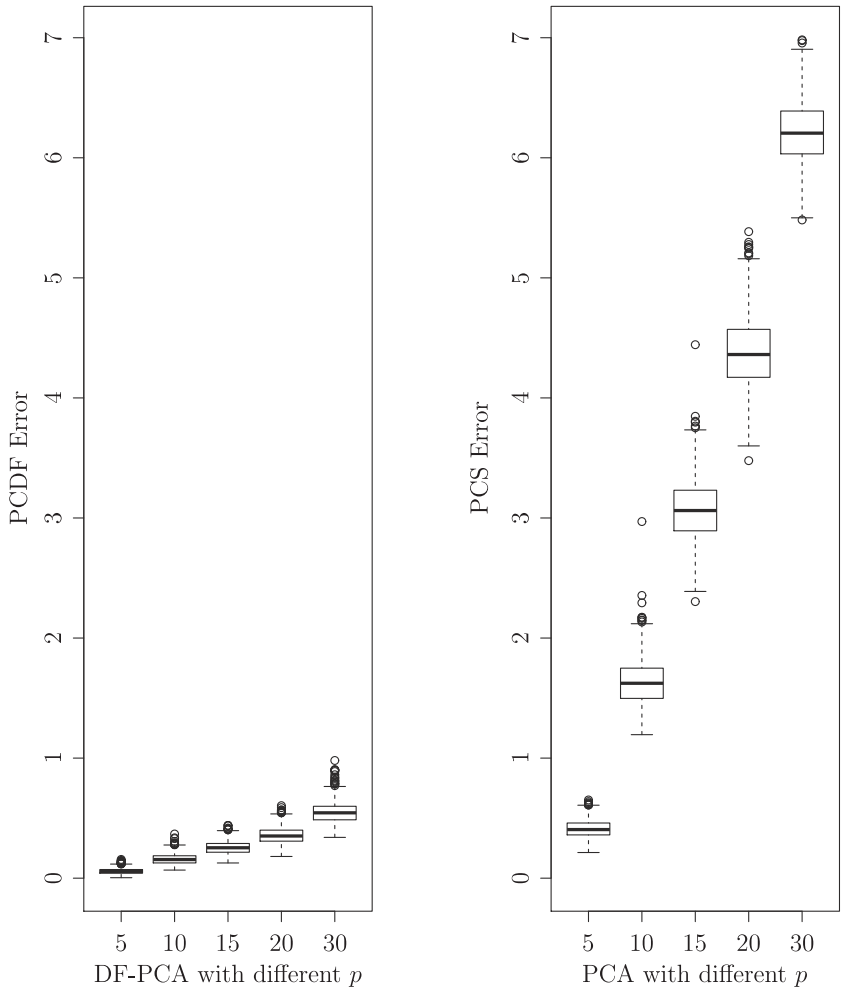


Figure 1. The comparison results of DF-PCA and PCA.

but dimension folding PCA had the slower error increase in both the mean and standard deviation.

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

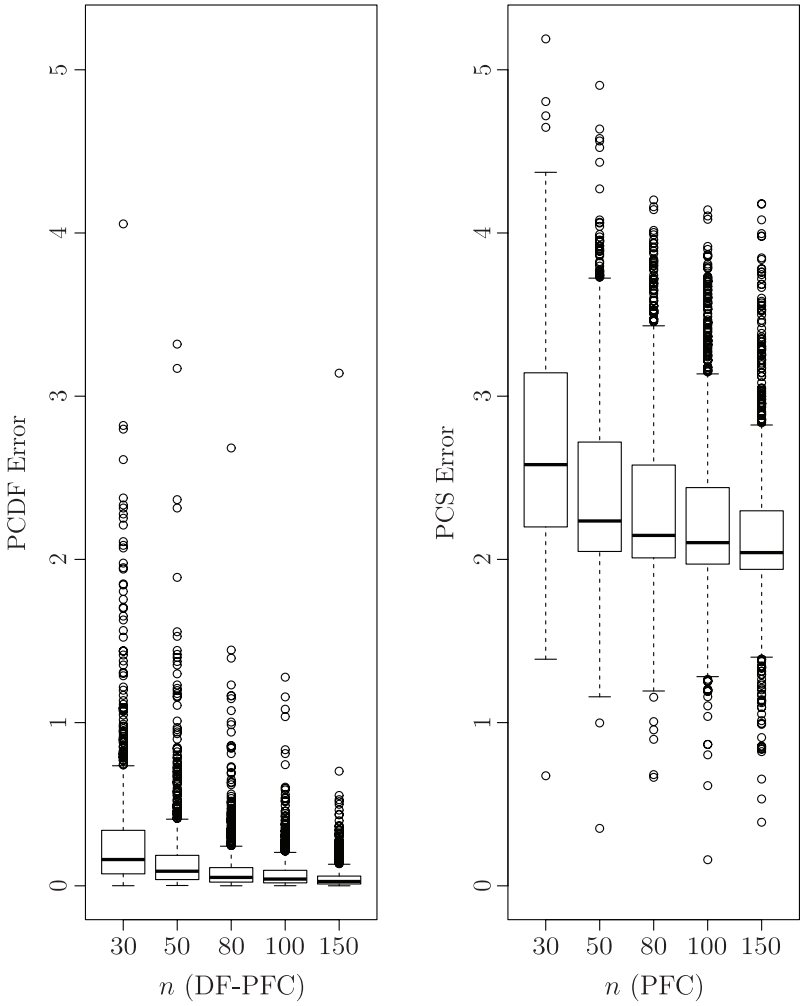
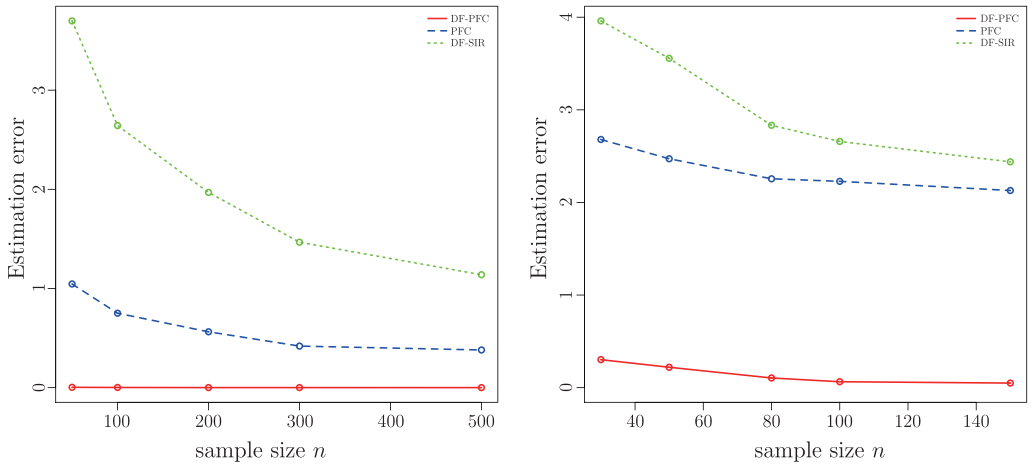


Figure 2. The comparison results of DF-PFC and PFC under general error.

The covariance matrices were

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

For the isotropic error case, we chose $p_L = p_R = 10$ and $\sigma = 0.8$, with sample size $n = 120, 150, 200, 300, 500$. The other parameters were kept the same as those in the general error case. We ran the simulation 1,000 times for each sample size. Figure 2 summarizes the results under the general error setting. It can be seen that the central dimension folding subspaces were estimated precisely based



(a) isotropic error (b) general error

Figure 3. The comparison results of DF-PFC, DF-SIR and PFC.

on the estimation procedures proposed in Section 3.2 except for some extreme outliers. Although the plots appear with dense outliers, the actual percentages of these outliers were less than 5% under 1,000 repetitions. Some outliers like the one with estimation error close to 3 at $n = 150$ could be due to the algorithm getting caught in a local minimum. Conventional PFC had much higher estimation errors for all sample sizes.

We further compared the model-based methods to dimension folding SIR. For the latter, 8 slices were selected for the response variable. Based on our simulation results, it was the best choice among 6, 8, 10 and 15 slices. The mean estimation errors are shown in Figure 3, based on 1,000 repetitions. It can be seen that dimension folding PFC provided the most accurate estimations for the central dimension folding subspace over all sample sizes. Although conventional PFC was less accurate than dimension folding PFC, it still beat dimension folding SIR to a large extent. Dimension folding SIR failed to obtain precise estimation because the conditional mean $\mathrm{E}(X|Y)$ was not adequately estimated by slicing the responses. The PFC methods benefitted from careful fitting of the inverse regression of X on Y.

6.2. Choice of d_L and d_R

In the previous sections, the reduced row and column dimensions d_L and d_R were assumed known. In applications, one can apply an information criterion, say AIC or BIC, to select optimal dimensions by minimizing the objective function $-2L(d_L, d_R) + h(n)g(d_L, d_R)$. Here $L(d_L, d_R)$ is the log likelihood function of the estimated model, $h(n)$ is $\log(n)$ for BIC and 2 for AIC, and $g(d_L, d_R)$ is the

Table 2. Percentages of correct identifications.

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

number of parameters to be estimated. One can also use the likelihood ratio test statistic $\Lambda(d_{L_0}, d_{R_0}) = 2(L(\min(r_L, p_L), \min(r_R, p_R)) - L(d_{L_0}, d_{R_0}))$ to perform sequential tests for increasing values of d_L and d_R.

We illustrate these procedures using the simulated samples obtained from the isotropic error setting. Here $d_L = d_R = 2$, $p_L = p_R = 10$, and $n = 200$ were chosen for both dimension folding PCA and dimension folding PFC. The simulations were repeated 1,000 times. All three methods were able to correctly identify the true dimensions over 95% of the time. When we took the true dimensions to be $(d_L, d_R) = (1, 1), (1, 2)$ and $(2, 1)$, the percentages of the precise identifications were over 90% for all methods.

6.3. Prediction

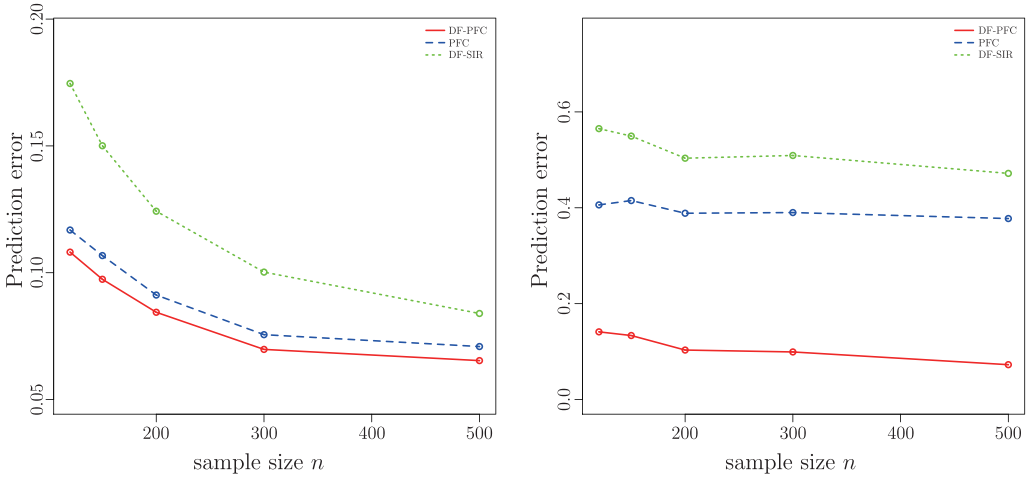
We evaluated the prediction performance of dimension folding PFC, conventional PFC, and dimension folding SIR using the simulated data under the isotropic error from Section 6.1. The two prediction methods in Section 5 were applied. For the first method, we fitted a generalized additive model of Y on the reduced predictor $(\hat{\Gamma}_1 \otimes \hat{\Gamma}_2)^T \mathrm{vec}(X)$ to the original data and then generated new data for prediction. The new data are denoted by (X_i^*, Y_i^*), $i = 1, \ldots, n_{\mathrm{new}}$, where $n_{\mathrm{new}} = n/4$. The average prediction error was calculated as:

$$\mathrm{PE} = \sum_{i=1}^{n_{\mathrm{new}}} \frac{(Y_i^* - \hat{\mathrm{E}}(Y|X = X_i^*))^2}{n_{\mathrm{new}}}. \tag{6.3}$$

This procedure was repeated for 1,000 data sets and the averaged prediction error $\sum_{i=1}^{1,000} \mathrm{PE}_i/1,000$ was used to assess the prediction accuracy of the three methods.

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

Figure 4(a) shows the prediction results with generalized additive model fitting. It illustrates the potential advantages of using an inverse regression model



(a) prediction with model fitting (b) nonparametric prediction

Figure 4. Prediction results under isotropic error.

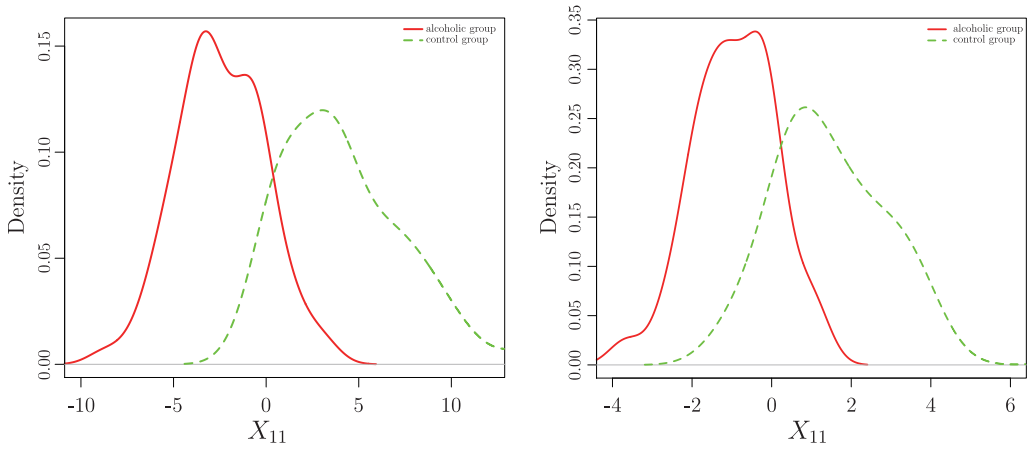
to estimate the conditional expectation $\mathrm{E}(X|Y)$, or $\mathrm{E}[\mathrm{vec}(X)|Y]$, instead of using a slicing method. Dimension folding PFC predicted best over all sample sizes. Though conventional PFC omits the predictor's matrix structure, it still gave more accurate results than did dimension folding SIR. Figure 4(b) shows the prediction performance according to the second prediction approach. It provided smaller prediction errors for dimension folding PFC and relatively large errors for conventional PFC and dimension folding SIR.

7. Data Analysis

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

7.1. EEG data

The primary goal of this study was to explore the relationship between alcoholism and the pattern of voltage values over times and channels. Let $(X_1, Y_1), \ldots, (X_{122}, Y_{122})$ denote the observed data, where X_i is a 256×64 matrix and Y_i is a binary univariate variable, $i = 1, \ldots, 122$. It is easy to see that error structure is not isotropic. In this case, conventional PFC is not applicable since $n \ll p_L \times p_R$. We applied dimension folding PFC with a general error to these data. Since our proposed estimation procedures circumvent vectorization of the predictors, we were able to handle the original EEG data without pre-screening



(a) DF-PFC without screen (b) DF-PFC with screen

Figure 5. Density plot with new reduced predictor X_{11}.

work, as in Li, Kim, and Altman (2010). In our case, the maximum dimension of a matrix inversion is 256 by 256 ($\hat{M}^{-1}$), instead of the 256×64 by 256×64 ($\hat{\Sigma}^{-1}$) required for the moment-based dimension folding methods. According to Remark 1 in Section 3.2.2, both inverse matrices $\hat{M}^{-1}$ and $\hat{\Omega}^{-1}$ exist for the original EEG data, because $n > \max(p_L/p_R, p_R/p_L) - 1$.

For a categorial response Y of h categories, $f(Y)$ can be naturally chosen as a diagonal matrix with its kth diagonal element $\text{diag}(f(Y))_k = I(Y \in H_k) - n_k/n$, $k = 1, \ldots, h-1$. Thus, for the EEG data, we have $d_L = d_R = r_L = r_R = r = 1$. Then the sufficient reduction $\hat{\Gamma}_2^T \hat{M}^{-1} X \hat{\Omega}^{-1} \hat{\Gamma}_1$ obtained by the dimension folding PFC model is a univariate variable, labeled as X_{11}. Figure 5(a) shows good separation of the two groups by X_{11} without pre-screening the original predictors. Figure 5(b) shows the corresponding result after pre-screening the predictors to smaller dimensions $(p_L^*, p_R^*) = (15, 15)$ with the screening method in Li, Kim, and Altman (2010). Pre-screening the predictors loses information about the original data as the two groups cannot be separated quite as well as in (a). To obtain classification results, we applied quadratic discriminant analysis and leave-one-out cross validation. Without pre-screening the original predictors, dimension folding PFC with a general error correctly classified 107 subjects out of the total 122 subjects based on X_{11}; after pre-screening the predictors, it classified 102 out of the 122 subjects. In comparison, dimension folding DR and dimension folding SIR provided 97 and 94 out of 122 correct decisions, using $(p_L^*, p_R^*) = (15, 15)$ and $(d_L, d_R) = (1, 2)$.

7.2. Dow Jones stock data

We used Dow Jones industrial stock data from January 2001 to December

Table 3. Prediction results ($\times 1,000$) with 10 folded cross validations.

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

2010. The response is the monthly Dow Jones industrial average index change rate. If m_i denotes the Dow Jones industrial average monthly index for the ith month, the responses $Y_i = (m_i - m_{i-1})/m_{i-1}$, $i = 1, \ldots, n$, are the index change rates, assumed to be independent. For each response (month), the predictor was formed by 19 daily stock price change rates over the 30 Dow Jones companies. We chose 19 daily stock price change rates because there are usually 19-23 trading days each month. Hence the predictor for each observation is a 19×30 matrix and the response is a univariate continuous variable. We deleted the observations in September 2001 and September 2008 due to the incidents of terrorism and the financial crisis, leaving $n = 118$ observation months. The final data set consisted of $(X_1, Y_1), \ldots, (X_{118}, Y_{118})$ observations. Primary interest was in association between monthly stock index change rates and the daily stock price change rates from the individual companies.

Dimension folding PFC with both isotropic and general errors, dimension folding SIR, and the Lasso were applied to our study. We evaluated the prediction performance for the first three methods using the prediction approach with OLS fitting of Y on the reduced predictor $\mathrm{vec}(\hat{\Gamma}_2^T X \hat{\Gamma}_1)$, as proposed in Section 5. Four sets of dimensions, $(d_L, d_R) = (1, 1)$, $(d_L, d_R) = (1, 2)$, $(d_L, d_R) = (2, 1)$, and $(d_L, d_R) = (2, 2)$, were selected. The function $f(Y)$ was chosen as a diagonal matrix with its diagonal elements formed by (Y, Y^2, Y^3, Y^4) for dimension folding PFC. Dimension folding SIR was studied with slicing numbers 6 and 8. We also applied the Lasso to select important signals in $\mathrm{vec}(X)$ and performed prediction. The 10-fold cross validation method was used to evaluate the prediction performance using (6.3) for all methods. The results are summarized in Table 3.

It can be seen that isotropic dimension folding PFC provided smaller prediction errors than all other methods. Since the dependence of the stock price change rates is not strong from day to day and from company to company, dimension folding PFC under a general error structure could be overparametrized and thus the prediction errors were likely to be increased. Dimension folding SIR presented less accurate results than the isotropic dimension folding PFC model over all selected dimensions and slicing numbers. Lasso showed relatively large prediction errors.

8 Discussion

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

There are different formulations for the dimension folding PFC model. Model (3.1) provides a multiplicative coefficient structure $\beta_2 f(Y)\beta_1^T$ for the fitted function. Instead, one can model dimension folding PFC with an additive coefficient structure, an interactive coefficient structure, or a general coefficient structure, respectively, as

$$X = \mu + \Gamma_2[\beta_2 f(Y)\mathbf{e}_{r_R,d_R} + \mathbf{e}_{d_L,r_L} f(Y)\beta_1^T]\Gamma_1^T + \varepsilon, \tag{8.1}$$

$$X = \mu + \Gamma_2[\beta_2 f(Y)\mathbf{e}_{r_R,d_R} + \mathbf{e}_{d_L,r_L} f(Y)\beta_1^T + \beta_2 f(Y)\beta_1^T]\Gamma_1^T + \varepsilon, \tag{8.2}$$

$$X = \mu + \Gamma_2 \mathrm{vec}^{-1}\{\beta g(Y)\}\Gamma_1^T + \varepsilon, \tag{8.3}$$

where $\mathbf{e}_{r_R,d_R}$ is a $r_R \times d_R$ matrix with all elements equal to one, and $\mathbf{e}_{d_L,r_L}$ is similarly defined. If $f(Y)$ is diagonal and its diagonal elements are formed by polynomial basis functions, then under (8.1) the folded conditional mean $[\Gamma_2^T \mathrm{E}(X|Y)\Gamma_1]_{ij} = \sum_{k=1}^{r}(\beta_{ik}^{(2)} + \beta_{kj}^{(1)})Y^k$, where the coefficients are additive. When the multiplicative or additive coefficient model itself is not sufficient to formulate the relationship between X and Y, (8.2) might be needed. In this case, $[\Gamma_2^T \mathrm{E}(X|Y)\Gamma_1]_{ij} = \sum_{k=1}^{r}(\beta_{ik}^{(2)} + \beta_{kj}^{(1)} + \beta_{ik}^{(2)}\beta_{kj}^{(1)})Y^k$. This is called the dimension folding PFC model with the interactive coefficient structure. More generally, one might not impose any constraints on the coefficients and adopt (8.3), where "vec^{-1}" stands for the matrixing operation. Then with polynomial basis functions as the components of $g(Y)$, the folded conditional mean $[\Gamma_2^T \mathrm{E}(X|Y)\Gamma_1]_{ij} = \sum_{k=1}^{r_L r_R} \beta_{(j-1)d_L+i,k} Y^k$, where $\beta_{(j-1)d_L+i,k}$ is the element in $[(j-1)d_L + i]$th row and kth column of β. The choice of a particular dimension folding PFC model depends on the intrinsic row and column structure of $X|Y$. To estimate model (8.3), one can apply the estimation procedure in Section 3.2, though the algorithm cannot be directly used for the dimension folding PFC model with the additive or the interactive coefficient structure. Instead, one can use numerical algorithms with least square iterations.

The proposed dimension folding models can also be generalized to array-valued predictors. If $\mathbf{X} = \{X_{i_1 \cdots i_m} : i_1 = 1, \ldots, p_1, \ldots, i_m = 1, \ldots, p_m\}$ is a m-way random array of dimension $p_1 \times \cdots \times p_m$ and Y is a univariate random response, dimension folding PCA and PFC are formulated as

$$\text{vec}\ (X) = \text{vec}(\mu) + (\Gamma_1 \otimes \Gamma_2 \otimes \cdots \otimes \Gamma_m)\text{vec}\ (\nu_i) + \text{vec}\ (\varepsilon), \tag{8.4}$$

$$\text{vec}(X) = \text{vec}(\mu) + (\Gamma_1 \otimes \Gamma_2 \otimes \cdots \otimes \Gamma_m) \cdot (\beta_1 \otimes \beta_2 \otimes \cdots \otimes \beta_m) \cdot \text{vec}(f(y)) + \text{vec}(\varepsilon), \tag{8.5}$$

respectively. Here $\Gamma_i \in \mathbb{R}^{p_{m-i} \times d_{m-i}}$, $i = 1, \ldots, m$, ν_i is a m-way array of dimension $d_1 \times \cdots \times d_m$, $\beta_i \in \mathbb{R}^{d_{m-i} \times r_{m-i}}$, $i = 1, \ldots, m$, $f(y)$ is a m-way array of dimension $r_1 \times \cdots \times r_m$, and vec (ε) has a multivariate normal distribution with mean $0_{p_1 \cdots p_m \times p_1 \cdots p_m}$ and covariance matrices $\Omega_1 \otimes \Omega_1 \otimes \cdots \otimes \Omega_m$. It can be shown that the dimension folding subspace with m-way array-valued predictors is $\text{Span}\{(\Omega_1 \otimes \Omega_1 \otimes \cdots \otimes \Omega_m)^{-1}(\Gamma_1 \otimes \Gamma_2 \otimes \cdots \otimes \Gamma_m)\}$, which can be estimated by adapting the numerical algorithms in Section 2.2 and Section 3.2.

Background and proofs of all propositions and corollaries are provided in a supplement file. The R codes and the EEG data are also provided in supplement files. These files are available in the web-appendix of the online journal.

Acknowledgement

Research for this article was supported in part by National Science Foundation Grant DMS-1007547. The authors are grateful to the Editor, an associate editor, and two referees for their insightful suggestions and comments that help to improve the paper substantially. The authors also thank Bing Li and Min Kyung Kim for sharing their code for dimension folding SIR.

References

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

Blondin, D. (2007). Rates of strong uniform consistency for local least squares kernel regression estimators. *Statist. Probab. Lett.* **77**, 1526-1534.

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

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

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

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

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

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

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

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

Dutilleul, P. (1999). The MLE algorithm for the matrix normal distribution. *J. Statist. Comput. Simulation* **64**, 105-123.

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

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

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

Kolda, T. G. and Bader, B. W. (2009). Tensor decomposition and application. *SIAM Rev.* **51**, 455-500.

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

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

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

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

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

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

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

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

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

Yang, J., Zhang, D., Frangi, A. F. and Yang, J. (2004). Two dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **26**, 131-137.

Ye, J. (2005). Generalized low rank approximation of matrices. *Mac. Learn.* **61**, 167-191.

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

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

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

Zhu, L. and Ng, K. W. (1995). Asymptotics of sliced inverse regression. *Statist. Sinica* **5**, 727-736.

School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.

E-mail: dingx056@stat.umn.edu

School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.

E-mail: dennis@stat.umn.edu

(Received May 2012; accepted January 2013)