Journal of Multivariate Analysis 133 (2015) 216–231

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

# Tensor sliced inverse regression

Shanshan Ding[a,*], R. Dennis Cook[b]

[a] Department of Applied Economics and Statistics, University of Delaware, 225 Townsend Hall 531 S College Ave, Newark, DE 19711, USA

[b] School of Statistics, University of Minnesota, 313 Ford Hall 224 Church St SE, Minneapolis, MN 55455, USA

ARTICLE INFO

*Article history:*
Received 6 November 2013
Available online 28 September 2014

*AMS subject classifications:*
62B05
62F12
62H12
62H30
62H35

*Keywords:*
Sufficient dimension reduction
Sliced inverse regression
Central subspace
Central dimension folding subspace
Tensor data
Tensor decomposition

ABSTRACT

Sliced inverse regression (SIR) is a widely used non-parametric method for supervised dimension reduction. Conventional SIR mainly tackles simple data structure but is inappropriate for data with array (tensor)-valued predictors. Such data are commonly encountered in modern biomedical imaging and social network areas. For these complex data, dimension reduction is generally demanding to extract useful information from abundant measurements. In this article, we propose higher-order sufficient dimension reduction mainly by extending SIR to general tensor-valued predictors and refer to it as tensor SIR. Tensor SIR is constructed based on tensor decompositions to reduce a tensor-valued predictor's multiple dimensions simultaneously. The proposed method provides fast and efficient estimation. It circumvents high-dimensional covariance matrix inversion that researchers often suffer when dealing with such data. We further investigate its asymptotic properties and show its advantages by simulation studies and a real data application.

## 1. Introduction

Sliced inverse regression was proposed by Li [16]. It is a major supervised dimension reduction technique in non-parametric regression problems. It assumes that the response variable $Y \in \mathbb{R}^1$ depends on the predictor $X \in \mathbb{R}^p$ only through $K(K < p)$ unknown linear combinations of the predictor. Let $B = (\beta_1, \beta_2, \ldots, \beta_K) \in \mathbb{R}^{p\times K}$. This relationship can be described as $Y \perp\!\!\!\perp X|B^TX$, where '$\perp\!\!\!\perp$' stands for independence. To build SIR into the sufficient dimension reduction (SDR) framework, $B^TX$ is called a sufficient reduction of $X$ [1,2]. The matrix $B$ itself is not identifiable since it can be replaced by any non-singular transformation of its columns. However, the linear space spanned by the columns of $B$ is identifiable, denoted as $\mathcal{S}_B$, or Span($B$). As a consequence of this structure one can reduce the dimension of the predictor $X$ by replacing it with its projection $\mathrm{P}_{\mathcal{S}_B}X$ onto the subspace $\mathcal{S}_B$, without loss of information on the conditional distribution of $Y|X$; that is,

$$Y \perp\!\!\!\perp X|\mathrm{P}_{\mathcal{S}_B}X. \tag{1}$$

When $K$ is the smallest column rank of $B$ such that (1) holds, the subspace $\mathcal{S}_B$ is called the central dimension reduction subspace (CS), denoted as $\mathcal{S}_{Y|X}$. The goal of SIR is to estimate $\mathcal{S}_{Y|X}$. We provide a brief review of the SIR procedure in Section 2.1.

Conventional SIR is simple and useful for dimension reduction of a vector-valued predictor $X \in \mathbb{R}^p$. However, it is inefficient to tackle problems with more general tensor-valued predictors, such as an $m$-mode tensor $\mathcal{X} \in \mathbb{R}^{p_1\times p_2\times\cdots\times p_m}$. This type of data is commonly encountered in applications. For instance, EEG (electroencephalography) signals in biomedical

[*] Corresponding author.
*E-mail addresses:* sdingg@udel.edu (S. Ding), dennis@stat.umn.edu (R.D. Cook).

http://dx.doi.org/10.1016/j.jmva.2014.08.015
0047-259X/© 2014 Elsevier Inc. All rights reserved.

engineering, gene expression in bioinformatics and images in pattern recognition are usually formed as two-mode tensors. Video sequences, spatial data and data in social networks often contain three- or multi-mode tensor predictors. Such data are often referred to as multivariate relational data because the tensor-valued predictors represent intrinsic spatial, repeated measured, or other correlated structure among variables. In the EEG data, for example, the brain signals of each subject forms a 256 × 64 matrix (two-mode tensor) with its rows and columns representing time and location information respectively. Due to the curse of dimensionality, SDR is desirable for such complex data. However, vectorizing these higher-order predictors could and typically does lose important information about the data structure and decrease estimation accuracy.

SDR for tensor-valued predictors has received increasing attention in recent literature. Pioneering work was done by Li et al. [17], where the authors proposed the idea of dimension folding and developed a class of moment-based dimension folding methods, including dimension folding SIR, to reduce a tensor predictor's multiple dimensions simultaneously. Their methods apply to many moment-based dimension reduction approaches but, as will be shown in later sections, are not very efficient, in operation, for dealing with higher-order tensor predictors. Other works include longitudinal SIR studied by Pfeiffer et al. [21] and dimension folding PCA and PFC developed by Ding and Cook [7]. These two studies focused only on two-mode tensor predictors, $\mathbf{X} \in \mathbb{R}^{p_1 \times p_2}$.

In this paper, we propose a higher-order SDR approach by extending SIR to general $m$-mode tensor-valued predictors; we refer to it as tensor SIR. The proposed method makes more efficient use of the tensor structure and leads to $\sqrt{n}$ consistent and asymptotically normal estimator of the sufficient reduction subspace. We further compare tensor SIR with the aforementioned methods in the two-mode tensor case. Tensor SIR outperforms dimension folding SIR by (i) circumventing high-dimensional covariance matrix inversion; (ii) alleviating computational cost and improving estimation accuracy; and (iii) having easy interpretation and good theoretical properties. In comparison to longitudinal SIR, tensor SIR places fewer restrictions on the covariance structure of $\mathrm{vec}(\mathcal{X})$. It provides the maximum likelihood estimation of the sufficient reduction when $\mathbf{X}|Y$ is matrix-normally distributed and $\mathrm{cov}[\mathrm{vec}(\mathbf{X})]$ has a Kronecker structure.

The rest of this paper is organized as follows. Section 2 introduces tensor SIR for two-mode tensor predictors, called two-tensor SIR. Section 3 is devoted to the development of tensor SIR for more general $m$-mode tensor predictors. We develop the asymptotic properties for the proposed methods in Section 4. Section 5 establishes connections between tensor SIR and other high-order SDR methods. Sections 6 and 7 contain simulation results and data analyses. Discussion is given in Section 8.

## 2. Two-tensor SIR

Without loss of generality, we assume that the predictors discussed in this paper have mean zero. Let $P_B = B(B^TB)^\dagger B^T$ be the projection onto Span$(B)$, and $P^T_{B(A)} = AB(B^TAB)^\dagger B^T$ be the projection onto Span$(B)$ relative to $A$, where $B \in \mathbb{R}^{p\times d}$ and $A \in \mathbb{R}^{p\times p}$ $(A > 0)$ are two matrices, and † is the Moore–Penrose inverse. Before introducing tensor SIR, we provide a brief review for the conventional SIR.

### 2.1. A review of SIR

In the classical setting, $X \in \mathbb{R}^p$ is a predictor vector and $Y \in \mathbb{R}^1$ is a response variable. SIR serves to reduce the predictor's dimension by finding the CS $\mathcal{S}_{Y|X}$ so that the projected predictor $P_{\mathcal{S}_{Y|X}}X$ retains the full information on $Y|X$. Let $\mathcal{S}_{Y|X} = \mathrm{Span}(\eta)$, where $\eta \in \mathbb{R}^{p\times d}$ $(d \le p)$. Let $\Sigma$ and $\hat{\Sigma}$ be the covariance and sample covariance matrices of $X$. Under the linearity condition (Condition 3.1 in [16]), $\mathrm{E}(X|\eta^TX)$ is a linear function of $\eta^TX$. That is, $\mathrm{E}(X|\eta^TX) = A\eta^TX$, where $A$ has an explicit expression $A = \Sigma\eta(\eta^T\Sigma\eta)^\dagger$ (Proposition 4.2, [2]). Therefore,

$$\mathrm{E}(X|Y) = \mathrm{E}[\mathrm{E}(X|\eta^TX, Y)|Y] = \mathrm{E}[\mathrm{E}(X|\eta^TX)|Y] = P^T_{\eta(\Sigma)}\mathrm{E}(X|Y), \tag{2}$$

which indicates $\mathrm{E}(X|Y) \in \mathrm{Span}(\Sigma\eta)$. Correspondingly, $\Sigma^{-1}\mathrm{Span}\{\mathrm{cov}[\mathrm{E}(X|Y)]\} \subseteq \mathcal{S}_{Y|X}$. Conventional SIR estimates $\mathcal{S}_{Y|X}$ by the sample estimate $\hat{\Sigma}^{-\frac{1}{2}}$ times the leading $d$ eigenvectors of $\widehat{\mathrm{cov}}[\hat{\Sigma}^{-\frac{1}{2}}\hat{\mathrm{E}}(X|Y)]$. To allow relatively easy estimation of the inverse mean $\mathrm{E}(X|Y)$, the response $Y$ is replaced with a discrete version by partitioning the range of $Y$ into certain slices. One estimates $\mathrm{E}(X|Y)$ by the intraslice mean.

### 2.2. Two-tensor SIR

To introduce the idea of tensor SIR, we first consider a simple case when the predictor $\mathbf{X} \in \mathbb{R}^{p_1\times p_2}$ is two-mode tensor-valued (matrix-valued) and the response $Y$ is univariate. We propose an SDR procedure called two-tensor SIR. It is a special case of tensor SIR dealing with matrix-valued predictors.

The sufficient dimension reduction for $\mathbf{X} \in \mathbb{R}^{p_1\times p_2}$ is defined as follows. Let '⊗' stand for the Kronecker product.

**Definition 1** (*Li et al. [17]*). Let $B_1 \in \mathbb{R}^{p_1\times d_1}$ $(d_1 \le p_1)$ and $B_2 \in \mathbb{R}^{p_2\times d_2}$ $(d_2 \le p_2)$ be two semi-orthogonal matrices that satisfy

$$Y \perp\!\!\!\perp \mathbf{X}|B_1^T\mathbf{X}B_2. \tag{3}$$

(i) Then $\text{Span}(B_2) \otimes \text{Span}(B_1)$ is called a dimension folding subspace. (ii) If $d_1$ and $d_2$ both are the smallest column dimensions such that (3) holds, then $\text{Span}(B_2) \otimes \text{Span}(B_1)$ is called the central tensor (dimension folding) subspace (CTS) for $Y|\mathbf{X}$, denoted as $\mathcal{S}_{Y|\circ\mathbf{X}\circ}$.

The key idea is to reduce the predictor's row and column dimensions simultaneously while preserving its matrix structure without loss of information on $Y|\mathbf{X}$. The column spans of the semi-orthogonal matrices $B_1$ an $B_2$ indicate the row and column SDR directions respectively. Li et al. [17] proposed dimension folding SIR to estimate the CTS for $\mathbf{X} \in \mathbb{R}^{p_1\times p_2}$. Their method relies on the linearity condition on $\text{vec}(\mathbf{X})$ and thus the estimation is built on $\text{vec}(\mathbf{X})$. Two-tensor SIR considers a matrix-formed linearity condition and leads to more efficient estimation for the CTS. We propose the methodology below and provide a connection between the two methods in Section 5.

Let $\alpha \in \mathbb{R}^{p_1\times d_1}(d_1 \le p_1)$ and $\beta \in \mathbb{R}^{p_2\times d_2}(d_2 \le p_2)$ be two full rank matrices. Assume that the conditional means $\text{E}(\mathbf{X}|\alpha^T\mathbf{X})$ and $\text{E}(\mathbf{X}|\mathbf{X}\beta)$ are linear functions of $\alpha^T\mathbf{X}$ and $\mathbf{X}\beta$ respectively. In other words, there exist two uniquely defined matrices $A \in \mathbb{R}^{p_1\times d_1}$ and $B \in \mathbb{R}^{p_2\times d_2}$ such that

$$\text{E}(\mathbf{X}|\alpha^T\mathbf{X}) = A\alpha^T\mathbf{X}, \qquad \text{E}(\mathbf{X}|\mathbf{X}\beta) = \mathbf{X}\beta B^T. \tag{4}$$

Condition (4) means that the linearity holds along each mode (row and column) of the two-mode tensor predictor. It is different than the linearity condition directly imposed on $\text{vec}(\mathbf{X})$, which is commonly used by other higher-order SDR methods in literature. We will provide a comparison between the different linearity conditions in Section 5.1.

Under (4), the matrices $A$ and $B$ are uniquely determined.

**Lemma 1.** *Let $\Omega_1 = \text{E}(\mathbf{X}\mathbf{X}^T)$ and $\Omega_2 = \text{E}(\mathbf{X}^T\mathbf{X})$ be the column and row covariance matrices of $\mathbf{X}$. If condition (4) holds for full rank matrices $\alpha$ and $\beta$, then $A = \Omega_1\alpha(\alpha^T\Omega_1\alpha)^{-1}$ and $B = \Omega_2\beta(\beta^T\Omega_2\beta)^{-1}$.*

Now suppose that $\mathcal{S}_{Y|\circ\mathbf{X}\circ} = \text{Span}(B_2 \otimes B_1)$. Two-tensor SIR only requires a special case $\alpha = B_1$ and $\beta = B_2$ for the linearity condition (4). Then according to Lemma 1, it follows that

$$\text{E}(\mathbf{X}|Y) = \text{E}[\text{E}(\mathbf{X}|B_1^T\mathbf{X}, Y)|Y] = \text{E}[\text{E}(\mathbf{X}|B_1^T\mathbf{X})|Y] = P^T_{B_1(\Omega_1)}\text{E}(\mathbf{X}|Y), \tag{5}$$

Similarly, we observe $\text{E}(\mathbf{X}|Y) = \text{E}(\mathbf{X}|Y)P_{B_2(\Omega_2)}$. Therefore,

$$\text{E}(\mathbf{X}|Y) = P^T_{B_1(\Omega_1)}\text{E}(\mathbf{X}|Y)P_{B_2(\Omega_2)}. \tag{6}$$

Let $\Gamma_1$ and $\Gamma_2$ be the bases of $\text{Span}(\Omega_1 B_1)$ and $\text{Span}(\Omega_2 B_2)$ respectively. Then the CTS is equivalent to $\mathcal{S}_{Y|\circ\mathbf{X}\circ} = (\Omega_1^{-1} \otimes \Omega_1^{-1})\text{Span}(\Gamma_1 \otimes \Gamma_2)$. Correspondingly, (6) can be reformulated as

$$\text{E}(\mathbf{X}|Y) = P_{\Gamma_1}\text{E}(\mathbf{X}|Y)P_{\Gamma_2}, \tag{7}$$

or equivalently,

$$\text{E}[\text{vec}(\mathbf{X})|Y] = P_{\Gamma_2\otimes\Gamma_1}\text{E}[\text{vec}(\mathbf{X})|Y]. \tag{8}$$

Eqs. (7) and (8) indicate that in addition to the fact $\text{Span}\{\text{E}[\text{vec}(\mathbf{X})|Y]\} \subseteq \text{Span}(\Gamma_2 \otimes \Gamma_1)$, the relations $\text{Span}\{\text{E}(\mathbf{X}|Y)P_{\Gamma_2}\} \subseteq \text{Span}(\Gamma_1)$ and $\text{Span}\{\text{E}(\mathbf{X}^T|Y)P_{\Gamma_1}\} \subseteq \text{Span}(\Gamma_2)$ hold. They suggest that after projecting the row (column) space of $\text{E}(\mathbf{X}|Y)$ onto $\text{Span}(\Gamma_2)$ ($\text{Span}(\Gamma_1)$), the column (row) space of the projected matrix is a subspace of $\text{Span}(\Gamma_1)$ ($\text{Span}(\Gamma_2)$). Let $\text{cov}_c[A] = \text{E}[AA^T]$ be the column covariance matrix for any random matrix $A$. Then the column spaces of $\text{cov}_c[\text{E}(\mathbf{X}|Y)P_{\Gamma_2}]$ and $\text{cov}_c[\text{E}(\mathbf{X}^T|Y)P_{\Gamma_1}]$ are contained in $\text{Span}(\Gamma_1)$ and $\text{Span}(\Gamma_2)$ respectively. These relationships provide the basic idea for tensor SIR to estimate the CTS and, as stated in the following proposition, they can be derived by minimizing the discrepancy function

$$\text{E}\|\text{E}(\mathbf{X}|Y) - P_{\Gamma_1}\text{E}(\mathbf{X}|Y)P_{\Gamma_2}\|^2_{\text{F}}, \tag{9}$$

where $\|\cdot\|_{\text{F}}$ stands for the Frobenius norm.

**Proposition 1.** *Let $(\Gamma_1, \Gamma_2)$ be the minimizers of the objective function*

$$\text{E}\|\text{E}(\mathbf{X}|Y) - P_{G_1}\text{E}(\mathbf{X}|Y)P_{G_2}\|^2_{\text{F}}, \tag{10}$$

*over all semi-orthogonal matrices $G_1 \in \mathbb{R}^{p_1\times d_1}$ and $G_2 \in \mathbb{R}^{p_2\times d_2}$. Then*

(i) *For fixed $G_1$, the columns of the minimizer $\Gamma_2$ over $G_2$ consist of the $d_2$ eigenvectors of the matrix $\Sigma_R = \text{E}[\text{E}(\mathbf{X}^T|Y)P_{G_1}\text{E}(\mathbf{X}|Y)]$ corresponding to its $d_2$ largest nonzero eigenvalues.*

(ii) *For fixed $G_2$, the columns of the minimizer $\Gamma_1$ over $G_1$ are given by the $d_1$ eigenvectors of the matrix $\Sigma_L = \text{E}[\text{E}(\mathbf{X}|Y)P_{G_2}\text{E}(\mathbf{X}^T|Y)]$, corresponding to its $d_1$ largest nonzero eigenvalues.*

According to Proposition 1, for an iid sample $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, by slicing the responses into $H$ categories, one can apply the following algorithm to estimate $\Gamma_1$, $\Gamma_2$ and the CTS.

1. Generate initial values of $\Gamma_{10}$ and let $\hat{\Gamma}_1 = \Gamma_{10}$.
2. Given fixed $\hat{\Gamma}_1$, for each slice $J_s$, $s = 1, \ldots, H$, compute the sample mean within the category by $\bar{\mathbf{X}}_s = \frac{\sum_{Y_i\in J_s}\mathbf{X}_i}{n_s}$, where $n_s$ is number of observations within category $s$. Compute the weighted column covariance matrix $\hat{\Sigma}_R = \sum_{s=1}^H \frac{n_s}{n}\bar{\mathbf{X}}_s^T\hat{\Gamma}_1\hat{\Gamma}_1^T\bar{\mathbf{X}}_s$ and take the $d_2$ eigenvectors of $\hat{\Sigma}_R$ corresponding to its $d_2$ largest eigenvalues to form the columns of $\hat{\Gamma}_2$.

3. For fixed $\hat{\Gamma}_2$, compute the weighted column covariance matrix $\hat{\Sigma}_L = \sum_{s=1}^{H} \frac{n_s}{n} \bar{\mathbf{X}}_s \hat{\Gamma}_2 \hat{\Gamma}_2^T \bar{\mathbf{X}}_s^T$ and take the $d_1$ eigenvectors of $\hat{\Sigma}_L$ corresponding to its $d_1$ largest eigenvalues to form the columns of $\hat{\Gamma}_1$.
4. Repeat 2 and 3 and iterate with the updated estimators until the objective function $\sum_{s=1}^{H} \frac{n_s}{n} \|\bar{\mathbf{X}}_s - P_{\hat{\Gamma}_1} \bar{\mathbf{X}}_s P_{\hat{\Gamma}_2}\|_F^2$ converges. Then the CTS $\mathcal{S}_{Y|\mathbf{X}}$ is estimated by $(\hat{\Omega}_2^{-1} \otimes \hat{\Omega}_1^{-1})\text{Span}(\hat{\Gamma}_2 \otimes \hat{\Gamma}_1)$, where $\hat{\Omega}_1 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T$ and $\hat{\Omega}_2 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{X}_i$.

Two-tensor SIR is well connected with conventional SIR and is easily interpreted. The algorithm shows that in order to reduce the dimension of each mode, one first needs to project the column space of the other mode into its sufficient reduction subspace and then apply SIR for the reduced predictors. Hence two-tensor SIR can be treated as an adaptive SIR procedure. We will show more advantages of this method in Sections 4 and 5.

## 3. Multiple mode tensor SIR

### 3.1. Methodology

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

**Definition 2** (*k-th Mode Product*). The product of an $m$-mode tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_m}$ and a matrix $A_k \in \mathbb{R}^{d_k \times p_k} (k \in \mathcal{M})$, is a $p_1 \times \cdots \times p_{k-1} \times d_k \times p_{k+1} \times \cdots \times p_m$ dimensional $m$-mode tensor, denoted by $\mathcal{X} \times_k A_k$, with its $i_1 \cdots i_{k-1} j_k i_{k+1} \cdots i_m$-th element defined as

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

**Definition 3** (*Tensor Matricization*). The $k$-th mode unfolding matrix of an $m$-mode tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_m}$ is defined as $\mathbf{X}_{(k)} \in \mathbb{R}^{p_k \times (p_{k+1} \cdots p_m p_1 \cdots p_{k-1})}$, $k \in \mathcal{M}$, where the $i$-th row of $\mathbf{X}_{(k)}$ contains all elements of $\mathcal{X}$ that have the $k$-th index equal to $i$.

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

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

then the unfolding along the third mode gives

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

Based on Definitions 2 and 3, the following properties hold. For $A_k \in \mathbb{R}^{p_k \times d_k}$ $(k \in \mathcal{M})$,

(i) $\mathcal{Y} = \mathcal{X} \times_k A_k^T \Leftrightarrow \mathbf{Y}_{(k)} = A_k^T \mathbf{X}_{(k)}$;
(ii) $\mathcal{Y} = \mathcal{X} \times_1 A_1^T \times_2 A_2^T \times_3 \cdots \times_m A_m^T \Leftrightarrow \mathbf{Y}_{(k)} = A_k^T \mathbf{X}_{(k)} (A_m \otimes \cdots \otimes A_{k+1} \otimes A_{k-1} \otimes \cdots \otimes A_1)$;
(iii) $\text{vec}(\mathcal{Y}) = \text{vec}(\mathbf{Y}_{(1)}) = (A_m \otimes \cdots A_1)^T \text{vec}(\mathcal{X}) = (\bigotimes_{j=m}^{1} A_j)^T \text{vec}(\mathcal{X})$.

For further background on tensor operations, see [12–15]. The vectorization of a tensor is usually defined by vectorizing its first mode unfolding matrix; that is, $\text{vec}(\mathcal{X}) = \text{vec}(\mathbf{X}_{(1)})$. Hence in (iii), the index order of $A_j$ $(j \in \mathcal{M})$ is from $m$ to 1. In general, the choice of the unfolding order is not important as one can always convert $\text{vec}(\mathbf{X}_{(k)})$ to $\text{vec}(\mathbf{X}_{(1)})$ by elementary row exchange.

The goal of SDR for an $m$-mode tensor predictor $\mathcal{X}$ is to reduce the predictor's multiple dimensions simultaneously so that the reduced $m$-mode tensor contains full information about the response $Y$ while preserving the tensor structure. The formal definition is given below.

**Definition 4.** Let $B_1 \in \mathbb{R}^{p_1 \times d_1}, B_2 \in \mathbb{R}^{p_2 \times d_2}, \ldots, B_m \in \mathbb{R}^{p_m \times d_m}$ be $m$ semi-orthogonal matrices. If $d_1, d_2, \ldots, d_m$ are the minimum column dimensions such that $Y \perp\!\!\!\perp \mathcal{X} \mid \mathcal{X} \times_1 B_1^T \times_2 B_2^T \cdots \times_m B_m^T$, then $\text{Span}(\bigotimes_{j=m}^{1} B_j)$ is the CTS for $\mathcal{X}$, denoted as $\mathcal{S}_{Y|\mathcal{X} \circ m}$.

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

$$\text{E}(\mathbf{X}_{(k)} | \alpha_k^T \mathbf{X}_{(k)}) = A_k \alpha_k^T \mathbf{X}_{(k)}, \quad k \in \mathcal{M}, \tag{11}$$

or equivalently,

$$\text{E}(\mathcal{X} | \mathcal{X} \times_k \alpha_k^T) = \mathcal{X} \times_k A_k \alpha_k^T, \quad k \in \mathcal{M}, \tag{12}$$

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

**Lemma 2.** *Let $\Omega_k = \mathrm{E}(\mathbf{X}_{(k)}\mathbf{X}_{(k)}^T)$ be the k-th mode covariance matrix of $\mathcal{X}$. If condition (11) holds for full rank matrices $\alpha_k$, then $A_k = \Omega_k\alpha_k(\alpha_k^T\Omega_k\alpha_k)^{-1}$, $k \in \mathcal{M}$.*

Let $\mathrm{Span}(\bigotimes_{j=m}^{1} B_j)$ be the CTS of $Y|\mathcal{X}$. Similar to the two-mode tensor case, Tensor SIR only requires that the linearity condition (11) holds for $\alpha_k = B_k$, $k \in \mathcal{M}$, but not all full rank $d_k$ matrices. Then according to Lemma 2, using the same argument in (5), we have $\mathrm{E}(\mathbf{X}_{(k)}|Y) = P_{B_k(\Omega_k)}^T\mathrm{E}(\mathbf{X}_{(k)}|Y)$, or equivalently, $\mathrm{E}(\mathcal{X}|Y) = \mathrm{E}(\mathcal{X}|Y) \times_k P_{B_k(\Omega_k)}^T$, $k \in \mathcal{M}$. Therefore, by continuing operation on $\mathrm{E}(\mathcal{X}|Y)$ over all $k \in \mathcal{M}$, we have

$$\mathrm{E}[\mathcal{X}|Y] = \mathrm{E}[\mathcal{X}|Y] \times_1 P_{B_1(\Omega_1)}^T \times_2 P_{B_2(\Omega_2)}^T \times_3 \cdots \times_m P_{B_m(\Omega_m)}^T. \tag{13}$$

Now let $\Gamma_k$ be the bases of $\mathrm{Span}(\Omega_k B_k)$, $k \in \mathcal{M}$. Then $\mathcal{S}_{Y|\mathcal{X}\circ m} = (\bigotimes_{j=m}^{1} \Omega_j^{-1}) \cdot \mathrm{Span}(\bigotimes_{j=m}^{1} \Gamma_j)$ and (13) can be reformulated as

$$\mathrm{E}(\mathcal{X}|Y) = \mathrm{E}(\mathcal{X}|Y) \times_1 P_{\Gamma_1} \times_2 P_{\Gamma_2} \times_3 \cdots \times_m P_{\Gamma_m}. \tag{14}$$

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

**Proposition 2.** *Let $(\Gamma_1, \Gamma_2, \ldots, \Gamma_m)$ be the minimizers of the objective function*

$$\mathrm{E}\|\mathrm{E}(\mathcal{X}|Y) - \mathrm{E}(\mathcal{X}|Y) \times_1 P_{G_1} \times_2 P_{G_2} \times_3 \cdots \times_m P_{G_m}\|_{\mathrm{F}}^2, \tag{15}$$

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

Correspondingly, for an iid sample $(\mathcal{X}_i, Y_i)$, $i = 1, \ldots, n$, supposed that the responses are sliced into $H$ categories. Let $\bar{\mathcal{X}}_s = \sum_{y_i\in J_s} \mathcal{X}_i/n_s$ be the sample mean within slice $J_s$ and let $\bar{\mathbf{X}}_{s_{(k)}}$ be the $k$-th unfolding matrix of $\bar{\mathcal{X}}_s$, where $n_s$ is the number of observations in $J_s$, $s = 1, \ldots, H$. Based on Proposition 2, tensor SIR estimates of the CTS can be obtained in the following way:

1. Generate initial values of $\hat{\Gamma}_k^{(0)} \in \mathbb{R}^{p_k\times d_k}$, $k = 2, \ldots, m$, such that the column space of $\hat{\Gamma}_k^{(0)}$ is chosen to be the dominant eigenspace of the sample estimate of $\mathrm{cov}_c[\mathrm{E}(\mathbf{X}_{(k)}|Y)]$. For notation convenience, let $\hat{\Gamma}_k = \hat{\Gamma}_k^{(0)}$.
2. Update $\hat{\Gamma}_1, \ldots, \hat{\Gamma}_m$ sequentially by forming the columns of $\hat{\Gamma}_k$ as the leading $d_k$ eigenvectors of
$$\hat{\Sigma}_k = n^{-1}\sum_{s=1}^{H} n_s\bar{\mathbf{X}}_{s_{(k)}}\left(\bigotimes_{j=m,j\neq k}^{1} P_{\hat{\Gamma}_j}\right)\bar{\mathbf{X}}_{s_{(k)}}^T, \quad k \in \mathcal{M},$$
given $\hat{\Gamma}_j, j \in \mathcal{M}, j \neq k$.
3. Iterate step 2 until the objective function $n^{-1}\sum_{s=1}^{H} n_s\|\bar{\mathcal{X}}_s - \bar{\mathcal{X}}_s \times_1 P_{\hat{\Gamma}_1} \times_2 \cdots \times_m P_{\hat{\Gamma}_m}\|_{\mathrm{F}}^2$ converges. The CTS is then estimated by $\mathrm{Span}(\bigotimes_{j=m}^{1} \hat{\Omega}_j^{-1}\hat{\Gamma}_j)$, where $\hat{\Omega}_j = n^{-1}\sum_{i=1}^{n} \mathbf{X}_{(j)_i}\mathbf{X}_{(j)_i}^T$ is the sample column covariance matrix of $\mathbf{X}_{(j)}$, $j \in \mathcal{M}$.

Similar to the discussion in Section 2, this algorithm can be treated as an adaptive SIR algorithm for multiple mode tensor predictors.

### *3.2. Kronecker tensor SIR*

In the conventional setting $X \in \mathbb{R}^p$, Cook [3] showed that SIR provides the MLE for the central subspace when $X|Y$ is multivariate normal. It would be interesting to see whether tensor SIR yields the MLE for the CTS. We propose an alternative tensor SIR procedure that requires a special structure for $\mathrm{cov}[\mathrm{vec}(\mathcal{X})]$ and leads to the MLE. The procedure is described below. A statistical justification is given in Section 5.4.

Assume that $\mathrm{cov}[\mathrm{vec}(\mathcal{X})]$ has the Kronecker structure

$$\mathrm{cov}[\mathrm{vec}(\mathcal{X})] = V_m \otimes V_{m-1} \otimes \cdots \otimes V_1 = \bigotimes_{j=1}^{m} V_j, \tag{16}$$

where $V_j \in \mathbb{R}^{p_j\times p_j}$, $j \in \mathcal{M}$. It can be shown that each separate covariance matrix $V_j$ corresponds to the $j$-th unfolding matrix with $V_j = \mathrm{E}[\mathbf{X}_{(j)}\mathbf{X}_{(j)}^T]/\prod_{i=1,i\neq j}^{m}\mathrm{tr}(V_i)$. Then similar to conventional SIR, tensor SIR can be developed based on the standardized scale $\mathcal{Z} = \mathcal{X} \times_1 V_1^{-1/2} \times_2 V_2^{-1/2} \times \cdots \times V_m^{-1/2}$. Suppose that $\mathcal{S}_{Y|\mathcal{Z}\circ m} = \mathrm{Span}(\bigotimes_{j=m}^{1}\beta_j)$. One can apply the algorithm in Section 3.1 to estimate $\beta$s using the standardized predictor $\mathcal{Z} = \mathcal{X} \times_1 \hat{V}_1^{-1/2} \times_2 \hat{V}_2^{-1/2} \times \cdots \times \hat{V}_m^{-1/2}$, where $\hat{V}_j = \hat{\Omega}_j = n^{-1}\sum_{i=1}^{n} \mathbf{X}_{(j)_i}\mathbf{X}_{(j)_i}^T$. The scalar $\prod_{i=1,i\neq j}^{m}\mathrm{tr}(V_i)$ is not essential for the CTS estimation. Therefore, by the equivalence property, the CTS of $Y|\mathcal{X}$ is estimated by $(\bigotimes_{j=m}^{1}\hat{\Omega}_j^{-1/2})\hat{\mathcal{S}}_{Y|\mathcal{Z}\circ m}$. Since this procedure relies on the Kronecker structure on $\mathrm{cov}[\mathrm{vec}(\mathcal{X})]$, we call it as Kronecker tensor SIR, shortened as tensor SIR-K.

When Condition (11) holds but $\mathrm{cov}[\mathrm{vec}(\mathcal{X})]$ does not have the Kronecker structure, tensor SIR-K tends to provide biased estimation because the transformation $\mathcal{Z} = \mathcal{X} \times_1 V_1^{-1/2} \times_2 V_2^{-1/2} \times \cdots \times V_m^{-1/2}$ does not standardize the predictor properly.

## 4. Large sample properties

In this section, we show that the tensor SIR estimator is $\sqrt{n}$ consistent for the CTS, and that it is asymptotically normal under certain regularity conditions. Let $\Gamma_1 = (\gamma_{1,1}, \ldots, \gamma_{1,d_1})$, $\Gamma_2 = (\gamma_{2,1}, \ldots, \gamma_{2,d_2}), \ldots, \Gamma_m = (\gamma_{m,1}, \ldots, \gamma_{m,d_m})$ be the column expressions of the minimizers of (15). Then the bases of the CTS can be represented as $\{\bigotimes_{k=m}^{1}(\Omega_k^{-1}\gamma_{k,j_k}), j_1 = 1, \ldots, d_1, \ldots, j_m = 1, \ldots, d_m\}$, which are the principal directions of tensor SIR. Since $\Omega_k$ are estimated at rate $\sqrt{n}$, the rate for estimating the CTS is determined by how well $\Gamma_k$ can be estimated. Therefore, we first study the asymptotic properties for $\Gamma_k$, $k \in \mathcal{M}$.

Let $\zeta_k = \{(i,j) : \mathrm{vec}(\mathbf{X}_{(k)}) = \prod_{i,j} T_{i,j}\mathrm{vec}(\mathcal{X})\}$ be a set of indexes to transform $\mathrm{vec}(\mathbf{X}_{(k)})$ to $\mathrm{vec}(\mathcal{X})$, where $T_{i,j}$ is an elementary matrix produced by exchanging row $i$ and row $j$ of the identity matrix $I_u$. Denote the transformation matrices $\prod_{(i,j)\in\zeta_k} T_{i,j}$ by $T_k$, $k \in \mathcal{M}$. It follows that $\mathrm{vec}(\mathbf{X}_{(k)}) = T_k\mathrm{vec}(\mathcal{X})$. The following lemma provides an alternative expression of $\Sigma_k$, $k \in \mathcal{M}$.

**Lemma 3.** *The matrix* $\Sigma_k = \mathrm{E}[\mathrm{E}(\mathbf{X}_{(k)}|Y)(\bigotimes_{j=m,j\neq k}^{1} P_{\Gamma_j})\mathrm{E}(\mathbf{X}_{(k)}^T|Y)]$ *is equal to*

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

*where* $\Omega = \mathrm{cov}\{\mathrm{E}[\mathrm{vec}(\mathcal{X}) \mid Y]\}$.

Let $\lambda_{k,1} > \lambda_{k,2} > \cdots > \lambda_{k,d_k} \geq 0$ be the first $d_k$ eigenvalues of $\Sigma_k$, $k \in \mathcal{M}$. According to Proposition 2, the columns of $\Gamma_k$ consist of the corresponding eigenvectors of $\Sigma_k$. Therefore, the following equation system

$$\left\{\sum_{j_1=1}^{d_1}\cdots\sum_{j_{k-1}=1}^{d_{k-1}}\sum_{j_{k+1}=1}^{d_{k+1}}\cdots\sum_{j_m=1}^{d_m}\left[\left(\bigotimes_{l=m,l\neq k}^{1}\gamma_{l,j_l}\right)\otimes I_{p_k}\right]^T T_k\Omega T_k^T\left[\left(\bigotimes_{l=m,l\neq k}^{1}\gamma_{l,j_l}\right)\otimes I_{p_k}\right]\right\}\gamma_{k,j_k} = \lambda_{k,j_k}\gamma_{k,j_k} \tag{17}$$

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

From (17), all of the leading eigenvalues and eigenvectors $\{\lambda_{k,j_k}, \gamma_{k,j_k}, j_k = 1, \ldots, d_k, k \in \mathcal{M}\}$ can be expressed as functions of $\Omega$. Correspondingly, the sample estimates $\hat{\Gamma}_1, \ldots, \hat{\Gamma}_m$ are functions of $\hat{\Omega}$, where $\hat{\Omega} = \sum_{s=1}^{H}\frac{n_s}{n}\mathrm{vec}(\bar{\mathcal{X}}_s)\mathrm{vec}(\bar{\mathcal{X}}_s)^T$. Hence if the asymptotic distribution of $\hat{\Omega}$ is obtainable, the statistical properties of $\hat{\Gamma}_k$ can be derived based on a delta method. We adopt the idea of Zhu and Ng [24] to establish the asymptotics for $\hat{\Omega}$.

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

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

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

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

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

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

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

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

**Assumption 4.** $g(Y)$ is non-expansive in the metric of $G(Y)$ on both sides of a positive number $\delta_0$, such that $G^{4+b}(t)P(y > t) \to 0$ as $t \to \infty$.

**Lemma 4.** *Given Assumptions 1–4 with* $b > 0$, *when* $c = O(n^\tau)$, *where* $\tau = 1/2 - \max\{2r, 2/(4+b)\} > 0$, $\sqrt{n}[\mathrm{vec}(\hat{\Omega} - \Omega)]$ *converges in distribution to a normal random vector* $W$ *with mean zero and covariance matrix*

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

Based on the relationship between $\hat{\Gamma}_1, \dots, \hat{\Gamma}_m$ and $\hat{\Omega}$ and the asymptotic results from Lemma 4, the asymptotic distribution of $\hat{\Gamma}_1, \dots, \hat{\Gamma}_m$ is then obtained.

**Theorem 1.** *Under the linearity condition (11) and the conditions in Lemma 4,*

$$\sqrt{n}[\mathrm{vec}(\hat{\Gamma}_1, \dots, \hat{\Gamma}_m) - \mathrm{vec}(\Gamma_1, \dots, \Gamma_m)]$$

*converges in distribution to* $J_m W$, *where* $J_m = \{\mathrm{vec}(\Gamma_1)/\partial \mathrm{vec}(\Omega)^T, \dots, \mathrm{vec}(\Gamma_m)/\partial \mathrm{vec}(\Omega)^T\}$ *with*

$$\partial \gamma_{k,j_k}/\partial \mathrm{vec}(\Omega) = \left\{ \gamma_{k,j_k} \otimes \mathrm{vec}\left( \bigotimes_{j=m,\, j\neq k}^{1} P_{\Gamma_j} \right) \otimes \left\{ \lambda_{k,j_k} I_{p_k} - \mathrm{E}\left[ \mathrm{E}(\mathbf{X}_{(k)} \mid Y) \left( \bigotimes_{j=m,\, j\neq k}^{1} P_{\Gamma_j} \right) \mathrm{E}(\mathbf{X}_{(k)}^T \mid Y) \right] \right\}^{+} \right\}^{T}$$
$$\times (K_{p_k, u_{-k}} T_k \otimes T_k) \quad (j_k = 1, \dots, d_k, k \in \mathcal{M}).$$

The important point of Theorem 1 is the consistency and asymptotic normality of the estimates $\hat{\Gamma}_k, k \in \mathcal{M}$. This proves that tensor SIR provides $\sqrt{n}$ consistent estimation for the CTS as $\hat{\Omega}_k$ converges to $\Omega_k$ at rate $\sqrt{n}$. The following theorem gives the asymptotic normality of the tensor SIR estimator. Let $\Sigma = \mathrm{cov}[\mathrm{vec}(\mathcal{X})]$ and let $Q = \mathrm{E}\{\mathrm{cov}[\mathrm{vec}(\mathcal{X}) \mid Y]\}$.

**Theorem 2.** *Under the linearity condition (11) and the conditions in Lemma 4,*

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

*converges in distribution to* $HW_1$, *where* $W_1$ *follows a multivariate normal distribution with mean zero and covariance matrix* $N(\mathbf{0}, \mathrm{cov}[(\mathrm{vec}(\mathcal{X})^T \otimes \mathrm{vec}(\mathcal{X})^T, \epsilon^T \otimes \epsilon^T)^T])$, *and*

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

The expression of $H$ is given in the Appendix. Theorem 2 shows an important statistical property of tensor SIR, since no asymptotic studies have been given in the literature for higher-order SDR, including dimension folding SIR and longitudinal SIR.

## 5. Connections with other higher-order SDR methods

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

### 5.1. Comparison of different linearity conditions

In literature, the higher-order SDR methods, such as dimension folding SIR, longitudinal SIR and dimension folding PCA and PFC, require a linearity condition imposed directly on $\mathrm{vec}(\mathbf{X})$. That is, $\mathrm{E}[\mathrm{vec}(\mathbf{X}) | \eta^T \mathrm{vec}(\mathbf{X})]$ is linear function of $\eta^T \mathrm{vec}(\mathbf{X})$ for the basis matrix $\eta \in \mathbb{R}^{p_1 p_2 \times d} (d \leq p_1 p_2)$ of the CTS. Since $\eta$ is usually unknown, this condition is generally satisfied when the distribution of $\mathrm{vec}(\mathbf{X})$ is elliptically symmetric [16]. In comparison, two-tensor SIR uses the tensor-formed linearity condition (4). It requires elliptical symmetry only along each mode of $\mathbf{X}$ but the different modes need not be jointly elliptically symmetric. However, joint elliptical symmetry is requisite when we directly impose the linearity condition on $\mathrm{vec}(\mathbf{X})$. In addition, longitudinal SIR further requires the Kronecker structure (16) on $\mathrm{cov}[\mathrm{vec}(\mathbf{X})]$ for $m = 2$. It can be shown that when the linearity condition holds for $\mathrm{vec}(\mathbf{X})$ and $\mathrm{cov}\{\mathrm{vec}(\mathbf{X})\}$ has the Kronecker structure, Condition (4) is satisfied. Yet the reverse direction does not necessarily hold. In other words, when the tensor-formed linearity condition is satisfied, $\mathrm{cov}[\mathrm{vec}(\mathbf{X})]$ needs not be Kronecker structured. The simulation in Section 6.1 can serve as a counter example. For higher-order tensors, the tensor-formed linearity condition is weaker.

### 5.2. Two-tensor SIR and dimension folding SIR

Dimension folding SIR relies on the linearity condition on $\mathrm{vec}(\mathbf{X})$. Under this condition, Li et al. [17] showed that $U(Y) = \Sigma^{-1}\mathrm{E}[\mathrm{vec}(\mathbf{X})|Y]$ is contained in a subspace of the CTS, where $\Sigma = \mathrm{cov}[\mathrm{vec}(\mathbf{X})]$. This subspace is called the Kronecker envelope of the $U(Y)$, denoted as $\mathcal{E}^{\otimes}(U)$. It is the Kronecker product of two smallest subspaces $\mathcal{S}_{\circ U}$ and $\mathcal{S}_{U\circ}$ such that $\mathrm{Span}\{U(Y)\} \subseteq \mathcal{S}_{U\circ} \otimes \mathcal{S}_{\circ U}$, for any $Y$. Dimension folding SIR then estimates $\mathcal{E}^{\otimes}(U)$ by minimizing

$$\mathrm{E} \parallel U(Y) - (b \otimes a)\omega_Y \parallel_{\mathrm{F}}^2, \tag{18}$$

over $a \in \mathbb{R}^{p_L \times u_L}(u_L \leq d_L)$, $b \in \mathbb{R}^{p_R \times u_R}(u_R \leq d_R)$, and $\omega_Y \in \mathbb{R}^{u_L u_R}$, where $\mathrm{Span}(b) = \mathcal{S}_{U\circ}$, $\mathrm{Span}(a) = \mathcal{S}_{\circ U}$, and $\omega_Y$ is a vector-valued latent function of $Y$. When $\Sigma = I_{p_L p_R}$, (18) can be equivalently formulated as

$$\mathrm{E} \parallel \mathrm{E}(\mathbf{X}|Y) - a\nu_Y b^T \parallel_{\mathrm{F}}^2, \tag{19}$$

where $\nu_Y \in \mathbb{R}^{u_L \times u_R}$ is the matrix form of $\omega_Y$. Given $a$ and $b$, the minimizer of (19) over all $\nu_Y$ is $\nu_Y = (a^T a)^{-1} a^T \mathrm{E}(\mathbf{X}|Y) b (b^T b)^{-1}$. Therefore, the objective function of dimension folding SIR reduces to

$$\mathrm{E} \parallel \mathrm{E}(\mathbf{X}|Y) - P_a \mathrm{E}(\mathbf{X}|Y) P_b \parallel_{\mathrm{F}}^2, \tag{20}$$

which is equivalent to the objective function of two-tensor SIR. Hence a key difference between the two methods lies in how the predictors are standardized. Dimension folding SIR standardizes the predictors by the large covariance matrix cov[vec($\mathbf{X}$)]. When cov[vec($\mathbf{X}$)] does not have a Kronecker structure, the objective function of dimension folding SIR generally cannot be converted to a matrix form. In contrast, two-tensor SIR performs standardization by using the two smaller matrices $\mathrm{E}(\mathbf{X}\mathbf{X}^T)$ and $\mathrm{E}(\mathbf{X}^T\mathbf{X})$. This standardization leads to the following desirable properties of two-tensor SIR: (1) It alleviates computation cost and avoids inversion of a large covariance matrix. (2) It provides eigenbased subspace estimation that can be easily generalized to $m$-mode tensor predictors. (3) Two-tensor SIR shows good theoretical properties. It provides $\sqrt{n}$ consistent estimator for the CTS and is asymptotically normal under certain regularity conditions. (4) When the predictor is vector-valued, tensor SIR coincides with conventional SIR.

### 5.3. Two-tensor SIR and longitudinal SIR

Longitudinal SIR [21] addresses dimension reduction for data with longitudinal predictors, a special form of matrix-valued predictors. It estimates the SIR directions, or the basis of the CTS, by $\mathrm{Span}(\hat{\Omega}_2^{-1/2}\hat{\eta}_2 \otimes \hat{\Omega}_1^{-1/2}\hat{\eta}_1)$, where the columns of $\hat{\eta}_1$ and $\hat{\eta}_2$ contain the leading $d_1$ and $d_2$ eigenvectors of the sample estimates of $\Psi_1 = \mathrm{E}[\mathrm{E}(\mathbf{Z}|Y)\mathrm{E}(\mathbf{Z}^T|Y)]$ and $\Psi_2 = \mathrm{E}[\mathrm{E}(\mathbf{Z}^T|Y)\mathrm{E}(\mathbf{Z}|Y)]$, respectively. Here $\mathbf{Z}$ denotes the standardized predictor. Longitudinal SIR requires the linearity condition on vec($\mathbf{X}$) and the Kronecker structure (16) on cov[vec($\mathbf{X}$)]. Thus, it is more restrictive than two-tensor SIR. In contrast, two-tensor SIR-K requires the equivalent conditions of longitudinal SIR, but it uses the leading eigenvectors of the sample estimates of $\Sigma_1 = \mathrm{E}[\mathrm{E}(\mathbf{Z}|Y)P_{\Gamma_2}\mathrm{E}(\mathbf{Z}^T|Y)]$ and $\Sigma_2 = \mathrm{E}[\mathrm{E}(\mathbf{Z}^T|Y)P_{\Gamma_1}\mathrm{E}(\mathbf{Z}|Y)]$ for estimation. Two-tensor SIR-K has more efficiency gains because it projects $\mathrm{E}(\mathbf{Z}|Y)$ onto each sufficient reduction direction before estimating the other direction, which can remove redundant information in estimation. This provides intuition regarding the asymptotic efficiency shown in Section 5.4.

### 5.4. Two-tensor SIR and dimension folding PFC

Dimension folding PFC [7] is a model-based SDR method for matrix-valued predictors. It gains efficiency by flexibly modeling the conditional mean function $\mathrm{E}(\mathbf{X}|Y)$. When $\mathbf{X}|Y$ is matrix normal, dimension folding PFC inherits optimal asymptotic properties from maximum likelihood estimation. The matrix normal distribution is formulated as $\mathbf{N}_{p_1 \times p_2}(0, M_1, M_2)$, where $M_1 = \mathrm{E}(\mathbf{X}\mathbf{X}^T)/\mathrm{tr}(M_2)$ and $M_2 = \mathrm{E}(\mathbf{X}^T\mathbf{X})/\mathrm{tr}(M_1)$ are the column and row covariance matrices of $\mathbf{X}$. For more background on this distribution, see [6]. The following proposition establishes the connection between two-tensor SIR and dimension folding PFC.

**Proposition 3.** *Under the matrix normality of* $\mathbf{X}|Y$*, when* $Y$ *is categorical and* cov[vec($\mathbf{X}$)] *has the Kronecker structure* (16) $(m = 2)$*, two-tensor SIR-K is equivalent to dimension folding PFC and thus provides the MLE of the CTS.*

Proposition 3 implies that when $\mathbf{X}|Y$ is matrix-normal and cov[vec($\mathbf{X}$)] satisfies the Kronecker condition (16), two-tensor SIR-K provides the optimal estimation for the CTS.

## 6. Simulation

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

$$\|P_{\hat{\mathcal{S}}} - P_{\mathcal{S}}\|_{\mathrm{F}}, \tag{21}$$

where $P_{\hat{\mathcal{S}}}$ is the projection onto the estimate $\hat{\mathcal{S}}$ of the CTS $\mathcal{S}$, to measure the distance between the estimated and true projection matrices.

### 6.1. Two-mode tensor predictors

We first consider the simulation setup from Li et al. [17]. Let $d_1 = d_2 = 2$ and $p_1 = p_2 = p = 5, 10$. The response $Y$ is a binary variable and was generated from the Bernoulli distribution with success probability equal to 0.5. The matrix-valued

**Table 1**
Comparison of the CTS estimation among different higher-order SDR methods for two-mode tensor predictors when $a = 4$. Each entry is the mean of the estimation errors (21) over 500 samples.

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

**Table 2**
Comparison of the CTS estimation among different higher-order SDR methods for two-mode tensor predictors when $a = 50$. Each entry is the mean of the estimation errors (21) over 500 samples.

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

predictor $\mathbf{X}$ was generated based on the conditional distribution of $\mathbf{X}$ given $Y$, which was taken to be multivariate normal with conditional mean

$$\mathrm{E}(\mathbf{X}|Y = 0) = \mathbf{0}_{p\times p}, \qquad \mathrm{E}(\mathbf{X}|Y = 1) = \begin{pmatrix} a\mathbf{I}_2 & \mathbf{0}_{2\times(p-2)} \\ \mathbf{0}_{(p-2)\times 2} & \mathbf{0}_{(p-2)\times(p-2)} \end{pmatrix}$$

and conditional variance

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

where $A$ is the index set $\{(1, 1), (1, 2), (2, 1)\}$. Let $e_i \in \mathbb{R}^p$ be the vector with $i$-th element equal to 1 and all other elements equal to zero. It can be shown that the CTS is $\Gamma_2 \otimes \Gamma_1$, where $\Gamma_1 = \Gamma_2 = (e_1, e_2)$, and the linearity condition (4) holds. However, $\mathrm{cov}[\mathrm{vec}(\mathbf{X})] = \mathrm{E}[\mathrm{cov}\{\mathrm{vec}(\mathbf{X}) \mid Y\}] + \mathrm{cov}[\mathrm{E}\{\mathrm{vec}(\mathbf{X}) \mid Y\}]$ does not exactly have a Kronecker structure.

We applied two-tensor SIR, two-tensor SIR-K, dimension folding SIR and longitudinal SIR for the simulated data and evaluated their estimation accuracy based on (21). The comparison results for $a = 4$ are listed in Table 1 with the shortened names 2-T SIR, 2-T SIR-K, DF-SIR and L-SIR, respectively. It can be seen that two-tensor SIR provides the most accurate estimation. Two-tensor SIR-K and longitudinal SIR perform similarly to Two-tensor SIR because $\mathrm{cov}[\mathrm{vec}(\mathbf{X})]$ has a close Kronecker structure when $a = 4$. All of the three methods outperform dimension folding SIR as the latter is computed based on $\mathrm{vec}(\mathbf{X})$ that requires more parameters in estimation. We also applied conventional SIR and observed much larger estimation errors compared to the other methods. Thus it is not a good competitor and is not listed in the table.

We now vary the conditional mean $\mathrm{E}(\mathbf{X}|Y = 1)$ by choosing $a = 50$ and keep all other settings the same. In this case, the signal of $\mathrm{cov}[\mathrm{E}\{\mathrm{vec}(\mathbf{X}) \mid Y\}]$ is strong and thus the Kronecker structure of $\mathrm{cov}[\mathrm{vec}(\mathbf{X})]$ is violated significantly. Table 2 shows that the performance of tensor SIR is not affected but the accuracy of tensor SIR-K and longitudinal SIR is dramatically decreased. The latter two methods highly rely on the Kronecker decomposition of $\mathrm{cov}[\mathrm{vec}(\mathbf{X})]$, resulting in less efficiency when this assumption is not well satisfied.

### 6.2. *Three-mode tensor predictors*

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

**Table 3**
Comparison of the CTS (or CS) estimation among different SDR methods for three-mode tensor predictors when $a = 4$. Each entry is the mean of the estimation errors over 500 samples.

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

**Table 4**
Comparison of the CTS (or CS) estimation among different SDR methods for three-mode tensor predictors when $a = 50$. Each entry is the mean of the estimation errors over 500 samples.

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

predictor $\mathcal{X}$ was generated based on the conditional distribution of $\mathcal{X}$ given $Y$ that is multivariate normal with conditional mean

$$\mathrm{E}\{\mathcal{X}[,,1]|Y=0\} = \mathrm{E}\{\mathcal{X}[,,2]|Y=0\} = \mathrm{E}\{\mathcal{X}[,,2]|Y=1\} = \mathbf{0}_{p\times 2p},$$
$$\mathrm{E}\{\mathcal{X}[,,1]|Y=1\} = \begin{pmatrix} a\mathbf{I}_2 & \mathbf{0}_{2\times(p-2)} \\ \mathbf{0}_{(p-2)\times 2} & \mathbf{0}_{(p-2)\times(p-2)} \end{pmatrix}$$

and conditional variance

$$\mathrm{var}\{\mathcal{X}[i,j,k]|Y=0\} = \begin{cases} 0.1 & \text{if } (i,j)\in A, \\ 1 & \text{if } (i,j)\notin A, \end{cases}$$
$$\mathrm{var}\{\mathcal{X}[i,j,k]|Y=1\} = \begin{cases} 1.5 & \text{if } (i,j)\in A, \\ 1 & \text{if } (i,j)\notin A, \end{cases}$$

where $A$ is the index set $\{(1, 1, 1), (1, 2, 1), (2, 1, 1), (1, 1, 2), (1, 2, 2), (2, 1, 2)\}$. It can be seen that the CTS of $\mathcal{X}|Y$ is $\Gamma_3 \otimes \Gamma_2 \otimes \Gamma_1$, where $\Gamma_1 = \Gamma_2 = (e_1, e_2)$, the same as that in the two-mode case, and $\Gamma_3 = (1, 0)^T$. Similar to the two-mode example, the tensor linearity condition (11) is satisfied for the data. However, cov[vec($\mathcal{X}$)] cannot be exactly decomposed into the Kronecker structure (16) ($m = 3$). The larger the value of $a$, the stronger the violation of the Kronecker assumption. As dimension folding SIR and longitudinal SIR were proposed only for matrix-valued predictors, we applied tensor SIR, tensor SIR-K to estimate the CTS and added the results of conventional SIR for comparison. When $p_1 = p_2 = 10$, the sample covariance matrix $\hat{\Sigma} = \widehat{\mathrm{cov}}[\mathrm{vec}(\mathcal{X})]$ is singular and the ridge-regression-type inverse $(\hat{\Sigma} + 0.001I_{200})^{-1}$ is used for conventional SIR. The results based on (21) were summarized in Table 3 for $a = 4$, where $\hat{\mathcal{S}}$ and $\mathcal{S}$ in (21) indicate the estimated and the true CS of $Y|\mathrm{vec}(\mathcal{X})$ for conventional SIR.

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

We now vary the conditional mean $\mathrm{E}\{\mathcal{X}[,,1]|Y = 1\}$ using $a = 50$ so cov[vec($\mathcal{X}$)] deviates significantly from the Kronecker structure. From Table 4, we see that tensor SIR outperforms tensor SIR-K noticeably as it does not impose any constraint on cov[vec($\mathcal{X}$)]. Tensor SIR-K highly depends on the Kronecker constraint. It can perform worse than conventional SIR when cov{vec($\mathcal{X}$)} deviates strongly from the Kronecker structure.

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

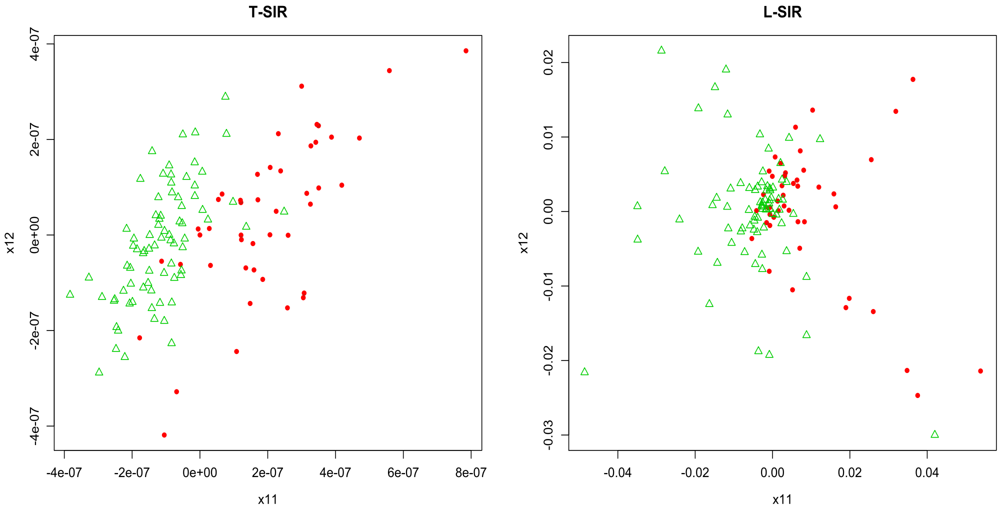

**Fig. 1.** Scatter plots by dimension reduced predictors $X_{11}, X_{12}$ with $(s_L, s_R) = (95, 15)$. The triangles indicate alcoholic subjects. The circles represent nonalcoholic subjects.

## 7. Application

We now analyze the EEG data using tensor SIR. The EEG data contains 122 subjects, which are divided into alcoholic and control groups with 77 subjects and 45 subjects respectively. For each subject, the predictor contains measurements from 64 channels of electrodes placed on the subject's scalp and sampled at 256 times. The 64 sites were matched among individuals. Thus, the predictor $\mathbf{X}$ is formed as a matrix of dimension $256 \times 64$, and the response $Y$ is a binary variable indicating groups. Since the row and column dimensions of the predictor are moderately large, it is very likely that only a few row linear combinations and a few column linear combinations are relevant to classify the response. Moreover, as $n \ll p_L \times p_R$, conventional classification methods, such as linear discriminate analysis (LDA) and quadratic discriminate analysis (QDA) cannot be directly applied to the data. Consequently, higher-order SDR tools are desirable to reduce the predictor's row and column dimensions.

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

Since two-tensor SIR circumvents vectorization of the predictors, it can be directly applied to the original EEG data without any prescreening work. In this case, it correctly classified 88 out of 122 subjects with $(d_L, d_R) = (1, 2)$, while longitudinal SIR classified 75 subjects under the same setting. Dimension folding SIR, however, cannot be directly applied due to the high dimension of $\text{vec}(\mathbf{X})$, which is equal to $256 \times 64 = 16, 384$. In order to make comparison with dimension folding SIR, we applied the procedure in Li et al. [17] to prescreen the predictor's row and column dimensions to $(s_L, s_R) = (15, 15)$ first and then performed higher-order SDR and QDA with $(d_1, d_2) = (1, 2)$. As a result, two-tensor SIR correctly classified 97 subjects out of total 122 subjects based on the reduced two-dimensional predictor vector $\hat{\Gamma}_1^{-1}\hat{\Omega}_1^{-1}X\hat{\Omega}_2^{-1}\hat{\Gamma}_1 = (x_{11}, x_{12})$. Both dimension folding SIR and longitudinal SIR provided 92 out of 122 correct decisions. We also tried to prescreen the predictors to other different dimensions, such as $(s_L, s_R) = (10, 10), (30, 30), (95, 15)$. In all cases, tensor SIR showed more accurate classification rates than the other two methods. Fig. 1 demonstrates the separation of the two groups by two-tensor SIR and longitudinal SIR when $(s_L, s_R) = (95, 15)$. Tensor SIR shows better separation.

## 8. Discussion

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

To determine the reduced dimensions, one can apply cross-validation to select $(d_1, d_2, \ldots, d_m)$ that provides the smallest prediction or classification error. One can also apply the procedure described in Dong and Li [8] that adapts the bootstrap idea in [23] to evaluate the multivariate correlation [9] between the original estimated principal tensor SIR components $C = (\bigotimes_{j=1}^{m} \hat{\Gamma}_j)\mathrm{vec}(\mathcal{X})$ and the estimated bootstrap tensor SIR components $C_b = (\bigotimes_{j=1}^{m} \hat{\Gamma}_j^b)\mathrm{vec}(\mathcal{X})$. The multivariate correlation is defined as

$$\{\mathrm{var}(C_b)\}^{-\frac{1}{2}}\mathrm{cov}(C_b, C)\{\mathrm{var}(C)\}^{-1}\mathrm{cov}(C, C_b)\{\mathrm{var}(C_b)\}^{-\frac{1}{2}}, \tag{22}$$

where $\hat{\Gamma}_j^b$ is the $b$-th bootstrap sample estimate of $\Gamma_j$. Let $\lambda_1, \dots, \lambda_l$ be the nonzero eigenvalues of (22) and let $r^2(C_b, C) = \prod_{i=1}^{l} \lambda_i$ be the eigenbased correlation coefficient. The optimal dimension combination is then selected to maximize the average bootstrap sample correlation $\bar{r}^2 = \frac{1}{B}\sum_{b=1}^{B} r^2(C_b, C)$.

As pointed out by a referee, some of the results in Section 2.2 overlap with results in Chapter 8 of Kim's Ph.D. thesis [11]. In particular, the two assumptions in (4) are the same as the two conditions in Assumption 3, Chapter 8, of [11] and our Lemma 1 is equivalent to Kim's Lemma 2. Although the problem settings are very similar, the work shown in Section 2.2 was proposed independently, and the two-tensor SIR approach is different from the method proposed in Kim's thesis. Two-tensor SIR does not impose the Kronecker covariance structure on the predictors, and it is an adaptive procedure that employs an iterative algorithm to estimate each component of the CTS based on all others.

The core of this article focuses on extending SIR to tensor-valued predictors. However, the same logic can be used to study tensor SAVE [5] and other tensor SDR methods (see, for instance, [4,18]) based on the tensor-formed linearity condition (11). Furthermore, we can relax the linearity condition (11) and use the recent results in [19,20] to develop semiparametric tensor SDR methods. These extensions are under investigation.

## Acknowledgments

This work was supported in part by National Science Foundation grant DMS-1007547. The authors are grateful to the Associate Editor and the reviewers for their insightful comments that have led to significant improvement of the article. In particular, the authors thank a referee for a reference to Kim's thesis [11] and for pointing out the connection between two-tensor SIR and dimension folding SIR described in Section 5.2. The authors also thank Bing Li and Min Kyung Kim for sharing their code and the EEG data.

## Appendix

**Proof of Lemmas 1 and 2.** We demonstrate the proof of 1 first. For convenience, we only show $A = \Omega_1\alpha(\alpha^T\Omega_1\alpha)^{-1}$ since the expression of $B$ can be similarly derived. Consider

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

Based on the fact that $\mathrm{E}(\mathbf{X}|\alpha^T\mathbf{X}) = A\alpha^T\mathbf{X}$, we have $\mathrm{E}\{\mathrm{E}(\mathbf{X}|\alpha^T\mathbf{X})(\alpha^T\mathbf{X})^T\} = \mathrm{E}(A\alpha^T\mathbf{X}\mathbf{X}^T\alpha) = A\alpha^T\Omega_1\alpha$. Therefore, $A = \Omega_1\alpha(\alpha^T\Omega_1\alpha)^{-1}$.

The proof of Lemma 2 can be done based on the same logic and thus is omitted.

**Proof of Propositions 1 and 2.** To prove Proposition 1, it is easy to see that the objective function (10) is equivalent to

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

Thus, minimizing (10) is the same as maximizing $L = \mathrm{tr}\{\mathrm{E}[P_{G_2}\mathrm{E}(\mathbf{X}^T|Y)P_{G_1}\mathrm{E}(\mathbf{X}|Y)]\} = \mathrm{tr}\{G_1^T\mathrm{E}[\mathrm{E}(\mathbf{X}|Y)P_{G_2}\mathrm{E}(\mathbf{X}^T|Y)]G_1\}$. Then for fixed $G_2$, the minimizer $\hat{\Gamma}_1$ over $G_1$ is obtained by choosing its columns to be the first $d_1$ eigenvectors of $\mathrm{E}[\mathrm{E}(\mathbf{X}|Y)P_{G_2}\mathrm{E}(\mathbf{X}^T|Y)]$. Similarly, $L$ can be written as $\mathrm{tr}\{G_2^T\mathrm{E}[\mathrm{E}(\mathbf{X}^T|Y)P_{G_1}\mathrm{E}(\mathbf{X}|Y)]G_2\}$ and thus the minimizer $\hat{\Gamma}_2$ can be similarly proved.

The proof of Proposition 2 can be similarly done since the objective function (15) is equivalent to

$$\mathrm{E}\left\|\mathrm{E}[\mathbf{X}_{(k)}|Y] - P_{\Gamma_k}\mathrm{E}[\mathbf{X}_{(k)}|Y]\left(\bigotimes_{j=m,\, j\neq k}^{1} P_{\Gamma_j}\right)\right\|_{\mathrm{F}}^2, \quad k \in \mathcal{M}. \tag{A.1}$$

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

**Proof of Lemmas 3 and 4.** For Lemma 3, we consider the column-wise expression of $\bigotimes_{j=m,\, j\neq k}^{1} \Gamma_j$,

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

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

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

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

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

Hence

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

The proof of Lemma 4 can be done by following the proofs of Theorems 1 and 2 in [24]. Note that

$$\sqrt{n}(\hat{\Omega}-\Omega) = \sqrt{n}[(\hat{\Sigma}-\hat{Q})-(\Sigma-Q)] = -M_1 - M_2 - M_3 + M_4^{(1)} - M_4^{(2)},$$

where $M_1, M_2, M_3$ and $M_4^{(2)}$ are the same defined as $T_1, T_2, T_3$ and $T_4^{(2)}$ in [24], only with the predictor $x$ replaced by $\mathrm{vec}(\mathcal{X})$, and $M_4^{(1)} = n^{-\frac{1}{2}}\sum_{i=1}^{n}[\mathrm{vec}(\mathcal{X}_i)\mathrm{vec}(\mathcal{X}_i)^T - \varepsilon_i\varepsilon_i^T - \Omega]$. Under the conditions in Lemma 4, the elements in $M_1, M_2, M_3$ and $M_4^{(2)}$ are all equal to $o_p(1)$, as shown in [24], and $\mathrm{vec}(M_4^{(1)})$ converges to $N(\mathbf{0}, \mathrm{cov}[\mathrm{vec}(\mathcal{X})\otimes\mathrm{vec}(\mathcal{X}) - \epsilon\otimes\epsilon])$.

**Proof of Theorem 1.** The main procedure is to show the gradient matrices $\partial\gamma_{k,j_k}/\partial\mathrm{vec}(\Omega)$. Inspired by Hung et al. [10], we apply the perturbation method to derive these results. Let $\Omega$ be perturbed to $\Omega(\varepsilon) = \Omega + \varepsilon\Omega^* + o(\varepsilon)$. With this perturbation, the eigenequation system becomes

$$\begin{aligned}\Sigma_k(\varepsilon)\gamma_{k,j_k}(\varepsilon) = &\left\{\sum_{j_1=1}^{d_1}\cdots\sum_{j_{k-1}=1}^{d_{k-1}}\sum_{j_{k+1}=1}^{d_{k+1}}\cdots\sum_{j_m=1}^{d_m}\left[\left(\bigotimes_{l=m,l\neq k}^{1}\gamma_{l,j_l}(\varepsilon)\right)\otimes I_{p_k}\right]^T T_k\Omega(\varepsilon)T_k^T\cdot\right. \\ &\left.\times\left[\left(\bigotimes_{l=m,l\neq k}^{1}\gamma_{l,j_l}(\varepsilon)\right)\otimes I_{p_k}\right]\right\}\gamma_{k,j_k}(\varepsilon) = \lambda_{k,j_k}(\varepsilon)\gamma_{k,j_k}(\varepsilon),\quad j_k = 1,\ldots,d_k,\ k\in\mathcal{M},\end{aligned} \tag{A.2}$$

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

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

where

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

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

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

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

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

$$\mathbf{X}_{(k)}\mid Y = \Gamma_k\nu_y\left(\bigotimes_{l=m,l\neq k}^{1}\Gamma_l\right)^T + e,$$

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

$$\mathrm{E}[\mathrm{E}(\mathbf{X}_{(k)}\mid Y)(\Lambda^*\Lambda^T+\Lambda\Lambda^{*^T})\mathrm{E}(\mathbf{X}_{(k)}^T\mid Y)] = \mathrm{E}\left\{\mathrm{E}(\mathbf{X}_{(k)}\mid Y)\left[\bigotimes_{l=m,l\neq k}^{1}(\Gamma_l\Gamma_l^{*^T}+\Gamma_l^*\Gamma_l^T)\right]\mathrm{E}(\mathbf{X}_{(k)}^T\mid Y)\right\}$$
$$= \mathrm{E}\left\{\left[\Gamma_k\nu_y\left(\bigotimes_{l=m,l\neq k}^{1}\Gamma_l\right)^T\right]\left[\bigotimes_{l=m,l\neq k}^{1}(\Gamma_l\Gamma_l^{*^T}+\Gamma_l^*\Gamma_l^T)\right]\left[\Gamma_k\nu_y\left(\bigotimes_{l=m,l\neq k}^{1}\Gamma_l\right)^T\right]^T\right\}$$
$$= \mathrm{E}\left\{\Gamma_k\nu_y\left[\bigotimes_{l=m,l\neq k}^{1}(\Gamma_l^{*^T}\Gamma_l+\Gamma_l^T\Gamma_l^*)\right]\nu_y^T\Gamma_k^T\right\}.$$

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

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

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

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

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

$$\Sigma_k^* = \sum_{j_1=1}^{d_1}\cdots\sum_{j_{k-1}=1}^{d_{k-1}}\sum_{j_{k+1}=1}^{d_{k+1}}\cdots\sum_{j_m=1}^{d_m}\left[\left(\bigotimes_{l=m,l\neq k}^{1}\gamma_{l,j_l}\right)\otimes I_{p_k}\right]^T T_k\Omega^* T_k^T\left[\left(\bigotimes_{l=m,l\neq k}^{1}\gamma_{l,j_l}\right)\otimes I_{p_k}\right]. \tag{A.3}$$

From (A.2), using the result of Lemma 2.1 in [22], we have

$$\gamma^*_{k,j_k} = \left\{\lambda_{k,j_k}I_{p_k} - \sum_{j_1=1}^{d_1}\cdots\sum_{j_{k-1}=1}^{d_{k-1}}\sum_{j_{k+1}=1}^{d_{k+1}}\cdots\sum_{j_m=1}^{d_m}\left[\left(\bigotimes_{l=m,l\neq k}^{1}\gamma_{l,j_l}\right)\otimes I_{p_k}\right]^T T_k\Omega T_k^T\cdot\left[\left(\bigotimes_{l=m,l\neq k}^{1}\gamma_{l,j_l}\right)\otimes I_{p_k}\right]\right\}^+\Sigma_k^*\gamma_{k,j_k}$$
$$= \left\{\lambda_{k,j_k}I_{p_k} - \mathrm{E}[\mathrm{E}(\mathbf{X}_{(k)}\mid Y)\left(\bigotimes_{j=m,j\neq k}^{1}P_{\Gamma_j}\right)\mathrm{E}(\mathbf{X}_{(k)}^T\mid Y)]\right\}^+\Sigma_k^*\gamma_{k,j_k}. \tag{A.4}$$

The combination of (A.3) and (A.4) gives

$$\begin{aligned}\partial \gamma_{k,j_k}/\partial \mathrm{vec}(\Omega) = & \left\{ \gamma_{k,j_k} \otimes \mathrm{vec}\left( \bigotimes_{j=m,\, j\neq k}^{1} P_{\Gamma_j} \right) \otimes \left\{ \lambda_{k,j_k} I_{p_k} - \mathrm{E}[\mathrm{E}(\mathbf{X}_{(k)} \mid Y) \left( \bigotimes_{j=m,\, j\neq k}^{1} P_{\Gamma_j} \right) \cdot \mathrm{E}(\mathbf{X}_{(k)}^T \mid Y)] \right\}^{+} \right\}^{T} \\ & \times (K_{p_k,u_{-k}} \otimes I_u)(T_k \otimes T_k),\end{aligned}$$

for $j_k = 1, \ldots, d_k$, $k =\in \mathcal{M}$. Then by applying the delta method and the result in Lemma 4, we finish the proof of Theorem 1.

**Proof of Theorem 2.** Theorem 2 is established based on the delta method. Since $\Gamma_k$, $k \in \mathcal{M}$, are functions of $\Omega$, they are functions of $(\mathrm{vec}(\Sigma)^T, \mathrm{vec}(Q)^T)^T$ as $\Omega = \Sigma - Q$. Moreover, it can be shown $\Omega_k$ are functions of $\Sigma$. Note that $\Omega_k = \mathrm{E}(\mathbf{X}_{(k)}\mathbf{X}_{(k)}^T) = \sum_{j=1}^{u_{-k}} \mathrm{E}(\mathbf{X}_{(k),j}\mathbf{X}_{(k),j}^T)$, where $\mathbf{X}_{(k),j}$ denotes the $j$-th column of $\mathbf{X}_{(k)}$. On the other hand,

$$\begin{aligned}\Sigma &= \mathrm{E}[\mathrm{vec}(\mathcal{X})\mathrm{vec}(\mathcal{X})^T] = \mathrm{E}[\mathrm{vec}(\mathbf{X}_{(1)})\mathrm{vec}(\mathbf{X}_{(1)})^T] = T_k \mathrm{E}[\mathrm{vec}(\mathbf{X}_{(k)})\mathrm{vec}(\mathbf{X}_{(k)})^T] T_k^T \\ &= T_k \mathrm{E}[(\mathbf{X}_{(k),1}^T, \ldots, \mathbf{X}_{(k),u_{-k}}^T)^T (\mathbf{X}_{(k),1}^T, \ldots, \mathbf{X}_{(k),u_{-k}}^T)] T_k^T.\end{aligned}$$

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

Under the conditions in Lemma 4, following the proof of Theorem 2 in [24], we have $\sqrt{n}\hat{\Omega} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \epsilon_i \epsilon_i^T$. Along with the fact that $\sqrt{n}\hat{\Sigma} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathrm{vec}(\mathcal{X}_i)\mathrm{vec}(\mathcal{X}_i)^T$, we have

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

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

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

converges in distribution to $HW_1$, where $H$ is the gradient matrix given in Theorem 2 with

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

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

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

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

**Proof of Proposition 3.** We consider the dimension folding PFC model (8.3) in [7]. When the range of the response is divided into $h$ slices, the fitting function $f(Y)$ in (8.3) is naturally determined as $(I(Y \in H_1) - n_1/n, I(Y \in H_2) - n_2/n, \ldots, I(Y \in H_h) - n_h/n)^T$. Let $\mathcal{S}_d(A)$ be the subspace spanned by the leading $d$eigenvectors of $A$ and $\mathcal{S}_d(A, B) = A^{-1/2}\mathcal{S}_d(A^{-1/2}BA^{-1/2})$. Based on Corollary 1 in Ding and Cook [7], the MLE of the CTS is equal to $\mathcal{S}_{d_2}(\hat{\Omega}_2, \hat{\Sigma}_{\mathrm{fit}_R}) \otimes \mathcal{S}_{d_1}(\hat{\Omega}_1, \hat{\Sigma}_{\mathrm{fit}_L})$, where

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

To prove Proposition 3, using the results in Section 3.2, it is sufficient to show that $\mathcal{S}_{d_1}(\hat{\Omega}_1, \hat{\Sigma}_{\mathrm{fit}_L}) = \mathrm{Span}(\hat{\Omega}_1^{-1/2}\hat{\beta}_1)$ and $\mathcal{S}_{d_2}(\hat{\Omega}_2, \hat{\Sigma}_{\mathrm{fit}_R}) = \mathrm{Span}(\hat{\Omega}_2^{-1/2}\hat{\beta}_2)$. We only demonstrate the first equation since the second one is satisfied based on the first equation. Since $\mathcal{S}_{d_1}(\hat{\Omega}_1, \hat{\Sigma}_{\mathrm{fit}_L}) = \hat{\Omega}_1^{-1}\mathrm{Span}_d\{n^{-1}\sum_{s=1}^{H} n_s\bar{\mathbf{X}}_s\hat{M}_2^{-1}\hat{\Gamma}_2\hat{\Gamma}_2^T\hat{M}_2^{-1}\bar{\mathbf{X}}_s^T\}$, it is equal to $\hat{\Omega}_1^{-1}\mathrm{Span}_d\{n^{-1}\sum_{s=1}^{H} n_s\bar{\mathbf{X}}_s\hat{\Omega}_2^{-1/2}\hat{\beta}_2\hat{\beta}_2^T\hat{\Omega}_2^{-1/2}\bar{\mathbf{X}}_s^T\}$ based on the equation $\hat{\Omega}_2^{-1/2}\hat{\beta}_2 = \hat{M}_2^{-1}\hat{\Gamma}_2$. This equation holds by initiating $\Gamma_{20}$, $\beta_{20}$, $\Omega_{20}$ and $M_{20}$ such that $\Omega_{20}^{-1/2}\beta_{20} = M_{20}^{-1}\Gamma_{20}$.

## References

[1] R.D. Cook, On the interpretation of regression plots, J. Amer. Statist. Assoc. 89 (1994) 177–190.
[2] R.D. Cook, Regression Graphics: Ideas for Studying Regressions Through Graphics, Wiley, New York, 1998.

[3] R.D. Cook, Fisher lecture: dimension reduction in regression (with discussion), Statist. Sci. 22 (2007) 1–26.

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

[5] R.D. Cook, S. Weisberg, Discussion of sliced inverse regression for dimension reduction, by k.-c. li, J. Amer. Statist. Assoc. 86 (1991) 328–332.

[6] D.J. De Waal, Matrix-valued distributions, in: S. Kotz, N.L. Johnson (Eds.), Encycl. Statist. Sci., 5, Wiley, New York, 1985, pp. 326–333.

[7] S. Ding, R.D. Cook, Dimension folding pca and pfc for matrix-valued predictors, Statist. Sinica 24 (2014) 463–492.

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

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

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

[11] M.K. Kim, On dimension folding of matrix or array valued statistical objects (Ph.D. thesis), Pennsylvania State University, 2010.

[12] T.G. Kolda, Multilinear Operators for Higher-order Decompositions. Tech. Report SAND2006-2081, Sandia National Laboratories, Albuquerque, New Mexico and Livermore, California, 2006.

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

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

[15] L.D. Lathauwer, B.D. Moor, J. Vandewalle, 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 (2001) 1324–1342.

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

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

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

[19] Y. Ma, L. Zhu, A semiparametric approach to dimension reduction, J. Amer. Statist. Assoc. 107 (2012) 168–179.

[20] Y. Ma, L. Zhu, Efficient estimation in sufficient dimension reduction, Ann. Statist. 41 (2013) 250–268.

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

[22] R. Sibson, Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling, J. R. Stat. Soc. Ser. B Stat. Methodol. 41 (1979) 217–229.

[23] Z. Ye, R. Weiss, Using the bootstrap to select one of a new class of dimension reduction methods, J. Amer. Statist. Assoc. 98 (2003) 968–978.

[24] L. Zhu, K.W. Ng, Asymptotics of sliced inverse regression, Statist. Sinica 5 (1995) 727–736.