2-2020

# Sufficient Dimension Folding in Regression via Distance Covariance for Matrix-valued Predictors

Wenhui Sheng
*Marquette University*, wenhui.sheng@marquette.edu

Qingcong Yuan
*Miami University - Oxford*

*Statistical Analysis and Data Mining*, Vol. 13, No. 1 (February 2020): 71-82. DOI. 

# Sufficient Dimension Folding in Regression via Distance Covariance for Matrix-valued Predictors

Wenhui Sheng
Department of Mathematical and Statistical Sciences, Marquette University, Milwaukee, Wisconsin
Qingcong Yuan
Department of Statistics, Miami University, Oxford, Ohio

## Abstract

In modern data, when predictors are matrix/array-valued, building a reasonable model is much more difficult due to the complicate structure. However, dimension folding that reduces the predictor dimensions while keeps its structure is critical in helping to build a useful model. In this paper, we develop a new sufficient dimension folding method using distance covariance for regression in such a case. The method works efficiently without strict assumptions on the predictors. It is model-free and nonparametric, but neither smoothing techniques nor selection of tuning parameters is needed. Moreover, it works for both univariate and multivariate response cases. In addition, we propose a new

method of local search to estimate the structural dimensions. Simulations and real data analysis support the efficiency and effectiveness of the proposed method.

## Keywords

central dimension folding subspace, distance covariance, sufficient dimension folding

## INTRODUCTION

Modern scientific technology could produce data with complicated structures, such as matrix- or array-valued predictors, for example, EEG (electroencephalography) data, longitudinal data, and neuroimaging data. Building models for such kind of data is difficult, due to its matrix structure. Li et al [15] proposed sufficient dimension folding methods to analyze such kind of data and pointed out that there could be two benefits of preserving the original structure of predictors: keeping important aspects of interpretation on the data structure and greatly reducing the number of parameters in helping to build a model, and thus enhancing the accuracy of estimation. They developed three inverse dimension folding methods, folded-SIR, folded-SAVE, and folded-DR, along with the concept of central dimension folding subspace (CDFS). Pfeiffer et al [21] developed an alternative inverse dimension folding method for longitudinal data. Ding and Cook [6] proposed two inverse approaches for dimension folding: dimension folding principal component analysis and dimension folding principal fitted component (DF-PFC). Further, Ding and Cook [7] introduced a tensor sliced inverse regression approach and Ding and Cook [8] discussed recently developed sufficient dimension folding methods with a special focus on sliced inverse regression. On the other hand, Xue and Yin [[29]] and Xue et al [31] proposed forward dimension folding methods. There are other techniques analyzing matrix- or array-valued data set, such as matrix logistic regression [11] and tensor regression [32].

All of the above methods require certain conditions on the distribution of predictors or tuning parameters in smoothing approaches. In this article, we propose a new sufficient dimension folding method using distance covariance (DCOV; [26]). DCOV is a measure of dependence between two random vectors with arbitrary dimensions. DCOV has many advantages, for example, it equals to zero if and only if the random vectors are independent; it can measure both linear and nonlinear dependence. These advantages enable our method to perform well under different kinds of regression relations. There are other merits of the proposed method, such as it is model-free and nonparametric, but needs no smoothing technique; it does not require any particular distributions on predictors to have a fast computing. Further, it works well for both univariate and multivariate response.

The rest of the article is organized as follows. In Section 2, we introduce the new dimension folding method, its estimation and properties, along with a method of estimating the structural dimensions. In Section 3, we make comparisons among different dimension folding methods, and apply our method to two real data sets. Section 4 contains a brief remark about extending the method to array-valued predictors and a short summary of this article.

# METHODOLOGY

## Overview of CDFS

Let $X$ be a $p_l \times p_r$ random matrix and $Y$ be a scalar (or vector) response. LKA [15] proposed the concept of CDFS, and we assume that CDFS exists. Conditions for the existence of CDFS can be found in LKA [15]. To be self-contained, a brief introduction of the CDFS is given below.

For two subspaces $S_1$ and $S_2$ in $R^m$, let $S_1 \otimes S_2$ be the linear subspace spanned by the vectors $\{v_1 \otimes v_2 \colon v_1 \in S_1,\ v_2 \in S_2\}$, where $\otimes$ is the Kronecker product. Then the CDFS $(\mathcal{S}_{Y|^\circ X^\circ})$ is defined as $\mathcal{S}_{Y|X^\circ} \otimes \mathcal{S}_{Y|^\circ X}$, where $\mathcal{S}_{Y|X^\circ}$ (or $\mathcal{S}_{Y|^\circ X}$ ) is the intersection of all right (or left) dimension folding subspaces for $Y|X$. Here, the right dimension folding subspace, denoted by $S_r$, and the left dimension folding subspace, denoted by $S_l$, are defined as the subspaces satisfying the relation: $Y \perp\!\!\!\perp X \mid P_{S_l} X P_{S_r}$, where $P_{Sl}$ $(P_{Sr})$ is the projection matrix onto the subspace $S_l$ $(S_r)$.

Our aim is to find a basis matrix of $\mathcal{S}_{Y|^\circ X}$, denoted by $\alpha \in R^{p_l \times d_l}$, $d_l \leq p_l$ and a basis matrix of $\mathcal{S}_{Y|X^\circ}$, denoted by $\beta \in R^{p_r \times d_r}$, $d_r \leq p_r$, such that

$$Y \perp\!\!\!\perp X \mid \alpha^T X \beta,$$

(1)

here $d_r$ and $d_l$ are the structural dimensions. Thus, $\mathcal{S}_{Y|^\circ X^\circ} = \mathrm{Span}(\beta \otimes \alpha)$. Relation (1) is equivalent to $Y \perp\!\!\!\perp vec(X) \mid (\beta \otimes \alpha)^T vec(X)$, which indicates that the conventional central dimension reduction subspace $\mathcal{S}_{Y|vecX} \subseteq \mathcal{S}_{Y|^\circ X^\circ}$.

## Sufficient dimension folding via DCOV

In this article, we use DCOV to estimate CDFS. Suppose that $U \in R^p$ and $V \in R^q$, where $p$ and $q$ are positive integers, then DCOV between $U$ and $V$ is the nonnegative number, $\mathcal{V}U, V$ [25], which can be written as

$$\mathcal{V}^2 U, V = E \mid U - U' \| V - V' \mid + E \mid U - U' \mid E \mid V - V' \mid - E \mid U - U' \| V - V'' \mid - E \mid U - U'' \| V - V' \mid,$$

where $(U', V')$ and $(U'', V'')$ are *i.i.d* copies of $(U, V)$ and $|\cdot|$ is the Euclidean distance. Here, the dimensions of $U$ and $V$ can be arbitrary, and we assume that $E|U|, E|V| < \infty$. DCOV is equal to 0 if and only if $U$ and $V$ are independent. Such a property makes it valuable in many research fields including the sufficient dimension reduction [24]. Suppose that $a \in R^{p_l \times d_1}$ and $b \in R^{p_r \times d_2}$, then the squared DCOV between $Y$ and $(b \otimes a)^T vec(X)$ is then defined as

$$\begin{aligned}
&\mathcal{V}^2\big(Y, (b \otimes a)^T vec(X)\big) \\
&= E \mid Y - Y' \| (b \otimes a)^T \big(vec(X) - vec(X')\big) \mid \\
&+ E \mid Y - Y' \mid E \mid (b \otimes a)^T \big(vec(X) - vec(X')\big) \mid \\
&- E \mid Y - Y' \| (b \otimes a)^T \big(vec(X) - vec(X'')\big) \mid \\
&- E \mid Y - Y'' \| (b \otimes a)^T v\big(ec(X) - vec(X')\big) \mid,
\end{aligned}$$

where $(Y', X')$ and $(Y'', X'')$ are *i.i.d* copies of $(Y, X)$. The following result indicates how we might be able to find a basis for CDFS.

**Proposition 1**

*Suppose* $Y \perp\!\!\!\perp X \mid \alpha^T X\beta$, *where* $\alpha \in R^{p_l \times d_l}$ *is a basis of* $S_{Y|^\circ X}$ *and* $\alpha^T\alpha = I_{d_l}$ ; $\beta \in R^{p_r \times d_r}$ *is a basis of* $S_{Y|X^\circ}$ *and* $\beta^T\beta = I_{d_r}$, *then* $\beta \otimes \alpha$ *is a basis of the CDFS,* $S_{Y|^\circ X^\circ}$ . *Under the independence condition,* $P_{(\beta\otimes\alpha)}vec(X) \perp\!\!\!\perp Q_{(\beta\otimes\alpha)}vec(X)$, *we have* $\mathcal{V}^2\big(Y(\beta\otimes\alpha)^T\, vec(X)\big) \geq \big(\mathcal{V}^2 Y, (b \otimes a)^T vec(X)\big)$, *where* $a \in R^{p_l\times d_1}, a^T a = I_{d_1}, 1 \leq d_1 \leq p_l$ *and* $b \in R^{p_l\times d_2}, b^T b = I_{d_2}, 1 \leq d_2 \leq p_r$. *The equality holds if and only if* $b \otimes a$ *is also a basis of the CDFS.*

In Proposition 1, $P_{\beta\otimes\alpha} = (\beta\otimes\alpha)(\beta\otimes\alpha)^T$, $Q_{\beta\otimes\alpha} = I_{plpr} - P_{\beta\otimes\alpha}$, where $I_{plpr}$ is the identity matrix with dimension $p_l p_r$. This proposition indicates that we can obtain a basis of the CDFS by maximizing $\mathcal{V}^2\big(Y, (b\otimes a)^T vec(X)\big)$ with respect to $a$ and $b$ under the constraints $a^T a = I_{d_1}$ and $b^T b = I_{d_2}$, where $a \in R^{p_l\times d_1}$, $1 \leq d_1 \leq p_l$ and $b \in R^{p_r\times d_2}$, $1 \leq d_2 \leq p_r$. The independence condition $P_{(\beta\otimes\alpha)}vec(X) \perp\!\!\!\perp Q_{(\beta\otimes\alpha)}vec(X)$ in Proposition 1 is not as strong as it seems to be. As discussed in Sheng and Yin [23], it could be satisfied asymptotically when $p_l p_r$ is reasonably large. Therefore, a pair of basis matrices for $S_{Y|^\circ X}$ and $S_{Y|X^\circ}$, say, $\alpha$ and $\beta$, is the solution of (2).

$$(\alpha,\beta) = \arg\max_{a,b:\{a^T a = I_{d_1}, b^T b = I_{d_2} 1 \leq d_1 \leq p_l, 1 \leq d_2 \leq p_r\}} V\mathcal{V}^2\big(Y, (b\otimes a)^T vec(X)\big).$$

(2)

## Estimation of CDFS

Assume that the dimensions of $S_{Y|X^\circ}$ and $S_{Y|^\circ X}$, $d_r$ and $d_l$, are known, we introduce an algorithm to estimate the CDFS. In the following, let $\tilde{X}, \tilde{Y} = \{(X_i, Y_i),\ i = 1,\ldots,n\}$ denote a random sample from $(X, Y)$. The sample version of $\mathcal{V}^2\big(Y, (b\otimes a)^T vec(X)\big)$ is denoted by $\mathcal{V}n2Y\tilde{}b \otimes aTvecX\tilde{}$ with the form:

$$\mathcal{V}_n^2\left(\tilde{Y}, (b\otimes a)^T vec\big(\tilde{X}\big)\right) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}(b\otimes a)B_{kl},$$

(3)

where, for $k,\ l = 1,\ldots,n$,

$$\begin{aligned} &= a_{kl}(b\otimes a) - \bar{a}_{k\cdot}(b\otimes a) - \bar{a}_{\cdot l}(b\otimes a) + \bar{a}_{\cdot\cdot}(b\otimes a), \\ A_{kl}(b\otimes a) &= |\ (b\otimes a)^T vec(X_k - X_l)\ |, \bar{a}_{k\cdot}(b\otimes a) \\ a_{kl}(b\otimes a) & \\ \bar{a}_{\cdot l}(b\otimes a) &= \frac{1}{n}\sum_{l=1}^{n} a_{kl}(b\otimes a), \\ &= \frac{1}{n}\sum_{k=1}^{n} a_{kl}(b\otimes a), \bar{a}_{\cdot\cdot}(b\otimes a) = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}(b\otimes a)\,. \end{aligned}$$

Similarly, $b_{kl} = |Y_k - Y_l|$ and $B_{kl} = b_{kl} - \bar{b}_{k\cdot} - \bar{b}_{\cdot l} + \bar{b}_{\cdot\cdot}$ . An estimate of the basis of CDFS is $\beta_n \otimes \alpha_n$, where

$$(\alpha_n,\beta_n) = \arg\max_{a,b:\{a^T a = I_{d_l}, b^T b = I_{d_r}\}} \mathcal{V}_n^2\left(\tilde{Y}, (b\otimes a)^T vec\big(\tilde{X}\big)\right).$$

We estimate $\alpha_n$ and $\beta_n$ iteratively. An outline of the algorithm is as follows:

1. Obtain the initials: we first use sufficient dimension reduction method to get a basis, denoted as $\eta$, for the central subspace $S_{Y|vec(X)}$ [[4], [14]], whose dimension is $d_l d_r$. Then we use the two-step matrix decomposition technique ([29], Web Appendix C) to decompose $\eta$ as a Kronecker product such that $\eta = \beta_0 \otimes \alpha_0$, and $\alpha_0$ and $\beta_0$ are the initials. To be self-contained, we put the two-step matrix decomposition method in the Appendix. (Note that there are other ways to choose the initials, for example, we can use existing sufficient dimension folding method, such as folded-SIR, to get the initials.)

2. Let $\alpha_{(k)}$ and $\beta_{(k)}$ in the *k*th iteration, respectively. First maximize the following objective function with respect to *a*:

$$\mathcal{V}_n^2\left(\tilde{Y}, (\beta_{(k)} \otimes a)^T vec(\tilde{X})\right) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}(\beta_{(k)} \otimes a)B_{kl}.$$

Denote the estimate as $\alpha_{(k+1)}$. Next, maximize the objective function below with respect to $b$:

$$\mathcal{V}_n^2\left(\tilde{Y}, \left(b \otimes \alpha_{(k+1)}\right)^T vec(\tilde{X})\right) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}(b \otimes \alpha_{(k+1)})B_{kl}.$$

Denote the estimate as $\beta_{(k+1)}$. Then $\alpha_{(k+1)}$ and $\beta_{(k+1)}$ are the estimates of $a$ and $b$ in the $(k+1)$th iteration, respectively. Maximization is carried out iteratively through the general nonlinear programming function *fmincon*, available in MATLAB, that implements the interior-point approach. In the interior-point approach, the original constrained nonlinear program is replaced by a sequence of barrier subproblems, which are solved approximately by two powerful tools: sequential quadratic programming and trust region techniques. In this process, one of two main types of steps is used at each iteration: a direct step or a conjugate gradient step. By default, the algorithm tries a direct step first. If it cannot, it attempts a conjugate gradient step. Specifically, if the condition of nonconvexity and Hessian or Jacobian rank deficiencies exist, the direct step is replaced by the conjugate gradient step in order to obtain global convergence. More extensive descriptions about the interior-point approach are in refs. [[2]]; Waltz et al [27]. Our codes in MATLAB are available in the Supplement file.

3. Check convergence. Let $\tau_{(K)} = \beta_{(K)} \otimes \alpha_{(K)}$, $\tau_{(K+1)} = \beta_{(K+1)} \otimes \alpha_{(K+1)}$. If the Frobenius norm [17]: $\Delta f = \|\tau_{(k+1)}\tau_{(k+1)}^T - \tau_{(k)}\tau_{(k)}^T\|$ is smaller than the preset tolerance value, such as $10^{-6}$, then stop the iteration and set $\alpha_N = \alpha_{(K+1)}$ and $\beta_N = \beta_{(K+1)}$; otherwise, set $K :\equiv K + 1$ and go to step 2

## Asymptotic properties

We now establish the asymptotic results for our estimator.

**Proposition 2**

*Assume $\alpha \in R^{p_l \times d_l}$ and $\beta \in R^{p_r \times d_r}$ are basis matrices of $S_{Y|^\circ X}$ and $S_{Y|X^\circ}$, respectively, with $\alpha^T\alpha = I_{d_l}$ and $\beta^T\beta = I_{d_r}$. Suppose the support of X is compact, $E(Y) < \infty$ and $P_{(\beta \otimes \alpha)}vec(X) \perp\!\!\!\perp Q_{(\beta \otimes \alpha)}vec(X)$. Let*

$(\alpha_n, \beta_n) = \arg\max_{a^T a = I_{d_l}, b^T b = I_{d_r}} \mathcal{V}_n^2\left(\tilde{Y}, (b \otimes a)^T vec(\tilde{X})\right)$, *then there exists rotation matrices* $Q_1$ *and* $Q_2$, *such that* $\alpha_n \xrightarrow{P} \alpha Q_1$ *and* $\beta_n \xrightarrow{P} \beta Q_2$.

**Proposition 3**

*Assume* $\alpha \in R^{p_l \times d_l}$ *and* $\beta \in R^{p_r \times d_r}$ *are basis matrices of* $S_{Y|^\circ X}$ *and* $S_{Y|X^\circ}$, *respectively, with* $\alpha^T \alpha = I_{d_l}$ *and* $\beta^T \beta = I_{d_r}$ . *Suppose the support of* $X$ *is compact,* $E(Y) < \infty$ *and*$P_{(\beta \otimes \alpha)} vec(X) \perp\!\!\!\perp Q_{(\beta \otimes \alpha)} vec(X)$. *Let* $(\alpha_n, \beta_n) = \arg\max_{a^T a = I_{d_l}, b^T b = I_{d_r}} \mathcal{V}_n^2\left(\tilde{Y}, (b \otimes a)^T vec(\tilde{X})\right)$, *then under the regularity conditions in the Appendix*, *there exists rotation matrices* $Q_1$ *and* $Q_2$ *such that* $\sqrt{n}\left(\begin{pmatrix} vec(\alpha_n) \\ vec(\beta_n) \end{pmatrix} - \begin{pmatrix} vec(\alpha Q_1) \\ vec(\beta Q_2) \end{pmatrix}\right) \to N(0, V_1)$, *where* $V_1$ *is the covariance matrix defined in the Appendix*.

Proposition 2 shows the consistency of the estimators and Proposition 3 proves the root-n consistency and asymptotic normality of the estimators. In Propositions 2 and 3, the basis matrices $\alpha$ and $\beta$ are not unique, therefore $Q_1$ and $Q_2$ are needed, but the subspace spanned by $\beta \otimes \alpha$ is unique, hence its projection $(\beta \otimes \alpha)(\beta \otimes \alpha)^T$ is unique. In the corollary below, we give the asymptotic normality of $(\beta_n \otimes \alpha_n)(\beta_n \otimes \alpha_n)^T$, which avoids rotation matrices. Note that in Proposition 2, we can also get rid of the rotation matrices by using $\alpha_n \alpha_n^T \xrightarrow{P} \alpha\alpha^T$ and $\beta_n \beta_n^T \xrightarrow{P} \beta\beta^T$.

**Corollary 1**

*Under the same assumptions of Proposition 3,* $\sqrt{n}\left(vec(\beta_n \otimes \alpha_n)(\beta_n \otimes \alpha_n)\right)^T - vec\left((\beta \otimes \alpha)(\beta \otimes \alpha)\right)^T \to N(0, V_2)$, *where* $V_2$ *is the covariance matrix*.

## Estimating $d_l$ and $d_r$

In practice, we need to estimate the dimensions of $S_{Y|^\circ X}$ and $S_{Y|X^\circ}$ . We propose a new method via local search below.

Consider *m* nearest neighbor points of $(X_i,\ Y_i)$, where *m* is prespecified. Then, we use the idea of local linear approximation (see, eg, [[9], [28]]) and set the local basis of $S_{Y|^\circ X}$ and $S_{Y|X^\circ}$ to be a vector, respectively. Here the objective function is similar to the one in (2), but *a* and *b* are set as vectors. Denote their estimates as $\hat{\alpha}_{L_i}$ and $\hat{\beta}_{L_i}$, which are both vectors. Next we obtain the eigenvalues $\lambda_1^\alpha > \cdots > \lambda_{pl}^\alpha$ of $\sum_{i=1}^n \hat{\alpha}_{L_i} \hat{\alpha}_{L_i}^T$, and calculate the ratios $\lambda_j^\alpha / \lambda_{j+1}^\alpha$, for $j = 1, \ldots, p_l - 1$. Suppose the largest ratio is $\lambda_s^\alpha / \lambda_{s+1}^\alpha$, then the estimated dimension of $S_{Y|^\circ X}$ is *s*. Similarly, obtain the eigenvalues $\lambda_1^\beta > \cdots > \lambda_{p_r}^\beta$ of $\sum_{i=1}^n \hat{\beta}_{L_i} \hat{\beta}_{L_i}^T$, and calculate the ratios $\lambda_j^\beta / \lambda_{j+1}^\beta$, for $j = 1, \ldots, p_l - 1$. Suppose the largest ratio is $\lambda_t^\beta / \lambda_{t+1}^\beta$, then the estimated dimension of $S_{Y|X^\circ}$ is *t*. This maximal eigenvalue ratio criterion was suggested by Luo et al [18] and was also used by Li and Yin [16].

# NUMERICAL STUDIES

In this section, we compare the performance of our method (folded-DCOV) with existing dimension folding methods: folded-SIR [15], folded-MAVE [29], and DF-PFC [6]. Note that folded-SIR is an inverse approach and it is based on the algorithm of SIR [14], which is one of the most popular dimension reduction method; folded-MAVE is a forward approach based on rMAVE [28] which is a very efficient

method; DF-PFC is a newly proposed efficient dimension folding approach. Five models are considered. Models 1, 2, 3, and 4 have univariate responses, while model 5 has a multivariate response. For models 2, 3, and 5, since $S_{Y|vecX} = S_{Y|\circ X\circ}$, it is also reasonable to compare dimension folding methods with conventional sufficient dimension reduction methods. Thus, we add results of a conventional sufficient dimension reduction method developed with DCOV in Sheng and Yin [24], denoted as DCOV.

To evaluate the accuracy of the estimates, we use the Frobenius norm $\Delta_f = \Delta(\beta \otimes \alpha, \beta_n \otimes \alpha_n) = \|(\beta \otimes \alpha)(\beta \otimes \alpha)^T - (\beta_n \otimes \alpha_n)(\beta_n \otimes \alpha_n)^T\|$[17]. The smaller value of this norm indicates a better estimate. For each model, we consider two sizes of $X$, $5 \times 5$ and $7 \times 7$, using two sample sizes $n = 200$ and $n = 400$. We generate 100 datasets for each model and calculate the average of $\Delta_f s$ and the standard deviation (in the parenthesis). Let $e_i$ be a vector whose *i*th element is 1 and other elements are 0. In the settings of the five models, the independence condition in Proposition 1 of Section 2.2 is satisfied. We also study the performance of the proposed method when the independence condition is unsatisfied and the simulation results are reported in the Appendix.

*Model 1*: This is Example 1 in LKA [15]. Let $d_l = d_r = 2$ and $p_l = p_r = p = 5, 7$. The response $Y$ follows a Bernoulli distribution with success probability equal to 0.5. The conditional distribution of $X$ given $Y$ is multivariate normal with conditional mean

$$E(X \mid Y = 0) = 0_{p\times p}, E(X \mid Y = 1) = \begin{pmatrix} \mu I_2 & 0_{2\times(p-2)} \\ 0_{(p-2)\times 2} & 0_{(p-2)\times(p-2)} \end{pmatrix}$$

and conditional variance

$$var(x_{ij} \mid Y = 0) = \begin{cases} 0.1 & (i,j) \in A, \\ 1 & (i,j) \notin A, \end{cases}$$
$$var(x_{ij} \mid Y = 1) = \begin{cases} 1.5 & (i,j) \in A, \\ 1 & (i,j) \notin A, \end{cases}$$

where *A* is the index set $\{(1,2),(2,1)\}$. The model further assumes cov $x_{ij}x_{i',j'} = 0$ whenever $(i,j) \neq (i', j')$. In model 1, $S_{Y|vec(X)} = \text{Span}(e_1 \otimes e_2 + e_2 \otimes e_2, e_1 \otimes e_2, e_2 \otimes e_1)$ and $S_{Y|\circ X\circ} = \text{Span}(e_1 \otimes e_1, e_1 \otimes e_2, e_2 \otimes e_1, e_2 \otimes e_2)$, therefore $S_{Y|vec(X)}$ is a subspace of $S_{Y|\circ X\circ}$. In the simulation, we choose $\mu = 2$. Table shows the simulation performance for each method and folded-DCOV is the best under this model.

Table 1. Accuracy of estimation for model 1

| $(n, p_l, p_r)$ | Folded-DCOV | Folded-SIR | Folded-MAVE | DF-PFC[a] |
|---|---|---|---|---|
| (200, 5, 5) | 0.4171(0.0811) | 0.8558(0.1755) | 1.2969(0.4593) | 0.9364(0.7323) |
| (200, 7, 7) | 0.5520(0.1086) | 1.1546(0.1757) | 1.5289(0.3602) | 1.1825(0.6985) |
| (400, 5, 5) | 0.2828(0.0581) | 0.5988(0.1217) | 1.0091(0.4722) | 0.7973(0.7112) |
| (400, 7, 7) | 0.3778(0.0590) | 0.7800(0.1280) | 1.2336(0.4377) | 0.9802(0.7264) |

[a] In DF-PFC, $f(y_i) = I(y_i = 0) - n_0/n$, where $I(\cdot)$ is indicator function and $n_0$ is the number of 0 in *Y* (see ref. [6] for more details).

*Model 2*: This is Example 2 in LKA [15]. The only difference between model 1 and model 2 is that the index set *A* is set to be {(1, 1), (1, 2), (2, 1)} in model 2. In this case, $S_{Y|vec(X)} = S_{Y|\circ X\circ} = \text{Span}(e_1 \otimes e_1, e_1 \otimes$

$e_2, e_2 \otimes e_1, e_2 \otimes e_2$). The simulation results are presented in Table . Again, folded-DCOV outperforms other methods.

Table 2. Accuracy of estimation for model 2

| $(\boldsymbol{n}, \boldsymbol{p_l}, \boldsymbol{p_r})$ | **Folded-DCOV** | **Folded-SIR** | **Folded-MAVE** | **DF-PFC**[a] | **DCOV** |
|---|---|---|---|---|---|
| (200, 5, 5) | 0.3386(0.0784) | 0.7988(0.1607) | 1.2514(0.4920) | 0.9024(0.7154) | 1.2086(0.1544) |
| (200, 7, 7) | 0.4612(0.0773) | 1.1013(0.1845) | 1.4693(0.4535) | 1.1171(0.7259) | 1.6324(0.1811) |
| (400, 5, 5) | 0.2361(0.0502) | 0.5699(0.1244) | 1.0261(0.5149) | 0.8694(0.8083) | 0.8918(0.1712) |
| (400, 7, 7) | 0.3140(0.0596) | 0.7613(0.1181) | 1.2386(0.4948) | 0.9999(0.7361) | 1.2418(0.1347) |

[a] In DF-PFC, $f(y_i) = I(y_i = 0) - n_0/n$, where $I(\cdot)$ is indicator function and $n_0$ is the number of 0 in $Y$ (see ref. [6] for more details).

In models 3, 4, and 5, $vec(X) \sim N0, I_{p_l p_r}$. The random error $\epsilon \sim N(0,1)$ and is independent of $X$. Additional simulations for correlated predictors are given in the Appendix.

Model 3: $Y = 0.5\sin_{(x_{11})} + x_{12}\epsilon$.
Model 4: $Y = x_{11}(x_{12} + x_{21}) + x_{11} + x_{22} + 0.2\epsilon$.
Model 5: The response $Y = (y_1, y_2)^T$, where

$$\begin{cases} y_1 = x_{11}^2/(0.5 + (x_{21} + 1.5)^2) + 0.2\epsilon_1, \\ y_2 = x_{11}(x_{21} + 1) + 0.2\epsilon_2. \end{cases}$$

In model 3, one direction is in the variance function and $S_{Y|^\circ X^\circ} = S_{Y|vecX} = \mathrm{Span}(e_1 \otimes e_1, e_1 \otimes e_2)$. Table  indicates that folded-DCOV outperforms the other methods in estimating the CDFS. In model 4, all the directions are in the mean function. The CDFS $S_{Y|^\circ X^\circ} = \mathrm{Span}(e_1 \otimes e_1, e_1 \otimes e_2, e_2 \otimes e_1, e_2 \otimes e2)$, while $S_{Y|vec(X)} = \mathrm{Span}(e_1 \otimes e_1, e_1 \otimes e_2 + e_2 \otimes e_1, e_1 \otimes e_1 + e_2 \otimes e_2)$, therefore $S_{Y|vec(X)} \subset S_{Y|^\circ X^\circ}$. Table shows that folded-DCOV still performs very well. The average of $\Delta_f s$ for folded-DCOV is a little larger than folded-MAVE, which is not surprising since folded-MAVE focuses on dimensions in the mean function. However, the standard error of $\Delta_f s$ for folded-DCOV is much smaller than that of folded-MAVE. Model 5 has a multivariate response and $S_{Y|^\circ X^\circ} = S_{Y|vec(X)} = \mathrm{Span}(e_1 \otimes e_1, e_2 \otimes e_1)$. Table reports the estimation accuracy, and again, folded-DCOV performs very well. Note that in model 5, we do not compare with other dimension folding methods, because their methods are developed for univariate response.

Table 3. Accuracy of estimation for model 3

| $(\boldsymbol{n}, \boldsymbol{p_l}, \boldsymbol{p_r})$ | **Folded-DCOV** | **Folded-SIR** | **Folded-MAVE** | **DF-PFC**[a] | **DCOV** |
|---|---|---|---|---|---|
| (200, 5, 5) | 0.7581(0.4348) | 1.3851(0.2053) | 1.5814(0.3053) | 1.2239(0.3225) | 1.6307(0.0783) |
| (200, 7, 7) | 1.2230(0.4596) | 1.5601(0.1508) | 1.7762(0.2091) | 1.4933(0.2452) | 1.7976(0.0733) |
| (400, 5, 5) | 0.4924(0.3628) | 1.3097(0.1973) | 1.4258(0.3869) | 1.0125(0.2740) | 1.5303(0.0494) |
| (400, 7, 7) | 0.6645(0.3944) | 1.3805(0.1493) | 1.6307(0.2941) | 1.2642(0.2712) | 1.6566(0.0621) |

[a] In DF-PFC, $f(y_i) = \mathrm{diag}\,(y_i y_i^2 y_i^3 y_i^4)$ (see ref. [6] for more details).

Table 4. Accuracy of estimation for model 4

| $(\boldsymbol{n}, \boldsymbol{p_l}, \boldsymbol{p_r})$ | **Folded-DCOV** | **Folded-SIR** | **Folded-MAVE** | **DF-PFC**[a] |
|---|---|---|---|---|
| (200, 5, 5) | 0.6141(0.2358) | 0.8640(0.3780) | 0.4088(0.3397) | 1.1805(0.3926) |

| | | | | |
|---|---|---|---|---|
| (200, 7, 7) | 0.8114(0.1654) | 1.2842(0.4295) | 0.7045(0.4871) | 1.5287(0.4064) |
| (400, 5, 5) | 0.4144(0.0946) | 0.5887(0.1299) | 0.3914(0.4352) | 0.8160(0.2079) |
| (400, 7, 7) | 0.5470(0.0899) | 0.6753(0.2002) | 0.4658(0.3505) | 1.1034(0.3967) |

[a] In DF-PFC, $f(y_i) = \text{diag}\,(y_i y_i^2 y_i^3 y_i^4)$ (see ref. [6] for more details).

Table 5. Accuracy of estimation for model 5

| **(n, $p_l$, $p_r$)** | **(200, 5, 5)** | **(200, 7, 7)** | **(400, 5, 5)** | **(400, 7, 7)** |
|---|---|---|---|---|
| Folded-DCOV | 0.2142(0.0574) | 0.2910(0.0535) | 0.1503(0.0375) | 0.1988(0.0394) |
| DCOV | 1.4427(0.0136) | 1.5184(0.0666) | 1.4273(0.0071) | 1.4480(0.0105) |

*Estimating* $(d_r,\ d_l)$: We use model 4 to illustrate the performance of the local search method to estimate the dimensions. Let $(n,\ p_l,\ p_r)\ =\ (400, 5, 5)$. For the local search method, we select different values of *m* to see whether the method is sensitive to the choice of *m*.
Let $m\ =\ 60, 70, 80, 100$ and the corresponding percentage of correct estimates are 90%, 88%, 89%, and 95%. The simulation results indicate that this method is quite reasonable. Note that selecting *m* can be tricky as it cannot be too large due to local approximation, or too small due to effective estimation. A rule-of-thumb for *m* is approximately 3-8 × the number of parameters.

*Comparison of computing time*: Based on model 2 and setting $(n,\ p_l,\ p_r)\ =\ (200, 5, 5)$, Table compares the computing time of different dimension folding methods. All simulations were done on the same computer. On the average, our method takes 5.2 seconds for one simulation, which is acceptable. Indeed, folded-SIR and PF-DFC are faster than our method, but our method achieves much better accuracy under model 2. On the other hand, the local method folded-MAVE is slow as expected.

Table 6. CPU time in seconds for 10 simulations

| **Folded-DCOV** | **Folded-SIR** | **Folded-MAVE** | **PF-DFC** |
|---|---|---|---|
| 52 | 2 | 243 | 20 |

*The primary biliary cirrhosis data*: This data is from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. It contains records of many clinical, biochemical, serologic, and histologic measurements for each of 312 patients. These measurements were recorded during the patients' regular visits to the Mayo Clinic after their initial referral. More information can be found in Fleming and Harrington [10]. Many research papers have studied the relationship between different biomarkers and the survival time of PBC patients based on this data, for example, Murtaugh et al [20], Kim et al [12], and Albert and Shih [1]. Xue and Yin [29] also studied this data, but unlike the previous research, they studied it from the aspect of dimension folding. In the data, the measurements of multiple markers for each patient were observed over different times and it is desirable to keep the time structure in this longitudinal data. Xue and Yin [29] considered time as one fold of the predictor and the prognostic variables as another fold of the predictor, thus the predictor is a matrix.

We analyze the data from the angle of dimension folding using the proposed method, but differs from Xue and Yin [29]: they only considered three prognostic variables, we study one more variable, the status of edema, which was already proved to be an important factor for predicting survival for PBC patient in many research papers [[5], [12], [20]]. We define the time fold of the matrix predictor the

same way as in Xue and Yin [29]: the visits between day 90 and day 270 from the enrollment are classified as at the time point 6 months, and visits between day 270 and day 550, visits between day 550 and day 910, visits between day 910 and day 1275 from the enrollments are identified as time point 1 year, 2 years, and 3 years, respectively. Another fold of the matrix predictor is the prognostic variables, which are the status of edema, bilirubin, albumin level and prothrombin time, thus the predictor is a 4 × 4 matrix with columns corresponding to time points and rows corresponding to prognostic variables. Among the prognostic variables, status of edema is an ordinal variable, which is coded as 0 for no edema and no diuretic therapy; 0.5 for edema present without diuretic or edema resolved by diuretic; and 1 for edema despite diuretic therapy. The other three variables are continuous. The response is the time in years between registration and the earlier of transplantation or death.

The proposed local search method suggests that $(d_r,\ d_l)\ =\ (1, 1)$. The respective estimated basis for $S_{Y|^\circ X}$ and $S_{Y|X^\circ}$ are $\alpha_n^T = (0.1281, 0.9804, -0.1451, 0.0362)$ and $\beta_n^T = (-0.3861 - 0.3951 - 0.4983 - 0.6683)$. The top panel of Figure shows the relationship between log(response) versus the reduced predictor $(\alpha_n^T X \beta_n)$. Since all the values in the matrix predictor $X$ are positive, taking the coefficient signs in $\alpha_n$ and $\beta_n$ into account, Figure implies that albumin level has a positive relation with response $Y$, the transplant-free or survival time, while bilirubin level, prothrombin time and status of edema have negative relation with the response. All of these findings are consistent with the medical outcome [[5], [20]]. We also add a weighted least squares fit between log(response) and the reduced predictor on the top panel of Figure. The resulting residual plot on the bottom panel of Figure indicates that the model fit is well.

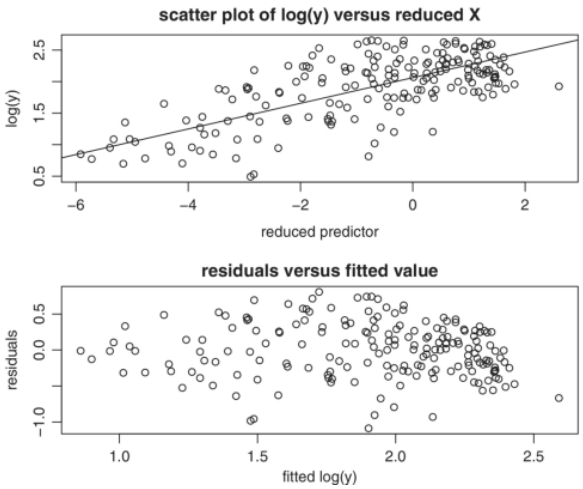

Figure 1. Summary plot and residual plot

*EEG data*: We also analyze the EEG data that was used in LKA [15]. The primary purpose of this study is to explore the association between alcoholism and the pattern of voltage values over times and channels.

The data consists of 122 observations, denoted as $(X_1,\ Y_1), \ldots, (X_{122},\ Y_{122})$, where $X_i$ is a $256\ \times\ 64$ matrix and $Y_i$ is a binary univariate variable indicating whether the *i*th subject is alcoholic $(Y_i\ =\ 0)$ or nonalcoholic $(Y_i\ =\ 1)$. LKA [15] used a prescreening procedure on the original predictors first, and

then estimated the reduced predictors. We use their prescreened data set $(X_1^*, Y_1), \ldots, (X_{122}^*, Y_{122})$, where $X_1^*$ is a $15 \times 15$ matrix, and then we apply the method in Section 2.5 to estimate the structural dimensions: the estimated result is $(d_l, d_r) = (1, 1)$ when choosing $m = 40, m = 45$, or $m = 50$. Considering the sample size is 122, the value of *m* cannot be too large, because the method of estimating dimensions is a local search approach. Under the set up $(d_l, d_r) = (1, 1)$, we apply folded-DCOV to obtain the reduced predictors. Finally, we classify the data based on our reduced predictors using leave-one-out cross validation and quadratic discriminant analysis. We correctly classify 94 out of 122 cases. In comparison, folded-SIR correctly identifies 89 out of 122 cases under $(p_l, p_r, d_l, d_r) = (15, 15, 1, 1)$ and the conventional SIR provided 76 out of 122 correct decisions under $(p_l, p_r, d) = (15, 15, 1)$.

# DISCUSSION

We propose a new sufficient dimension folding method for regression with matrix-valued predictors in this article. The proposed method works well with both univariate and multivariate response. Its advantage is supported by simulation studies and applications in real data. Note that in the PBC data, some response values are censored at the time of transplantation. Although we ignore the censored variable in the analysis, the results are reasonable. Thus, we consider extending the folding method to censored data, which could be an interesting future topic.

We can also extend the proposed method to array-valued predictors ([29], section 6). Let $\mathbf{X} = \{X_{j_1,\ldots,j_k} : j_1 = 1 \ldots p_1 \ldots j_k = 1 \ldots p_k\}$ be a k-way random array of dimension $p_1 \times \cdots \times p_k$ and let $Y$ be a scalar random response. Then a basis of the CDFS can be estimated by maximizing $V^2 Y, \left( \left( a^{(k)} \otimes \cdots \otimes a^{(1)} \right)^T vec(\mathbf{X}) \right)$ with respect to $a^{(k)}, \ldots, a^{(1)}$ under the constraints $a^{(1)T} a^{(1)} = I_{d_1}, \ldots, a^{(k)T} a^{(k)} = I_{d_k}$. Here we need to point out that although the proposed method is straightforward to be extended to array-valued predictor, the algorithm should be further polished in the extension, for example, the method to choose initial values should be changed. Also, as the dimension of the predictor increases, more computation are involved in the algorithm, which will slow down the computational speed.

# ACKNOWLEDGMENTS

We thank Dr. Bing Li and Dr. Shanshan Ding for sharing their data and codes. We are thankful to Editor, Associate Editor, and two referees their careful and insightful reviews that led to considerable improvement of this paper.

# APPENDIX A

## Proof of Proposition 1

The proof of Proposition 1 is similar to the proofs of Propositions 1 and 2 in Sheng and Yin [24]. Therefore, we omit some similar details and just mention important differences. We can prove the Proposition 1 under two cases: one is $\mathrm{Span}(b \otimes a) \subseteq \mathrm{Span}(\beta \otimes \alpha)$ and another is $\mathrm{Span}(b \otimes a) \nsubseteq Span(\beta \otimes \alpha)$. In both cases, we need to apply the Lemma A.1 in Sheng and Yin [24]. In that lemma, the authors considered the partition of a basis of the central subspace, while in this article, we

need to consider the partition of a basis of the CDFS, for example, the partition of $\beta \otimes \alpha$ and we denote the partition as $\beta_{s_1} \otimes \alpha_{s_2}, \beta_{\overline{s}_1} \otimes \alpha_{\overline{s}_2}$. Here, we need to emphasize that the partition $\beta_{s_1} \otimes \alpha_{s_2}, \beta_{\overline{s}_1} \otimes \alpha_{\overline{s}_2}$ must be meaningful in the setup of dimension folding problem. Here the "meaningful" partition means the $x'_{ij}s$ in both $\left(\beta_{s_1} \otimes \alpha_{s_2}\right)^T vec(X)$ and $\left(\beta_{\overline{s}_1} \otimes \alpha_{\overline{s}_2}\right)^T vec(X)$ can form a submatrix of $X$. The reason is that we aim to find the minimum matrix/vector $\alpha^T X\beta$ such that $Y \perp\!\!\!\perp X \mid \alpha^T X\beta$, therefore we need to keep the matrix/vector structure. The other parts of the proofs can be done with a similar logic that being used in the proofs of Propositions 1 and 2 in Sheng and Yin [24].

## Proof of consistency

**1 Lemma**

*If support of X is compact,* $E|Y| < \infty$ *and* $\alpha_n \xrightarrow{P} \alpha, \beta_n \xrightarrow{P} \beta$, *then* $V_n^2\left(\tilde{Y}(\beta_n \otimes \alpha_n)^T vec(\tilde{X})\right) - V_n^2\left(\tilde{Y}(\beta \otimes \alpha)^T\, vec(\tilde{X})\right) \xrightarrow{P} 0.$

**Proof of Lemma 1**

The proof is similar to the proof of Lemma A in the online supplementary material in Sheng and Yin [24], so we omit the details here.

**Proof of Proposition 2**

Without loss of generality, we assume $Q_1 = I_{dl}$ and $Q_2 = I_{dr}$. Suppose $\alpha_n$ is not a consistent estimator of $S_{Y|^\circ X}$ and $\beta_n$ is not a consistent estimator of $S_{Y|X^\circ}$, then there exists subsequences, still to be indexed by $n$, such that $\alpha_n \xrightarrow{P} \alpha^*$ and $\beta_n \xrightarrow{P} \beta^*$, but $\mathrm{Span}(\alpha^*) \neq \mathrm{Span}(\alpha)$, and $\mathrm{Span}(\beta^*) \neq \mathrm{Span}(\beta)$.

By Lemma, we have $V_n^2\left(\tilde{Y}, (\beta_n \otimes \alpha_n)^T vec(X)\right) - V_n^2\left(\tilde{Y}, (\beta^* \otimes \alpha^*)^T vec(\tilde{X})\right) \xrightarrow{P} 0$ and by Theorem 2 in Székely et al [26], we have $V_n^2\left(Y, (\beta^* \otimes \alpha^*)^T vec(\tilde{X})\right) \xrightarrow{a.s.} V^2\left(Y, (\beta^* \otimes \alpha^*)^T vec(X)\right)$, therefore $V_n^2\left(\tilde{Y}, (\beta_n \otimes \alpha_n)^T vec(\tilde{X})\right) \xrightarrow{P} V^2\left(Y, (\beta^* \otimes \alpha^*)^T vec(X)\right)$.

On the other hand, because $(\alpha_n, \beta_n) = \arg\max_{a^T a = I_{d_l}, b^T b = I_{d_r}} V_n^2\left(\tilde{Y}, (b \otimes a)^T vec(\tilde{X})\right)$, we have that $V_n^2\left(\tilde{Y}, (\beta_n \otimes \alpha_n)^T vec(\tilde{X})\right) \geq V_n^2\left(\tilde{Y}, (\beta \otimes \alpha)^T vec(\tilde{X})\right)$. If we take limit on both sides of the above inequality, we get $V^2\left(Y, (\beta^* \otimes \alpha^*)^T vec(X)\right) \geq V^2\left(Y, (\beta \otimes \alpha)^T vec(X)\right)$, however, we have proved that under the assumption $P_{(\beta\otimes\alpha)}^T vec(X) \perp\!\!\!\perp Q_{(\beta\otimes\alpha)}^T vec(X)$, $(\alpha, \beta) = \arg\max_{a^T a = I_{d_l}, b^T b = I_{d_r}} V^2\left(Y, (b \otimes a)^T vec(X)\right)$, and further because we assume that the spaces $S_{Y|^\circ X}$ and $S_{Y|X^\circ}$ are unique, so there are conflicts, hence $\alpha_n$ and $\beta_n$ are consistent estimators of the basis matrices of $S_{Y|^\circ X}$ and $S_{Y|X^\circ}$ .

## Proof of $\sqrt{n}$-consistency

In the proof, we need derivatives of $V^2\left(Y, (\beta \otimes \alpha)^T vec(X)\right)$ and $V_n^2\left(\tilde{Y}, (\beta \otimes \alpha)^T vec(\tilde{X})\right)$ with respect to $vec(\alpha)$ and $vec(\beta)$. For the simplicity of notation, we use $V'(\alpha), V'(\beta)$ denote the first derivatives of $V^2\left(Y, (\beta \otimes \alpha)^T vec(X)\right)$ with respect to $vec(\alpha)$ and $vec(\beta)$ respectively and we

use $V_n'(\alpha), V_n'(\beta)$ denote the first derivatives of $V_n^2\left(\tilde{Y}, (\beta \otimes \alpha)^T vec(\tilde{X})\right)$ with respect to $vec(\alpha)$ and $vec(\beta)$ respectively. For the notations of higher order derivatives, we use similar pattern.

Some additional notations are as follows: $I_{(m,n)}$ denotes a vec-permuation matrix. $I_m$ is an identity matrix with rank $m$. $A \otimes B$ denotes Kronecker product between matrices $A$ and $B$. $vec(\cdot)$ is a vec operator. Furthermore, we give the following definition and assumptions.

**Definition 1**

Let $\Delta(\beta \otimes \alpha) = \{b \otimes a: ||b \otimes a - \beta \otimes \alpha|| \leq c\}$ where $b$ is a $p_r \times d_r$ matrix and $a$ is a $p_l \times d_l$ matrix and $a^T a = I_{d_l}, b^T b = I_{d_r}$, $c$ is a fixed small constant, $||\cdot||$ is the Frobenius norm. We define an indicator function

$$\rho(X, X') = \begin{cases} 0 & \text{if } \|a^T(X - X')b\| \leq \varepsilon_0, \text{for } b \otimes a \in \Delta(\beta \otimes \alpha) \\ 1 & \text{if } \|a^T(X - X')b\| \leq \varepsilon_0, \text{for } b \otimes a \in \Delta(\beta \otimes \alpha)' \end{cases}$$

where $X'$ is an i.i.d. copy of $X$ and $\epsilon_0$ is a small number. We define the second and third derivatives of $\mathcal{V}^2\left(Y, (\beta \otimes \alpha)^T vec(X)\right)$ with respect to $vec(\alpha)$ and $vec(\beta)$ as $\mathcal{V}''(\alpha)\rho(X, X')$,$\mathcal{V}''(\beta)\rho(X, X')$,$\mathcal{V}'''(\alpha)\rho(X, X')$ and $\mathcal{V}'''(\beta)\rho(X, X')$. For the simplicity of notation, we will still use $\mathcal{V}''(\alpha)$, $\mathcal{V}''(\beta)$, $\mathcal{V}'''(\alpha)$ and $\mathcal{V}'''(\beta)$.

The reason we use this definition is that under Definition, the second and third derivatives of $\mathcal{V}^2\left(Y, (\beta \otimes \alpha)^T vec(X)\right)$ and $\mathcal{V}_n^2\left(\tilde{Y}, (\beta \otimes \alpha)^T vec(\tilde{X})\right)$ are bounded, near the neighborhood of the CDFS.

**Assumption 1**

$Var(\Phi^{(i)}) < \infty$, $i = 1, 2, 3$. $Var(\Psi^{(i)}) < \infty$, $i = 1, 2, 3$. Here

$$\Phi^{(1)}\left((X_1 Y_1)(X_2 Y_2)\right)$$

$$= \frac{I_{(d_l p_l)}^T\left((X_1 - X_2)\beta \otimes I_{d_l}\right)}{\left|\left(\beta^T(X_1 - X_2)^T \otimes I_{d_l}\right) I_{(d_l p_l)} vec\alpha\right|} |Y_1 - Y_2|,$$

$$\Phi^{(2)}\left((X_1 Y_1)(X_2 Y_2)(X_3 Y_3)\right)$$

$$= \frac{1}{6}\sum_{P_1} \frac{I_{(d_l p_l)}^T\left(\left(X_{i_1} - X_{i_2}\right)\beta \otimes I_{d_l}\right)}{\left|\left(\beta^T\left(X_{i_1} - X_{i_2}\right)^T \otimes I_{d_l}\right) I_{(d_l p_l)} vec\alpha\right|} \left|Y_{i_1} - Y_{i_3}\right|$$

$$\Phi^{(3)}\left((X_1 Y_1)(X_2 Y_2)(X_3 Y_3)(X_4 Y_4)\right)$$

$$= \left\{ \begin{array}{c} I_{(d_l p_l)}^T \left( (X_{i_1} - X_{i_2})\beta \otimes I_{d_l} \right) \\ \frac{1}{24} \sum_{P_2} \frac{\left( \beta^T (X_{i_1} - X_{i_2})^T \otimes I_{d_l} \right) I_{(d_l p_l)} vec\alpha}{\left| \left( \beta^T (X_{i_1} - X_{i_2})^T \otimes I_{d_l} \right) I_{(d_l p_l)} vec\alpha \right|} |Y_{i_3} - Y_{i_4}| \end{array} \right\},$$

$$\Psi^{(1)}\left((X_1 Y_1)(X_2 Y_2)\right)$$

$$\left( I_{d_r}^T \otimes ((X_1 - X_2)^T \alpha) \right)$$

$$= \frac{\left( I_{d_r} \otimes \left( \alpha^T (X_1 - X_2) \right) \right) vec\beta}{\left| \left( I_{d_r} \otimes \left( \alpha^T (X_1 - X_2) \right) \right) vec\beta \right|} |Y_1 - Y_2|,$$

$$\Psi^{(2)}\left((X_1 Y_1)(X_2 Y_2)(X_3 Y_3)\right)$$

$$\left( I_{d_r}^T \otimes \left( (X_{i_1} - X_{i_2})^T \alpha \right) \right)$$

$$= \frac{1}{6} \sum_{P_1} \frac{\left( I_{d_r} \otimes \left( \alpha^T (X_{i_1} - X_{i_2}) \right) \right) vec\beta}{\left| \left( I_{d_r} \otimes \left( \alpha^T (X_{i_1} - X_{i_2}) \right) \right) vec\beta \right|} |Y_{i_1} - Y_{i_3}|,$$

$$\Psi^{(3)}\left((X_1 Y_1)(X_2 Y_2)(X_3 Y_3)(X_4 Y_4)\right)$$

$$\left( I_{d_r}^T \otimes \left( (X_{i_1} - X_{i_2})^T \alpha \right) \right)$$

$$= \frac{1}{6} \sum_{P_2} \frac{\left( I_{d_r} \otimes \left( \alpha^T (X_{i_1} - X_{i_2}) \right) \right) vec\beta}{\left| \left( I_{d_r} \otimes \left( \alpha^T (X_{i_1} - X_{i_2}) \right) \right) vec\beta \right|} |Y_{i_3} - Y_{i_4}|$$

and $\sum_{P_1}$ denotes summation over the 3! permutations $(i_1, i_2, i_3)$ of (1, 2, 3), $\sum_{P_2}$ denotes summation over the 4! permutations $(i_1, i_2, i_3, i_4)$ of (1, 2, 3, 4).

**Assumption 2**

The second derivative of $\mathcal{L}n(\zeta)$ at $\zeta = \theta$ is nonsingular. The specific expression of $\mathcal{L}n(\zeta)$, $\zeta$ and $\theta$ are introduced below in the proof.

Assumption 1 is needed by the asymptotic properties of U-statistics ([13], Ch. 6).

Assumption 2 is in the spirit of von Mises proposition ([22], section 6.1). In this proposition, it claims that if the first nonvanishing term of Taylor expansion is the linear term, then the root-n consistency of the differentiable statistical function can be achieved. In our case, we assume the corresponding matrix is nonsingular, which guarantees the root-n consistency. If the matrix is singular, then $n$ or higher order consistency of some parts of our estimates can be proved.

**Proof of Proposition 3**

The proof of Proposition 3 is under the Assumptions 1 and 2.

For a random sample $\tilde{X}, \tilde{Y} = \{(X_k, Y_k),\ k = 1, \ldots, n\}$ from the joint distribution of random matrix $X \in R^{p_l \times p_r}$ and $Y \in R$. We consider the Lagrange function $\mathcal{L}(\zeta) = \mathcal{V}^2\big(Y, (b \otimes a)^T vec(X)\big) + \lambda_1^T \left(vec(a^T a) - vec\big(I_{d_l}\big)\right) + \lambda_2^T\, vec(b^T b) - vec\big(I_{d_r}\big)$. Here $\zeta = \begin{pmatrix} vec(a) \\ vec(b) \\ \lambda_1 \\ \lambda_2 \end{pmatrix} \in \mathbb{R}^{p_l d_l + p_r d_r + d_l^2 + d_r^2}$, where $a \in R^{p_l \times d_l}, b \in R^{p_r \times d_r}, \lambda_1 \in R^{d_l^2}, \lambda_2 \in R^{d_r^2}$. Let $\beta \otimes \alpha$ be a basis of CDFS, then under the independence condition $P_{(\beta \otimes \alpha)}^T vec(X) \perp\!\!\!\perp Q_{(\beta \otimes \alpha)}^T vec(X)$, there exists a $\lambda_{10},\ \lambda_{20}$, such that $\theta = \begin{pmatrix} vec(\alpha) \\ vec(\beta) \\ \lambda_1 \\ \lambda_2 \end{pmatrix}$ is a stationary point for $\mathcal{L}(\zeta)$, thus $\mathcal{L}'(\theta) = 0$. Let $\alpha_n$ and $\beta_n$ be the maximizer of $\mathcal{V}_n^2\left(\tilde{Y}(b \otimes a)^T vec\big(\tilde{X}\big)\right)$ under the constraints $a^T a = I_{d_l}$ and $b^T b = I_{d_r}$, then there exists a $\lambda_{1n}$ and $\lambda_{2n}$ such that $\begin{pmatrix} vec(\alpha_n) \\ vec(\beta_n) \\ \lambda_{1n} \\ \lambda_{2n} \end{pmatrix}$ is a stationary point for $\mathcal{L}_n\zeta = \mathcal{V}_n^2\left(\tilde{Y}(b \otimes a)^T vec\big(\tilde{X}\big)\right) + \lambda_1^T vec(a^T a) - vec\big(I_{d_l}\big) + \lambda_2^T vec(b^T b) - vec\big(I_{d_r}\big)$. Let $\theta_n = \begin{pmatrix} vec(\alpha_n) \\ vec(\beta_n) \\ \lambda_{1n} \\ \lambda_{2n} \end{pmatrix}$, then $\mathcal{L}_n'(\theta_n) = 0$. Without loss of generality, according to Proposition 2, we assume that $\alpha_n \xrightarrow{P} \alpha$ and $\beta_n \xrightarrow{P} \beta$, thus $\theta_n \xrightarrow{P} \theta$.

The Taylor expansion of $\mathcal{L}_n'(\theta_n)$ at $\theta$ is $0 = \mathcal{L}_n'(\theta_n) = \mathcal{L}_n'(\theta) + \mathcal{L}_n''(\theta)(\theta_n - \theta) + \mathcal{R}_1(\theta_n^*)$, where $\mathcal{L}_n'(\theta_n) = \begin{pmatrix} \mathcal{V}_n'(\alpha) + \big(I_{d_l} \otimes \alpha\big) I_{d_l^2} + I_{(d_l, d_l)}^T \lambda_{10} \\ \mathcal{V}_n'(\beta) + \big(I_{d_r} \otimes \beta\big) I_{d_r^2} + I_{(d_r, d_r)}^T \lambda_{20} \\ vec(\alpha^T \alpha) - vec\big(I_{d_l}\big) \\ vec(\beta^T \beta) - vec\big(I_{d_r}\big) \end{pmatrix}$ and the remainder term $\mathcal{R}_1(\theta_n^*)$ involves the third derivative of $\mathcal{L}_n(\zeta)$ at $\theta_n^*$, which has the form

$$\mathcal{R}_1(\theta_n^*) = \frac{1}{2}\begin{pmatrix} (\theta_n - \theta)^T T_n(1, :, :)(\theta_n - \theta) \\ (\theta_n - \theta)^T T_n(2, :, :)(\theta_n - \theta) \\ \vdots \\ (\theta_n - \theta)^T T_n(p_l d_l + p_r d_r + d_l^2 + d_r^2, :, :)(\theta_n - \theta) \end{pmatrix}.$$

where $T_n = \mathcal{L}_n'''(\theta_n^*), \theta_n^* = \begin{pmatrix} vec(\alpha_n^*) \\ vec(\beta_n^*) \\ \lambda_{1n}^* \\ \lambda_{2n}^* \end{pmatrix}$, $\|\theta_n^* - \theta\| \le \|\theta_n - \theta\|$, and $\|\cdot\|$ is the Frobenius norm.

Therefore, $-\left(\mathcal{L}_n''(\theta)\right)^{-1}\sqrt{n}\mathcal{L}_n'(\theta) = I + \frac{1}{2}\left(\mathcal{L}_n''(\theta)\right)^{-1} \times$
$\begin{pmatrix} (\theta_n-\theta)^T T_n(1,:,:) \\ (\theta_n-\theta)^T T_n(2,:,:) \\ \vdots \\ (\theta_n-\theta)^T T_n(p_l d_l + p_r d_r + d_l^2 + d_r^2,:,:) \end{pmatrix} \times \sqrt{n}(\theta_n-\theta).$

By Slutsky's theorem, we know $\sqrt{n}(\theta_n-\theta) \underset{=}{D} \left[ I + \frac{1}{2}\left(\mathcal{L}_n''(\theta)\right)^{-1} \times \right.$
$\left. \begin{pmatrix} (\theta_n-\theta)^T T_n(1,:,:) \\ (\theta_n-\theta)^T T_n(2,:,:) \\ \vdots \\ (\theta_n-\theta)^T T_n(p_l d_l + p_r d_r + d_l^2 + d_r^2,:,:) \end{pmatrix} \right] \times \sqrt{n}(\theta_n-\theta)$, therefore we have $-\left(\mathcal{L}_n''(\theta)\right)^{-1}\sqrt{n}\left(\mathcal{L}_n'(\theta)\right)^{-1} \underset{=}{D} \sqrt{n}(\theta_n-\theta)$. Further, because of $\mathcal{L}'(\theta) = 0$, we have $\left(I_{d_l} \otimes \alpha\right)\left(I_{d_l^2} + I_{d_l d_l}^T\right)\lambda_{10} = -\mathcal{V}'(\alpha), \left(I_{d_r} \otimes \beta\right)\left(I_{d_r^2} + I_{d_r d_r}^T\right)\lambda_{20} = -\mathcal{V}'(\beta)$, $vec(\alpha^T\alpha) - vec\left(I_{d_l}\right) = 0$ and $vec(\beta^T\beta) - vec\left(I_{d_r}\right) = 0$, therefore it is sufficient to study the relationship below:

$$-(\mathcal{L}_{11n})^{-1}\sqrt{n}\begin{pmatrix} \mathcal{V}_n'(\alpha) - \mathcal{V}'(\alpha) \\ \mathcal{V}_n'(\beta) - \mathcal{V}'(\beta) \end{pmatrix} \underset{=}{D} \sqrt{n}\begin{pmatrix} vec(\alpha_n) - vec(\alpha) \\ vec(\beta_n) - vec(\beta) \end{pmatrix},$$

where $(\mathcal{L}_{11n})^{-1}$ denotes the upper-left corner submatrix of $\left(\mathcal{L}_n''(\theta)\right)^{-1}$, which consists of the first $p_l d_l + p_r d_r$ rows and the first $p_l d_l + p_r d_r$ columns of $\left(\mathcal{L}_n''(\theta)\right)^{-1}$.

With some calculation, we can show both $\mathcal{V}_n'(\alpha)$ and $\mathcal{V}_n'(\beta)$ are linear combinations of U-statistics, that is, $\mathcal{V}_n'(\alpha) = \frac{(n-1)\left(n^2-2n+2\right)}{n^3}U_{1n}(\alpha) - \frac{2(n-1)(n-2)^2}{n^3}U_{2n}(\alpha) + \frac{(n-1)(n-2)(n-3)}{n^3}U_{3n}(\alpha)$ and $\mathcal{V}_n'(\beta) = \frac{(n-1)\left(n^2-2n+2\right)}{n^3}U_{1n}(\beta) - \frac{2(n-1)(n-2)^2}{n^3}U_{2n}(\beta) + \frac{(n-1)(n-2)(n-3)}{n^3}U_{3n}(\beta)$. Furthermore, according to Theorem 6.1.6 ([13], Ch. 6),

$$\sqrt{n}\begin{pmatrix} U_{1n}(\alpha) - \mu_1(\alpha) \\ U_{2n}(\alpha) - \mu_2(\alpha) \\ U_{3n}(\alpha) - \mu_3(\alpha) \\ U_{1n}(\beta) - \mu_1(\beta) \\ U_{2n}(\beta) - \mu_2(\beta) \\ U_{3n}(\beta) - \mu_3(\beta) \end{pmatrix} \xrightarrow{D} N(0, \sum),$$

where

$$U_{1n}(\alpha) = \binom{n}{2}^{-1} \sum_{1\le k<l\le n} \times \frac{I_{(d_l p_l)}^T\left((X_k - X_l)\beta \otimes I_{d_l}\right)\left(\beta^T(X_k - X_l)^T \otimes I_{d_l}\right)I_{(d_l p_l)}vec\alpha}{\left|\left(\beta^T(X_k - X_l)^T \otimes I_{d_l}\right)I_{(d_l p_l)}vec\alpha\right|}|Y_k - Y_l|,$$

$$U_{2n}(\alpha)=\binom{n}{3}^{-1}\sum_{1\le k<l<m\le n}\times\left\{\frac{1}{6}\sum_{P_1}\frac{I_{(d_lp_l)}^T\left((X_{i_1}-X_{i_2})\beta\otimes I_{d_l}\right)\left(\beta^T\left(X_{i_1}-X_{i_2}\right)^T\otimes I_{d_l}\right)I_{(d_lp_l)}vec\alpha}{\left|\left(\beta^T\left(X_{i_1}-X_{i_2}\right)^T\otimes I_{d_l}\right)I_{(d_lp_l)}vec\alpha\right|}\left|Y_{i_1}-Y_{i_3}\right|\right\},$$

$$U_{3n}(\alpha)=\binom{n}{4}^{-1}\sum_{1\le k_1<l_1<k_2<l_2m\le n}\times\left\{\frac{1}{24}\sum_{P_2}\frac{I_{(d_lp_l)}^T\left((X_{i_1}-X_{i_2})\beta\otimes I_{d_l}\right)\left(\beta^T\left(X_{i_1}-X_{i_2}\right)^T\otimes I_{d_l}\right)I_{(d_lp_l)}vec\alpha}{\left|\left(\beta^T\left(X_{i_1}-X_{i_2}\right)^T\otimes I_{d_l}\right)I_{(d_lp_l)}vec\alpha\right|}\left|Y_{i_3}-Y_{i_4}\right|\right\},$$

$$U_{1n}(\alpha)=\binom{n}{2}^{-1}\sum_{11\le k<l\le n}\times\frac{\left(I_{d_r}^T\otimes((X_k-X_l)^T\alpha)\right)\left(I_{d_r}\otimes\left(\alpha^T(X_k-X_l)\right)\right)vec\beta}{\left|\left(I_{d_r}\otimes\left(\alpha^T(X_k-X_l)\right)\right)vec\beta\right|}|Y_k-Y_l|,$$

$$U_{2n}(\beta)=\binom{n}{3}^{-1}\sum_{1\le k<l<m\le n}\times\left\{\frac{1}{6}\sum_{P_1}\frac{\left(I_{d_r}^T\otimes\left(\left(X_{i_1}-X_{i_2}\right)^T\alpha\right)\right)\left(I_{d_r}\otimes\left(\alpha^T\left(X_{i_1}-X_{i_2}\right)\right)\right)vec\beta}{\left|\left(I_{d_r}\otimes\left(\alpha^T\left(X_{i_1}-X_{i_2}\right)\right)\right)vec\beta\right|}\left|Y_{i_1}-Y_{i_3}\right|\right\},$$

$$U_{3n}(\beta)=\binom{n}{4}^{-1}\sum_{1\le k1<l1<k2<l2\le n}\times\left\{\frac{1}{24}\sum_{P_2}\frac{\left(I_{d_r}^T\otimes\left(\left(X_{i_1}-X_{i_2}\right)^T\alpha\right)\right)\left(I_{d_r}\otimes\left(\alpha^T\left(X_{i_1}-X_{i_2}\right)\right)\right)vec\beta}{\left|\left(I_{d_r}\otimes\left(\alpha^T\left(X_{i_1}-X_{i_2}\right)\right)\right)vec\beta\right|}\left|Y_{i_3}-Y_{i_4}\right|\right\},$$

and

$$\mu_1(\alpha)=E\frac{I_{(d_lp_l)}^T\left((X-X')\beta\otimes I_{d_l}\right)\left(\beta^T(X-X')^T\otimes I_{d_l}\right)I_{(d_lp_l)}vec\alpha}{\left|\left(\beta^T(X-X')^T\otimes I_{d_l}\right)I_{(d_lp_l)}vec\alpha\right|}|Y-Y'|,$$

$$\mu_2(\alpha)=E\frac{I_{(d_lp_l)}^T\left((X-X')\beta\otimes I_{d_l}\right)\left(\beta^T(X-X')^T\otimes I_{d_l}\right)I_{(d_lp_l)}vec\alpha}{\left|\left(\beta^T(X-X')^T\otimes I_{d_l}\right)I_{(d_lp_l)}vec\alpha\right|}|Y-Y''|,$$

$$\mu_3(\alpha)=E\frac{I_{(d_lp_l)}^T\left((X-X')\beta\otimes I_{d_l}\right)\left(\beta^T(X-X')^T\otimes I_{d_l}\right)I_{(d_lp_l)}vec\alpha}{\left|\left(\beta^T(X-X')^T\otimes I_{d_l}\right)I_{(d_lp_l)}vec\alpha\right|}|Y''-Y'''|,$$

$$\mu_1(\beta)=E\frac{\left(I_{d_r}^T\otimes((X-X')^T\alpha)\right)\left(I_{d_r}\otimes\left(\alpha^T(X-X')\right)\right)vec\beta}{\left|\left(I_{d_r}\otimes\left(\alpha^T(X-X')\right)\right)vec\beta\right|}|Y-Y'|,$$

$$\mu_2(\beta) = E\frac{\left(I_{d_r}^T \otimes ((X-X')^T\alpha)\right)\left(I_{d_r} \otimes \left(\alpha^T(X-X')\right)\right)vec\beta}{\left|\left(I_{d_r} \otimes \left(\alpha^T(X-X')\right)\right)vec\beta\right|}|Y-Y''|,$$

$$\mu_3(\beta) = E\frac{\left(I_{d_r}^T \otimes ((X-X')^T\alpha)\right)\left(I_{d_r} \otimes \left(\alpha^T(X-X')\right)\right)vec\beta}{\left|\left(I_{d_r} \otimes \left(\alpha^T(X-X')\right)\right)vec\beta\right|}|Y''-Y'''|,$$

Here $(X,\ Y), (X',\ Y'), (X'',\ Y''), (X''',\ Y''')$ are i.i.d copies.

Let $\mathbf{B} = \begin{pmatrix} \boldsymbol{I}_{p_l d_l}(-2)\boldsymbol{I}_{p_l d_l}\boldsymbol{I}_{p_l d_l} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0}^T & \mathbf{0}^T & \mathbf{0}^T & \boldsymbol{I}_{p_r d_r} - 2\boldsymbol{I}_{p_r d_r}\boldsymbol{I}_{p_r d_r} \end{pmatrix}$, where $\mathbf{0}$ is a $p_l d_l \times p_r d_r$ zero matrix, then

$$\sqrt{n}\mathbf{B}\begin{pmatrix} \mathbf{U}_{1n}(\alpha) - \mu_1(\alpha) \\ \mathbf{U}_{2n}(\alpha) - \mu_2(\alpha) \\ \mathbf{U}_{3n}(\alpha) - \mu_3(\alpha) \\ \mathbf{U}_{1n}(\beta) - \mu_1(\beta) \\ \mathbf{U}_{2n}(\beta) - \mu_2(\beta) \\ \mathbf{U}_{3n}(\beta) - \mu_3(\beta) \end{pmatrix} = \sqrt{n}\begin{pmatrix} \mathbf{U}_{1n}(\alpha) - 2\mathbf{U}_{2n}(\alpha) + \mathbf{U}3_n(\alpha) - \mathcal{V}'(\alpha) \\ \mathbf{U}_{1n}(\beta) - 2\mathbf{U}_{2n}(\beta) + \mathbf{U}3_n(\beta) - \mathcal{V}'(\beta) \end{pmatrix}.$$

Note that

$$\begin{aligned}\sqrt{n}\begin{pmatrix} \mathcal{V}_n'(\alpha) - \mathcal{V}'(\alpha) \\ \mathcal{V}_n'(\beta) - \mathcal{V}'(\beta) \end{pmatrix} &= \sqrt{n}\Bigg(\frac{(n-1)(n^2-2n+2)}{n^3}\mathbf{U}_{1n}(\alpha) - \frac{2(n-1)(n-2)^2}{n^3}\mathbf{U}_{2n}(\alpha) \\ &\quad + \frac{(n-1)(n-2)(n-3)}{n^3}\mathbf{U}_{3n}(\alpha) - \mathcal{V}'(\alpha)\frac{(n-1)(n^2-2n+2)}{n^3}\mathbf{U}_{1n}(\beta) \\ &\quad - \frac{2(n-1)(n-2)^2}{n^3}\mathbf{U}_{2n}(\beta) + \frac{(n-1)(n-2)(n-3)}{n^3}\mathbf{U}_{3n}(\beta) - \mathcal{V}'(\beta)\Bigg).\end{aligned}$$

Since under Assumption 1,

$$\begin{aligned}&\sqrt{n}\Bigg(\frac{(n-1)(n^2-2n+2)}{n^3}\mathbf{U}_{1n}(\alpha) - \frac{2(n-1)(n-2)^2}{n^3}\mathbf{U}_{2n}(\alpha) + \frac{(n-1)(n-2)(n-3)}{n^3}\mathbf{U}_{3n}(\alpha) \\ &\quad - \mathcal{V}'(\alpha)\frac{(n-1)(n^2-2n+2)}{n^3}\mathbf{U}_{1n}(\beta) - \frac{2(n-1)(n-2)^2}{n^3}\mathbf{U}_{2n}(\beta) \\ &\quad + \frac{(n-1)(n-2)(n-3)}{n^3}\mathbf{U}_{3n}(\beta) - \mathcal{V}'(\beta)\Bigg) \\ &\quad - \sqrt{n}\begin{pmatrix} \mathbf{U}_{1n}(\alpha) - 2\mathbf{U}_{2n}(\alpha) + \mathbf{U}3_n(\alpha) - \mathcal{V}'(\alpha) \\ \mathbf{U}_{1n}(\beta) - 2\mathbf{U}_{2n}(\beta) + \mathbf{U}3_n(\beta) - \mathcal{V}'(\beta) \end{pmatrix} \xrightarrow{P} 0.\end{aligned}$$

therefore by Slutsky's theorem, $\sqrt{n}\begin{pmatrix}\mathcal{V}_n'(\alpha)-\mathcal{V}'(\alpha)\\ \mathcal{V}_n'(\beta)-\mathcal{V}'(\beta)\end{pmatrix}\overset{D}{=}\sqrt{n}\mathbf{B}\begin{pmatrix}\mathbf{U}_{1n}(\alpha)-\mu_1(\alpha)\\ \mathbf{U}_{2n}(\alpha)-\mu_2(\alpha)\\ \mathbf{U}_{3n}(\alpha)-\mu_3(\alpha)\\ \mathbf{U}_{1n}(\beta)-\mu_1(\beta)\\ \mathbf{U}_{2n}(\beta)-\mu_2(\beta)\\ \mathbf{U}_{3n}(\beta)-\mu_3(\beta)\end{pmatrix}$. On the other hand, under Assumption 2 and the definition of second derivatives of $\mathcal{V}^2\big(Y,(\beta\otimes\alpha)^T vec(X)\big)$, by SLLN of U-Statistics, $\mathcal{L}_{11n}^{-1}\overset{a.s}{\to}\mathcal{L}_{11}^{-1}$, where $\mathcal{L}_{11}^{-1}$ denotes the upper-left corner submatrix of $(\mathcal{L}''(\theta))^{-1}$, which consists of the first $p_l d_l + p_r d_r$ rows and the first $p_l d_l + p_r d_r$ columns of $(\mathcal{L}''(\theta))^{-1}$.

Therefore, $\sqrt{n}\begin{pmatrix}vec(\alpha_n)-vec(\alpha)\\ vec(\beta_n)-vec(\beta)\end{pmatrix}\overset{D}{=}-(\mathcal{L}_{11n})^{-1}\sqrt{n}\mathbf{B}\begin{pmatrix}\mathbf{U}_{1n}(\alpha)-\mu_1(\alpha)\\ \mathbf{U}_{2n}(\alpha)-\mu_2(\alpha)\\ \mathbf{U}_{3n}(\alpha)-\mu_3(\alpha)\\ \mathbf{U}_{1n}(\beta)-\mu_1(\beta)\\ \mathbf{U}_{2n}(\beta)-\mu_2(\beta)\\ \mathbf{U}_{3n}(\beta)-\mu_3(\beta)\end{pmatrix}\to N(0,V_1)$, where $V_1=\mathcal{L}_{11}^{-1}\mathbf{B}\sum\mathbf{B}^T\mathcal{L}_{11}^{T-1}$.

**Proof of Corollary 1**

The proof of Corollary 1 is similar to the proof of Corollary 1 in Sheng and Yin [24], thus we omit some details here, especially some tedious calculations and only sketch the proof.

Since $\alpha_n$ and $\beta_n$ has the following asymptotic expansion forms:

$$\alpha_n=\alpha Q_1+E_n\{A^*\}+o_p\big(n^{-1/2}\big),\beta_n=\beta Q_2+E_n\{B^*\}+o_p\big(n^{-1/2}\big).$$

where $Q_1$ and $Q_2$ are orthogonal matrices. Then we can expand

$$\begin{aligned}&(\beta_n\otimes\alpha_n)(\beta_n\otimes\alpha_n)^T\\ &\quad=(\beta\otimes\alpha)(\beta\otimes\alpha)^T+(\beta\beta^T\otimes\alpha Q_1E_n^T\{A^*\})+(\beta\beta^T\otimes E_n\{A^*\}Q_1^T\alpha^T)\\ &\quad+(\beta Q_2E_n^T\{B^*\}\otimes\alpha\alpha^T)+(E_n\{B^*\}Q_2^T\beta^T\otimes\alpha\alpha^T)+o_p\big(n^{-1/2}\big).\end{aligned}$$

Since $vec\{(\beta\beta^T\otimes\alpha Q_1E_n^T\{A^*\})\}, vec\{(\beta\beta^T\otimes E_n\{A^*\}Q_1^T\alpha^T)\}$, $vec\{(\beta Q_2E_n^T\{B^*\}\otimes\alpha\alpha^T)\}$ and $vec\{(E_n\{B^*\}Q_2^T\beta^T\otimes\alpha\alpha^T)\}$ can be written as linear combinations of U-statistics, according to Theorem 6.1.6 ([13], Ch. 6), $\sqrt{n}\Big(vec\big((\beta_n\otimes\alpha_n)(\beta_n\otimes\alpha_n)^T\big)-vec\big((\beta\otimes\alpha)(\beta\otimes\alpha)^T\big)\Big)$ is asymptotically normal whose covariance matrix $V_2$ can be calculated.

*More simulation*. We also consider correlated $x_{ij}'$s in the matrix predictor. In the following, we use models 3 and 5 to illustrate the performance of different methods when the $(j_1,\ j_2)$th entry of $\sum=\mathrm{cov}(vec(x))$ equals to $0.5^{|j_1-j_2|}$. The simulation results are reported in Tables 7 and 8 accordingly.

Table A1 Accuracy of estimation for model 3 with correlated $x_{ij}'$s

| $(\boldsymbol{n},\ \boldsymbol{p_l},\ \boldsymbol{p_r})$ | **Folded-DCOV** | **Folded-SIR** | **Folded-MAVE** | **DF-PFC** | **DCOV** |
|---|---|---|---|---|---|
| (200, 5, 5) | 0.8101 (0.4178) | 1.4768 (0.2218) | 1.5817 (0.3284) | 1.3490 (0.2760) | 1.7009 (0.0880) |
| (200, 7, 7) | 1.3228 (0.3833) | 1.6120 (0.1521) | 1.7525 (0.2313) | 1.5216 (0.2381) | 1.8341 (0.0564) |
| (400, 5, 5) | 0.4820 (0.3447) | 1.3355 (0.1935) | 1.4640 (0.3846) | 1.2260 (0.2474) | 1.5860 (0.0624) |

| | | | | | |
|---|---|---|---|---|---|
| (400, 7, 7) | 0.8732 (0.3912) | 1.4654 (0.1699) | 1.5869 (0.2843) | 1.4118 (0.2213) | 1.7280 (0.0594) |

*Two-step matrix decomposition method*. Xue and Yin [29] proposed this method to decompose a semi-orthogonal matrix to a Kronecker product. To be self-contained, we introduce the details here.

Table A 2. Accuracy of estimation for model 5 with correlated xij's

| $(\boldsymbol{n}, \boldsymbol{p_l}, \boldsymbol{p_r})$ | **Folded-DCOV** | **DCOV** |
|---|---|---|
| (200, 5, 5) | 0.3502 (0.0913) | 1.4732 (0.0420) |
| (200, 7, 7) | 0.4343 (0.1604) | 1.6045 (0.1119) |
| (400, 5, 5) | 0.2880 (0.0644) | 1.4435 (0.0168) |
| (400, 7, 7) | 0.3135 (0.0680) | 1.4888 (0.0215) |

Suppose $\eta \in \mathbb{R}^{p_l p_r \times d_l d_r}$ is a semi-orthogonal matrix (that is, $\eta^T\eta = I_{d_l d_r}$) and $\eta = \beta^* \otimes \alpha^*$, where $\alpha^* \in \mathbb{R}^{p_l \times d_l}$, $\beta^* \in \mathbb{R}^{p_r \times d_r}$ and $\alpha^{*T}\alpha^* = I_{d_l}$, $\beta^{*T}\beta^* = I_{d_r}$ . The two-step decomposition method is to decompose $\eta$ to obtain matrices $\alpha^*$ and $\beta^*$. Suppose such $\eta$ is a basis of $\mathcal{S}_{Y|vec(X)}$ and $\mathcal{S}_{Y|vec(X)} = \mathcal{S}_{Y|^\circ X^\circ}$, then the two-step decomposition method can estimate $\mathcal{S}_{Y|^\circ X}$ and $\mathcal{S}_{Y|X^\circ}$ through $\mathcal{S}_{Y|vec(X)}$ . Here two-step means first, we estimate the basis of $\mathcal{S}_{Y|vec(X)}$, $\eta$; secondly, we decompose $\eta$ to $\alpha^*$ and $\beta^*$, which are initial values of bases of $\mathcal{S}_{Y|^\circ X}$ and $\mathcal{S}_{Y|X^\circ}$ in the algorithm described in Section 2.3 of this article. Let $\|\cdot\|$ be the Frobenius norm and let $h_{ij}$ be the $i$-dimensional vector with $j$th element being 1 and otherwise 0. The outline of the procedure is given below.

1. Generate the initial values of $\alpha^*_{(0)} \in \mathbb{R}^{p_l \times d_l}$ from $N(0,1)$.
2. Let $\alpha^*_{(k)}$ be the estimate of $\alpha^*$ in the $k$th iteration, the $ij$th element of $\beta^*_{(k)}$ in $k$th iteration is the minimizer, $\hat{b}$, of the objective function $\left\|\left(h^T_{p_r i} \otimes I_{p_l}\right)\eta\left(h_{p_r j} \otimes I_{p_l}\right) - b\alpha^*_{(k)}\right\|$. Normalize $\beta^*_{(k)}$ so that $\beta^{*T}_{(k)}\beta^*_k = I_{d_r}$.
3. Given $\beta^*_{(k)}$, the minimizer $\hat{a}$ of the matrix norm $\left\|\left(h^T_{p_l i} \otimes I_{p_r}\right)\left(\mathbb{K}_{p_l,p_r}\eta\mathbb{K}_{d_r,d_l}\right)\left(h_{d_l j} \otimes I_r\right) - a\beta^*_{(k)}\right\|$ is the $ij$th element of $\alpha^*_{(k+1)}$ in the $(k+1)$th iteration, where $\mathbb{K}$ is a commutation matrix. More details of commutation matrix can be found in Magnus and Neudecker [19]. Here we use: if $\alpha^* \in \mathbb{R}^{p_l \times d_l}$ and $\beta^* \in \mathbb{R}^{p_r \times d_r}$, then $\alpha^* \otimes \beta^* = \mathbb{K}_{p_l,p_r}\beta^* \otimes \alpha^*\mathbb{K}_{d_r,d_l}$ . Normalize $\alpha^*_{(k+1)}$ so that $\alpha^{*T}_{(k+1)}\alpha^*_{(k+1)} = I_{d_l}$.
4. Check the convergence. Let $\tau_{(k)} = \beta^*_{(k)} \otimes \alpha^*_{(k)}$ and $\tau_{(k+1)} = \beta^*_{(k+1)} \otimes \alpha^*_{(k+1)}$. If $\tau_{(k+1)}\tau^T_{(k+1)} - \tau_{(k)}\tau^T_{(k)}$ is smaller than the preset tolerance value, such as $10^{-6}$, then stop the iteration and set $\hat{\alpha}^* = \alpha^*_{(k+1)}$ and $\hat{\beta}^* = \beta^*_{(k+1)}$ ; otherwise, set $k :\equiv k + 1$ and go to step 2.

## REFERENCES


1. S. P. Albert and H. J. Shih, An approach for jointly modeling multivariate longitudinal measurements and discrete time-to-event data, *Ann. Appl. Stat*. 4 (2010), 1517 – 1532.
2. R. H. Byrd, J. C. Gilbert, and J. Nocedal, A trust region method based on interior point techniques for nonlinear programming, *Math. Program*. 89 (2000), 149 – 185.
3. R. H. Byrd, E. H. Mary, and J. Nocedal, An interior point algorithm for large-scale nonlinear programming, *SIAM J. Optim*. 9 (1999), 877 – 900.

4. R. D. Cook, Graphics for regressions with a binary response, *J. Am. Stat. Assoc*. 91 (1996), 983 – 992.
5. E. R. Dickson et al., Prognosis in primary biliary cirrhosis: Model for decision making, *Hepatology* 10 (1989), 1 – 7.
6. S. Ding and R. D. Cook, Dimension folding PCA and PFC for matrix-valued predictors, *Stat. Sin.* 24 (2014), 463 – 492.
7. S. Ding and R. D. Cook, Tensor sliced inverse regression, *J. Multivar. Anal*. 133 (2015a), 216 – 231.
8. S. Ding and R. D. Cook, Higer-order sliced inverse regression, *Wiley Interdiscip. Rev. Comput. Stat*. 7 (2015b), 249 – 257.
9. J. Fan and I. Gijbels, *Local polynomial and its applications*, Chapman & Hall, London, 1996.
10. R. T. Fleming and P. D. Harrington, *Counting process and survival analysis*, Wiley, New York, 1991.
11. H. Hung and C.-C. Wang, Matrix variate logistic regression model with application to EEG data, *Biostatistics* 14 (2013), 189 – 202.
12. R. W. Kim et al., Adaptation of the Mayo primary biliary cirrhosis natural history model for application in liver transplant candidates, *Liver Transplant*. 6 (2000), 489 – 494.
13. E. L. Lehmann, *Elements of large-sample theory*, Springer-Verlag, New York, 1999.
14. K.-C. Li, Sliced inverse regression for dimension reduction, *J. Am. Stat. Assoc*. 86 (1991), 316 – 327.
15. B. Li, M. Kim, and N. Altman, On dimension folding of matrix- or array-valued statistical objects, *Ann. Stat*. 38 (2010), 1094 – 1121.
16. L. Li and X. Yin, Longitudinal data analysis using sufficient dimension reduction method, *Comput. Stat. Data* Anal. 53 (2009), 4106 – 4115.
17. B. Li, H. Zha, and F. Chiaromonte, Contour regression: A general approach to dimension reduction, *Ann. Stat*. 33 (2005), 1580 – 1616.
18. R. Luo, H. Wang, and C.-L. Tsai, Contour projected dimension reduction, *Ann. Stat*. 37 (2009), 3743 – 3778.
19. R. J. Magnus and H. Neudecker, *Matrix differential calculus with applications in statistics and econometrics*, 2nd ed., Wiley, New York, 1999.
20. A.P. Murtaugh et al., Primary biliary cirrhosis: Prediction of short-term survival based on repeated patient visits, *Hepatology* 20 (1994), 126 – 136.
21. M. R. Pfeiffer, L. Forzani, and E. Bura, Sufficient dimension reduction for longitudinally measured predictors, *Stat. Med*. 22 (2011), 2414 – 2427.
22. R. J. Serfling, *Approximation theorems of mathematical statistics*, Wiley, New York, 1980.
23. W. Sheng and X. Yin, Direction estimation in single-index models via distance covariance, *J. Multivar. Anal*. 122 (2013), 148 – 161.
24. W. Sheng and X. Yin, Sufficient dimension reduction via distance covariance, *J. Comput. Graph. Stat.* 25 (2016), 91 – 104.
25. G. J. Székely and M. L. Rizzo, Brownian distance covariance, *Ann. Appl. Stat*. 3 (2009), 1236 – 1265.
26. G. J. Székely, M. L. Rizzo, and N. K. Bakirov, Measuring and testing dependence by correlation of distances, *Ann. Stat*. 35 (2007), 2769 – 2794.

27. R. A. Waltz et al., An interior algorithm for nonlinear optimization that combines line search and trust region steps, *Math. Program*. 107 (2006), 391 – 408.
28. Y. Xia et al., An adaptive estimation of dimension reduction space, *J. R. Stat. Soc. B* 64 (2002), 363 – 410.
29. Y. Xue and X. Yin, Sufficient dimension folding for regression mean function, *J. Comput. Graph. Stat.* 23 (2014), 1028 – 1043.
30. Y. Xue and X. Yin, Sufficient dimension folding for a functional of conditional distribution of matrix- or array-valued objects, *J. Nonparametric Stat*. 27 (2015), 253 – 269.
31. Y. Xue, X. Yin, and X. Jiang, Ensemble sufficient dimension folding methods for analyzing matrix-valued data, *Comput. Stat. Data Anal*. 103 (2016), 193 – 205.
32. H. Zhou, L. Li, and H. Zhu, Tensor regression with applications in neuroimaging data analysis, *J. Am. Stat. Assoc*. 108 (2013), 540 – 552.