Parsimonious Tensor Dimension Reduction

Xin Xing[*], Peng Zeng [†], Youhui Ye[*], Wenxuan Zhong[‡]

Abstract

Tensor data is emerging in many scientific applications, such as multi-tissue transcriptomics. In such cases, the covariates for each individual are no longer a vector. To apply traditional vector-based methods to this type of data, we need to either do the vectorization or analyze data marginally, which suffers a significant information loss. We propose a novel parsimonious tensor dimension reduction (pTDR) approach to directly link the response and tensor covariate through an unknown function g. In pTDR, the response variable, continuous or discrete, depends on K rank-one projections of the covariates, with the projections estimated via a sequential iterative dimension reduction algorithm. We further propose an asymptotic sequential statistical test to select the correct number of rank-one tensors. In contrast to the classic low-rank tensor regression, pTDR model is not restricted to the linear relationship between response and covariates. We apply pTDR to two modern genomic studies. We find that the gene expression of multiple tissues has a stronger association

[*] Department of Statistics, Virginia Tech

[†] Department of Mathematics & Statistics, Auburn University

[‡] Department of Statistics, University of Georgia.

with aging and obesity than was apparent using previous approaches. Numerical results demonstrate the advantages of pTDR over competitors in terms of prediction accuracy and computing efficiency. Our software is publicly available on GitHub (https://github.com/BioAlgs/pTDR).

Keywords: asymptotics, sufficient dimension reduction, tensor dimension reduction, multi-tissue transcriptome

1 Introduction

A tensor is a multidimensional array that can be expressed as sum of the outer product of some vectors (Merris (1997)), which is widely used in medical imaging (Li et al. (2010); Zhou et al. (2013); Zhou and Li (2014)) and bioinformatics (Zhong et al. (2005); Kessler et al. (2014); Hore et al. (2016)). Analyzing tensor data is beginning to emerge as a new way to uncover the new dependence relationships among different dimensions, which provide more insight into complex biological processes (Hore et al., 2016; Yang et al., 2015; Gamazon et al., 2019) and complex diseases (Erola et al., 2020; Talukdar et al., 2016; Kaminsky et al., 2012). This paper aims at developing advanced analysis methods for tensor data. We introduce two concrete examples which motivate our study.

Example 1: TwinsUK RNA-seq dataset. Aging is one of the most complex biological processes related to transcriptomic changes in tissues across the body and is one of the known risk factors for many age-related diseases in humans (Szilard, 1959; Moody and Sasser, 2020). Some genetic syndromes or chromosomal abnormalities can cause people to appear younger or older than their chronological age (Walker et al., 2009). TwinsUK RNA-seq dataset (`http://www.twinsuk.ac.uk/`) consists of 884 age-related genes and 4 tissues in blood, adipose, lymphoblastoid cell, and skin measured on 262 related individuals, which makes it an ideal cohort to study the process of aging. Denote Y as the biological age, and X as the gene expression profile, which can be represented as an 884×4-dimensional input. A primary focus is to infer biological age using multiple tissue gene expression data as biomarkers.

Example 2: *The Genotype-Tissue Expression (GTEx) project.* This project is an ongoing initiative aiming to create a comprehensive public database for investigating gene expression specific to different tissues (Lonsdale et al., 2013). To achieve this, the project collected samples from 54 healthy tissue sites obtained from nearly 1000 individuals, covering over 250 human traits. In our specific case, we examined the correlation between body mass index (BMI), denoted as Y, and multi-tissue gene expression represented by a two-dimensional tensor $\mathbf{X}$, with dimensions labeled as "Genes" and "Tissues".

Gene expression data illustrates a molecular portrait of biological processes. Recently, it has become feasible to generate large-scale multiple-tissue gene expression data from hundreds to thousands of individuals. The multiple tissue gene expression datasets can infer the interaction among tissues and provide an ideal source to identify complex cellular mechanisms underlying human traits and diseases (Aguet and Muñoz Aguirre, 2017). However, the large number of covariates causes the problem referred to as the "curse of dimensionality". For example, in the TwinUK dataset, we have 884 age-related genes and 4 tissues which generates 884 × 4 covariates much larger than the sample size 262.

A wide range of methods have been proposed in the literature for dimension reduction to mitigate this issue. Among them, sufficient dimension regression (SDR), which assumes that the response Y only depends on a lower-dimensional projection of $X \in \mathcal{X}$, is a popular approach. Let $P_{\mathcal{S}}$ be the projection operator from $\mathcal{X}$ to a linear space $\mathcal{S}$ in the standard inner product, where $\mathcal{S} \subset \mathcal{X}$. If

$$Y \perp\!\!\!\perp X \mid P_{\mathcal{S}}X, \tag{1}$$

where $\perp\!\!\!\perp$ means statistical independence, then it is said that $P_{\mathcal{S}}X$ is sufficient for the dependence of Y on X. In other words, the projection $P_{\mathcal{S}}X$ captures all the information contained in X regarding Y. Model (1) is referred to as the sufficient dimension reduction (SDR) regression model, and $\mathcal{S}$ is referred to as a dimension reduction subspace. Many methods have been proposed to estimate sufficient dimension reduction subspace; see, for example, Li (1991); Chen and Li (1998); Cook (1998); Li and Wang (2007); Nilsson et al. (2007), Cook and Weisberg (1994); Cook (1996, 1998).

Although the SDR model is a rich and flexible framework, it cannot be directly applied to tensor data. Existing works on SDR largely ignore the tensor structure by simply vectorizing each tensor observation into a vector and offering solutions using vector-based statistical methods. The disadvantages of this approach are as follows. First, the vectorization of tensor data destroys the original design information and leads to difficulties in interpretation. Second, vectorization significantly aggravates the curse of dimensionality. For example, the regression model between a scalar response Y and a matrix-valued predictor $\mathbf{X} \in \mathbb{R}^{p\times q}$ may assume $Y = \alpha + \beta_1^\top \mathbf{X}\beta_2 + \varepsilon$, which has only $p + q + 1$ parameters as one of β_1, β_2 must be taken to have a fixed scale for identifiability. If ignoring the tensor structure, assume that $Y = \alpha + \gamma^\top \mathrm{vec}(\mathbf{X}) + \varepsilon$, which has $pq + 1$ parameters, where $\mathrm{vec}(\mathbf{X})$ is the vectorized $\mathbf{X}$. New statistical methods and theories directly utilizing the intrinsic tensor structure are highly desirable.

A pioneer work along this line of thinking is the dimension folding (DF) method proposed in Li et al. (2010). Using matrix predictors as an example, we summarize the main idea of the DF method as finding two subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ such that their tensor product can include $\mathcal{S}$, where $\mathcal{S}$ satisfies model (1) with X being the vectorized tensor. That is, we want to find $\mathcal{S}_1$ and $\mathcal{S}_2$ such that $\mathcal{S} \subseteq \mathcal{S}_1 \otimes \mathcal{S}_2$ where $\mathcal{S}_1 \otimes \mathcal{S}_2$ is called the dimension folding space. Although this method can successfully impose a tensor structure on $\mathcal{S}$, it has some theoretical and empirical difficulties. *First*, using $\mathcal{S}_1$ and $\mathcal{S}_2$ as estimation targets will naturally create some redundant projection directions for regression analysis. For example, if we have a regression model $y = \cos[(\beta_1 \otimes \beta_2)^\top \mathrm{vec}(\mathbf{X})] + \sin[(\beta_3 \otimes \beta_4)^\top \mathrm{vec}(\mathbf{X})] + \varepsilon$, where $\mathbf{X}$ is a $p \times q$ matrix, the ideal results that DF can generate are $\mathcal{S}_1 = \mathrm{span}\{\beta_1, \beta_3\}$ and $\mathcal{S}_2 = \mathrm{span}\{\beta_2, \beta_4\}$. Consequently, $\mathcal{S}_1 \otimes \mathcal{S}_2$ can generate four projection directions $\beta_1 \otimes \beta_2$, $\beta_3 \otimes \beta_4$, $\beta_1 \otimes \beta_4$ and $\beta_3 \otimes \beta_2$, where the last two are actually regression irrelevant. Including the extra two directions increases the risk of overfitting, especially when the sample size is small. Way to choose three directions in this example which might need clarification in real applications with restrictions in choosing the number of directions. *Second*, empirically, estimating $\mathcal{S}_1$ and $\mathcal{S}_2$ requires a good estimate of $\mathcal{S}$, which involves estimating the inverse of the variance-covariance of vectorized tensor predictors. The dimension of the variance-covariance matrix of the tensor predictor increases quadratically as the number of coordinates increases. For a high-order tensor, estimating the inverse of a variance-covariance matrix will be extremely computationally challenging for the DF method. Recent work in Ding and Cook (2015) targets recovering the same dimension folding space with a matrix-formed linear condition, which makes the computational more efficient on high-order tensors.

To bypass the limitations mentioned above, we propose a novel *parsimonious* tensor dimension reduction regression (pTDR) model leveraging on a sequential iterative dimension reduction algorithm (SIDR). Instead of directly vectorizing the tensor predictors, pTDR preserves the data structure and dramatically reduces the dimensionality of the parameter space. The key idea of pTDR is that we sequentially search a collection of rank-one tensors via the proposed sequential rank test, such that the space spanned by the rank-one tensors is a subspace of the one found in DF that covers $\mathcal{S}$ in (1). We establish the general asymptotic theory and a sequential rank test for estimating the tensor subspace and selecting the correct number of directions. In practice, we propose a sequential iterative dimension reduction (SIDR) algorithm to prevent calculating the inverse of the large covariance matrix, making it more appealing to analyze data with high-order tensor observations.

We further applied the proposed pTDR approach to the TwinsUK RNA-seq dataset and GTEx Project. We also studied the association between the aging process/human traits and gene expression based on multiple-tissue transcriptomics data. As illustrated in Figure 1, we have Y represents

$$Y \perp\!\!\!\perp X \mid \langle X, \beta_1^{\circ(1\to m)} \rangle, \ldots, \langle X, \beta_K^{\circ(1\to m)} \rangle, \tag{2}$$

where $\beta_k^{\circ(1\to m)} = \beta_k^{(1)} \circ \cdots \circ \beta_k^{(m)}$ for $k = 1, \ldots, K$ is the outer product of m vectors. More generally, let U and V be two tensors with dimensions $m_1 \times m_2 \times \cdots \times m_p$ and $n_1 \times n_2 \times \cdots \times n_q$, respectively. The outer product of U and V, denoted by W, is a tensor of order $p + q$ with dimensions $m_1 \times \cdots \times m_p \times n_1 \times \cdots \times n_q$. The elements of W are given by:

$$W_{i_1,\ldots,i_p,j_1,\ldots,j_q} = U_{i_1,\ldots,i_p} \cdot V_{j_1,\ldots,j_q}.$$

For example, $m = 2$ denotes dimension along "Genes" and "Tissues" in Figure 1. In more general cases, m could be larger than 2 by increasing the dimensions by adding multiple conditions or groups. Compared to the single tissue study or vector-based methods, our proposed model can preserve the data structure, reduce the parameter space, and achieve lower prediction error, which supports the biological assumption that age can be predicted by the gene expression of multiple tissues with higher accuracy. Also, our results support the biological assumption that complex biological processes such as aging and obesity are related to the interaction of multiple tissues, which include that cross-tissue synchronization of gene expression changes (Yang et al., 2015) and interactions among genes across multiple tissues (Grundberg et al., 2012; Glastonbury et al., 2016).

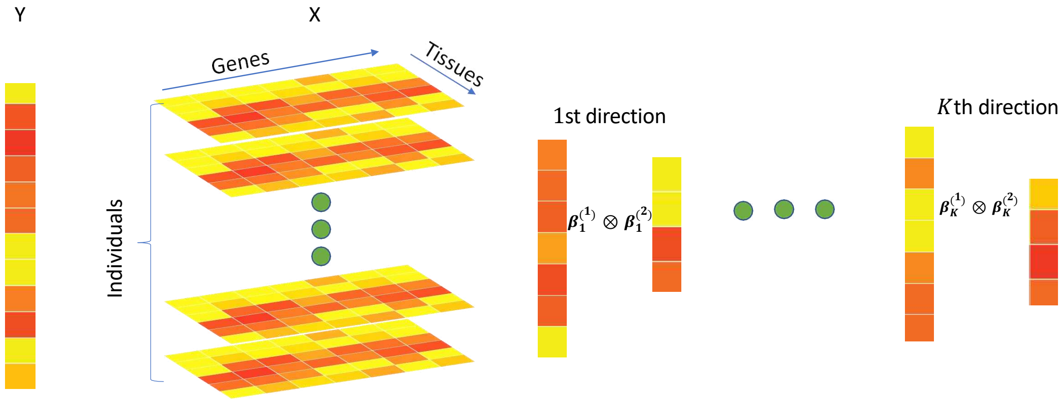

Y
X
Genes
Tissues
Individuals
1st direction
$\beta_1^{(1)} \otimes \beta_1^{(2)}$
Kth direction
$\beta_K^{(1)} \otimes \beta_K^{(2)}$

Figure 1: Graphical representation of the method. The response Y is the age and predictor variables X is a 2-d tensor. Through pTDR, we obtained K rank one tensors $\beta_1^{(1)} \otimes \beta_1^{(2)}, \ldots, \beta_K^{(1)} \otimes \beta_K^{(2)}$.

2 Parsimonious Tensor Dimension Reduction

2.1 Model setup

This section introduces the pTDR model and defines tensor sufficient dimension reduction subspace. The concepts and notations of tensor we used in this section mostly follow Merris (1997), Kolda (2001), and Kolda and Bader (2009).

Let $\mathbf{X} \in \mathcal{X} \equiv \mathbb{R}^{p_1 \times \cdots \times p_m}$ be a random tensor and $\mathrm{vec}(\mathbf{X})$ be the vectorized $\mathbf{X}$. For example, when $m = 2$, X is a 2-d tensor representing the gene expressions across p_1 genes and p_2 tissues in our motivating examples. Here we allow $m > 2$ for more general cases with extra dimensions along different conditions or groups. Let $\beta^{\circ(1\to m)} = \beta^{(1)} \circ \cdots \circ \beta^{(m)}$ represent the outer product of m vectors and $\beta^{(i)} \in \mathbb{R}^{p_i}$ is called the ith component of $\beta^{\circ(1\to m)}$. We have our pTDR model as

$$Y = g(\langle \mathbf{X}, \beta_1^{\circ(1\to m)} \rangle, \ldots, \langle \mathbf{X}, \beta_K^{\circ(1\to m)} \rangle, \varepsilon), \tag{3}$$

with g as an unknown function. Notice that (2) and (3) are equivalent models. If (3) holds, Y depends on X only through K indies, $\langle X, \beta_1^{\circ(1\to m)} \rangle, \ldots, \langle X, \beta_K^{\circ(1\to m)} \rangle$ which indicates (2). Conversely, if (2) holds, there exists g and ε such that (3) holds. When $\mathbf{X}$ is a vector, model (3) is precisely the generalized index model Xia (2008). The ith component of $\beta_k^{\circ(1\to m)}$, which is $\beta_k^{(i)}$, reflects the projection of $\mathbf{X}$ along the ith coordinate on the k-th direction. The pTDR model naturally incorporates the predictor's tensor structure and alleviates the curse of dimensionality. For example, assuming $K = 1$ and $\beta^{\circ(1\to m)}$ in the model (3) has only $\sum_{i=1}^{m} p_i$ parameters as one of $\beta_1^{(1)}, \ldots, \beta_1^{(m)}$ must be taken to have a fixed scale for identifiability, while there are $\prod_{i=1}^{m} p_i$ parameters if we vectorize $\mathbf{X}$.

Similar as the vector based SDR model, $\beta_1^{\circ(1\to m)}, \ldots, \beta_K^{\circ(1\to m)}$ in model (3) are not identifiable. For example, if $\beta_1^{(1)} \circ \beta_1^{(2)}$ and $\beta_1^{(1)} \circ \beta_2^{(2)}$ satisfy model (3), with some reparametrization $\beta_1^{(1)} \circ (\beta_1^{(2)} + \beta_2^{(2)})$ and $\beta_1^{(1)} \circ (\beta_1^{(2)} - \beta_2^{(2)})$ satisfy model (3) too. We define the space spanned by $\{\beta_1^{\circ(1\to m)}, \ldots, \beta_K^{\circ(1\to m)}\}$ as tensor dimension reduction space (TDS) as our target. TDS may not be unique. To bypass this ambiguity, we define the intersection of all TDS as the central tensor dimension reduction subspace (CTDS) if it is a TDS itself. The CTDS is unique and identifiable. In this paper, we consider the cases where CTDS exists. To ease the description, we use $\mathcal{S}_{Y|\mathbf{X}}$ to denote the CTDS and use $\mathcal{S}_{Y|\mathbf{X}}^{(i)}$ to denote the space spanned by $\{\beta_k^{(i)}\}$s for $k = 1, \ldots, K$.

2.2 Estimation of pTDR model

Considering the following motivating example, in which $Y = g(\langle \mathbf{X}, \beta_1^{\circ(1\to m)} \rangle + \varepsilon)$ and g is an invertible function, $\beta_1^{\circ(1\to m)}$ can be obtained by maximizing the squared correlation between $g^{-1}(Y)$ and $\langle \mathbf{X}, \eta \rangle$ with respect to η, where $\eta \in \mathbb{R}^{p_1 \times \cdots \times p_m}$. As a result, $\beta_1^{\circ(1\to m)}$ can be thought of as the most suitable rank-one tensor that, when the predictor is projected onto it, correlates with the best transformation of the response. Hence, from the perspective of projection pursuit, we don't require model (3) and CTDS for interpreting $(\beta_1^{\circ(1\to m)}, \ldots, \beta_K^{\circ(1\to m)})$ if they are calculated by maximizing the squared correlation, since they naturally represent the most effective directions to illustrate the relationship between the transformed response and the tensor predictors.

Recall that $\beta_1^{\circ(1\to m)} = \beta_1^{(1)} \circ \cdots \circ \beta_1^{(m)}$, where $\beta_1^{(i)} \in \mathbb{R}^{p_i}$. Let $T(Y)$ represent a transformation function applied to the response variable Y. Finding $\beta_1^{\circ(1\to m)}$ is equivalent to finding m vectors that maximize

$$L_1(\beta^{(1)}, \ldots, \beta^{(m)}) = \max_T \operatorname{corr}^2(T(Y), \langle \mathbf{X}, \beta^{\circ(1\to m)} \rangle). \tag{4}$$

For any fixed $\beta^{\circ(1\to m)}$, it can be shown that $\operatorname{corr}^2(T(Y), \langle \mathbf{X}, \beta^{\circ(1\to m)} \rangle)$ is maximized at $T(Y) = \mathrm{E}(\langle \mathbf{X}, \beta^{\circ(1\to m)} \rangle \mid Y)$ in the population level. Thus, $L_1(\cdots)$ has an explicit form

$$L_1(\beta^{(1)}, \ldots, \beta^{(m)}) = \frac{\operatorname{var}[\mathrm{E}(\langle \mathbf{X}, \beta^{\circ(1\to m)} \rangle \mid Y)]}{\operatorname{var}(\langle \mathbf{X}, \beta^{\circ(1\to m)} \rangle)}. \tag{5}$$

In practice, it is also possible to estimate $T(\cdot)$ directly using nonparametric regression methods, such as the one proposed in Fung et al. (2002). However, we will mainly focus on the estimation of $\beta^{(1)}, \ldots, \beta^{(m)}$ in this article. As a remark, a special case of (5) was studied in Chen and Li (1998) and Zhong et al. (2012) by letting $m = 1$. However, the commonly used orthogonality constraint that is assumed in Chen and Li (1998) and Zhong et al. (2012) for $m = 1$ cannot be assumed for $m \geq 2$.

In order to find the rest of rank-one tensor $\beta_2^{\circ(1\to m)}, \ldots, \beta_K^{\circ(1\to m)}$, we assume that $\beta_1^{\circ(1\to m)}, \ldots, \beta_K^{\circ(1\to m)}$ are not linearly dependent. Since the tensor parameters are identified sequentially, if we have already found k rank-one tensor parameters, we only require that the $(k+1)$th tensor parameter should not fall in the space spanned by the previous k tensor parameters. We notice that the PARAFAC in Harshman and Lundy (1984) or Tucker decomposition of m-model tensor Tucker (1951, 1966) is not unique using different algorithms Smilde et al. (2004); Kolda and Bader (2009) when $m > 2$.

Therefore, we propose a novel algorithm to estimate the rank-one tensors instead. Suppose that we have already obtained $\beta_1^{\circ(1\to m)}, \dots, \beta_k^{\circ(1\to m)}$. Let $\mathcal{F}_k \triangleq \text{span}\{ \beta_1^{\circ(1\to m)}, \dots, \beta_k^{\circ(1\to m)}\}$ be the linear space spanned by these k rank-one tensors. To find $\beta_{k+1}^{\circ(1\to m)}$, we first remove all the information that is contained in $\mathcal{F}_k$ from $\mathbf{X}$, and then find the best rank-one tensor direction in the same fashion as we did for $\beta_1^{\circ(1\to m)}$. Let $\Sigma_X = \mathrm{E}[(\mathrm{vec}(\mathbf{X}) - \mathrm{E}[\mathrm{vec}(\mathbf{X})])(\mathrm{vec}(\mathbf{X}) - \mathrm{E}[\mathrm{vec}(\mathbf{X})])^T]$ be the covariance matrix of $\mathrm{vec}(\mathbf{X})$. More specifically, let $P_{(k)} = \Sigma_{\mathbf{X}}\Gamma_{(k)}(\Gamma_{(k)}^\top\Sigma_{\mathbf{X}}\Gamma_{(k)})^{-1}\Gamma_{(k)}^\top$ be the projection matrix from $\mathbb{R}^{p_1\cdots p_m}$ onto $\mathcal{F}_k$, where $\Gamma_{(k)} \in \mathbb{R}^{p_1\cdots p_m\times k}$ is defined as $(\mathrm{vec}(\beta_1^{\circ(1\to m)}), \dots, \mathrm{vec}(\beta_k^{\circ(1\to m)}))$. Let $\mathbf{X}_{(k)}$ be the tensor counterpart of $\mathrm{vec}(\mathbf{X}_{(k)})$, where

$$\mathrm{vec}(\mathbf{X}_{(k)}) = (I - P_{(k)})\mathrm{vec}(\mathbf{X}).$$

It can be shown that $\langle \mathbf{X}_{(k)}, \eta\rangle = \mathrm{vec}(X)^\top(I - P_{(k)}^\top)\mathrm{vec}(\eta) = 0$ for any $\eta \in \mathcal{F}_k$ since η is in the subspace spanned by the columns of $P_{(k)}$.

Thus, the $(k+1)$th rank-one tensor $\beta_{k+1}^{\circ(1\to m)}$ can be obtained by maximizing

$$L_{k+1}(\beta^{(1)}, \dots, \beta^{(m)}) = \frac{\mathrm{var}[\mathrm{E}(\langle \mathbf{X}_{(k)}, \beta^{\circ(1\to m)}\rangle \mid Y)]}{\mathrm{var}(\langle \mathbf{X}_{(k)}, \beta^{\circ(1\to m)}\rangle)}. \tag{6}$$

Let $\beta_{k+1}^{(1)}, \dots, \beta_{k+1}^{(m)}$ be the maximizer of $L_{k+1}(\cdot)$, λ_{k+1}^* be the maximum value of $L_{k+1}(\cdot)$, and $\beta_{k+1}^{\circ(1\to m)} = \beta_{k+1}^{(1)} \circ \cdots \circ \beta_{k+1}^{(m)}$. The above procedure continues in a sequential way until it finds an integer K such that λ^*_{K+1} equal to zero, i.e.,

$$K := \arg\min_k \{\lambda^*_{k+1} = 0\}, \tag{7}$$

which indicates that there is no rank-one tensor associated with the response variable Y.

Recall that $\mathcal{F}_k$, where $k = 1, \ldots, K$, is the space that is spanned by $\{\beta_1^{\circ(1\to m)}, \ldots, \beta_k^{\circ(1\to m)}\}$. Therefore,

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_K.$$

The above procedure is referred to as sequential iterative dimension reduction algorithm (SIDR) because this procedure estimates a collection of well-defined rank-one tensor parameters

$$\beta_1^{\circ(1\to m)}, \ldots, \beta_K^{\circ(1\to m)} \tag{8}$$

that maximize the squared correlation between the indexes and a transformed response sequentially. Under certain mild conditions, the objective functions (6) have unique maximizers. This is due to the fact that Rayleigh's quotient in (6) can be connected to the eigenvalues of the empirical estimate of $\mathrm{var}[\mathrm{E}(\langle \mathbf{x}_{(k)}, \boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(m)} \rangle \mid \mathbf{Y})]$. In this case, $\beta_j^{(i)}$s defined by pTDR are identifiable. Also, under certain regularity conditions (Condition 1-3 in section 3.1), (8) is one solution satisfying model (2) as shown in our Theorem 1.

2.3 Implementation of the SIDR algorithm

In this part, we introduce the implementation of the SIDR algorithm. The estimate of the pTDR model is obtained through maximizing (5) and (6) which involve conditional expectation. We estimate the conditional expectation using slicing strategy, which is a popular technique in dimension reduction literature such as Li (1991). In addition, we update $\beta^{(i)}$ sequentially from $i = 1$ to m in each iteration. Thus, we only need to calculate the inverse of a $p_i \times p_i$ matrix in each step. In the following algorithm, we use $\otimes$ to denote the Kronecker product. We have $\mathrm{vec}(\beta^{\circ(1\to m)}) = \beta^{(m)} \otimes \cdots \otimes \beta^{(1)}$. The SIDR algorithm is summarized below.

Algorithm 1: Sequential iterative dimension reduction algorithm (SIDR)

```
1(a)[Continous response] Let $(\mathbf{X}_j, Y_j)$ denote the $j$th observation. Divide the range of
    the responses $\{Y_j\}$ $(j = 1, \ldots, n)$ into several disjoint intervals $I_1, \cdots, I_H$ and let
    $n_h$ denote the number of observations falling in $I_h$ and set $k = 0$;
1(b)[Discrete Response] Let $(\mathbf{X}_j, Y_j)$ denote the $j$th observation. For the discrete
    responses $Y_j$ $(j = 1, \ldots, n)$, categorize them into groups $I_1, \cdots, I_H$ based on their
    values, and let $n_h$ signify the count of observations in group $I_h$ and initialize
    $k = 0$.
2. If $k = 0$, we set $\mathbf{X}_{j(0)} = \mathbf{X}_j$. Otherwise, we set $\text{vec}(\mathbf{X}_{j(k)}) = (I - P_{(k)})\text{vec}(\mathbf{X}_j)$
    where $P_{(k)} = \widehat{\Sigma}_{\mathbf{X}}\Gamma_{(k)}(\Gamma^{\top}_{(k)}\widehat{\Sigma}_{\mathbf{X}}\Gamma_{(k)})^{-1}\Gamma^{\top}_{(k)}$ and $\widehat{\Sigma}_{\mathbf{X}}$ is the sample covariance matrix of
    $\text{vec}(\mathbf{X})$. Randomly initialize $\widehat{\beta}^{(i)}_{k+1,0}$ and set $t = 1$.
3. We maximize the empirical version of (6) with respect to $\beta^{(i)}$ sequentially to get
    the estimate $\widehat{\beta}^{(i)}_{k+1,t}$ for $i = 1, \ldots, m$, i.e.,
        $\widehat{\beta}^{(i)}_{k+1,t} = \text{argmax}_{\beta^{(i)}}\, L_{k+1,i}(\beta^{\circ(1\to m)}) = \frac{{\beta^{(i)}}^{\top}\text{var}[\text{E}(Z^{(i)}_{(k)} \mid Y)]\beta^{(i)}}{{\beta^{(i)}}^{\top}\text{var}[Z^{(i)}_{(k)}]\beta^{(i)}}$,
    where $Z^{(i)}_{j(k)} = \text{vec}(\mathbf{X}_{j(k)})^{\top}(\beta^{(m)}_{k+1,t-1} \otimes \cdots \otimes I_{p_i} \otimes \cdots \otimes \beta^{(1)}_{k+1,t})$. Estimate
    $\text{E}(Z^{(i)}_{(k)} \mid Y \in I_h)$ by $\bar{Z}^{(i)}_{(k)h} = \frac{1}{n_h}\sum_{\{j:Y_j \in I_h\}} Z^{(i)}_{j(k)}$ and $\text{var}[\text{E}(Z^{(i)}_{(k)} \mid Y)]$ by
    $\text{var}[\bar{Z}^{(i)}_{(k)h}] = \sum_{h=1}^{H} \frac{n_h}{n}\text{vec}(\bar{Z}^{(i)}_{(k)h} - \bar{Z}^{(i)}_{(k)})\text{vec}(\bar{Z}^{(i)}_{(k)h} - \bar{Z}^{(i)}_{(k)})^{\top}$. Return the largest
    eigenvalue of $\text{var}[\bar{Z}^{(i)}_{(k)h}](\text{var}[Z^{(i)}_{(k)}])^{-1}$ as $\widehat{\beta}^{(i)}_{k+1,t}$. Then, we have
    $\widehat{\beta}^{\circ(1\to m)}_{k+1,t} = \widehat{\beta}^{(1)}_{k+1,t} \circ \cdots \circ \widehat{\beta}^{(m)}_{k+1,t}$.
4. While $\|\widehat{\beta}^{\circ(1\to m)}_{k+1,t} - \widehat{\beta}^{\circ(1\to m)}_{k+1,t-1}\| > \epsilon$ for a predefined threshold $\epsilon > 0$. Update $t \leftarrow t + 1$
    and perform Step 3. Otherwise, Set $\widehat{\beta}^{\circ(1\to m)}_{k+1} = \widehat{\beta}^{\circ(1\to m)}_{k+1,t}$.
5. Test the hypothesis $H_0 : \lambda^{*}_{k+1} = 0$. If reject the $H_0$, we set k=k+1 and return to
    step 2. If $H_0$ is not rejected, we output $\widehat{\beta}^{\circ(1\to m)}_{1}, \ldots, \widehat{\beta}^{\circ(1\to m)}_{k+1}$. The details of the
    hypothesis testing and test rule are given in Theorem 3.3.
```

Note that in step 3, $E(Z_{(k)}^{(i)}|Y)$ is a general notation for the conditional expectation of $Z_{(k)}^{(i)}$ given Y, which is a function of Y. On the other hand, $\mathrm{E}(Z_{(k)}^{(i)}|Y \in I_h)$ represents a specific value of this function when Y falls within the interval I_h which can be treated as a discretized version of $E(Z_{(k)}^{(i)}|Y)$ which has a sample estimate $\bar{Z}_{(k)h}^{(i)} = \frac{1}{n_h}\sum_{\{j:Y_j \in I_h\}} Z_{j(k)}^{(i)}$.

When $m = 1$, i.e., the $\mathbb{X}$ is a vector, Algorithm 1 is equivalent to apply SIR sequentially to all modes of the tensor. For our motivating example introduced in Section 1, the jth subject has the gene expressions data $X_j \in \mathbb{R}^{p_1 \times p_2}$. Y_j is the trait of the jth subject. Plug in to the algorithm, we output $\widehat{\beta}_1^{\circ(1\to 2)}, \ldots, \widehat{\beta}_K^{\circ(1\to 2)}$ as the K rank-one tensors to span the TDS, which extract the information related to Y from X. The pTDR algorithm enables us to reduce number of parameters from $p_1 \times p_2$ to $K \times (p_1 + p_2)$. Also, we will introduce a sequential hypothesis testing for the choice of K in Theorem 3.3 in Section 3.

3 Theoretical results

3.1 Main Conditions

It is important to point out that $\beta_j^{\circ(1\to m)}$s obtained in Section 2 are identifiable parameters as maximizers of (5) and (6), while $\beta_j^{\circ(1\to m)}$s in model (3) are not identifiable because only the space that is spanned by them is identifiable. Thus, establishing the connection between the maximizers of (5) and (6) and the parameter in model (3) is highly desirable. Under some mild conditions, we show in this section that the space spanned by the maximizers of (5) and (6), $\mathcal{F}_K$, contains CTDS, which is spanned by the rank-one tensors in model (3). We further show that $\mathcal{F}_K$ is contained in the dimension folding subspace. In order to establish the above conclusion, we first state some conditions.

Condition 1 (Linear Condition). *A random tensor* $\mathbf{X}$ *is said to satisfy a linear condition with respect to tensor parameters* $\{\beta_1^{\circ(1\to m)}, \ldots, \beta_K^{\circ(1\to m)}\}$ *if there exist constants* $r_0, r_1, \ldots, r_K$ *such that*

$$\begin{aligned}&E(\langle b, \mathbf{X}\rangle \mid \langle \beta_1^{\circ(1\to m)}, \mathbf{X}\rangle, \ldots, \langle \beta_K^{\circ(1\to m)}, \mathbf{X}\rangle)\\=&r_0 + r_1\langle \beta_1^{\circ(1\to m)}, \mathbf{X}\rangle + \cdots + r_K\langle \beta_K^{\circ(1\to m)}, \mathbf{X}\rangle\end{aligned}$$

for any tensor b*.*

Condition 2 (Decomposable Variance Condition). *A random tensor* $\mathbf{X}$ *is said to satisfy the decomposable variance condition if there exists* $\Sigma_1, \cdots, \Sigma_m$*, where* $\Sigma_k \in \mathbb{R}^{p_k \times p_k}$*, such that the variance-covariance matrix of* $vec(\mathbf{X})$*, denoted by* $\Sigma_{\mathbf{X}}$*, have the following expression,* $\Sigma_{\mathbf{X}} = \Sigma_m \otimes \cdots \otimes \Sigma_1$*.*

Condition 3 (Coverage Condition). *For any tensor* $\gamma \in \mathcal{S}_{Y|\mathbf{X}}$ *there always exists a tensor* $\eta \in span\{E(\mathbf{X} \mid Y)\}$ *such that* $\Sigma_{\mathbf{X}} vec(\gamma) = vec(\eta)$*.*

The linear condition is essentially a tensor version of the linear condition that is defined in Li (1991). This condition is a sufficient condition for the consistency of most of SDR procedures. The linear condition implies that $\mathbf{X}$ follows an elliptically contoured distribution, which includes multivariate normal distribution as a special case. As discussed in Hall and Li (1993), the linear condition is a weak condition in the sense that it holds approximately for any distribution when the dimension is high. The decomposable variance condition is a sufficient condition to ensure $\mathcal{F}_K \subseteq \mathcal{S}_{Y|\mathbf{X}}^{(1)} \circ \cdots \circ \mathcal{S}_{Y|\mathbf{X}}^{(m)}$. The outer product of two space A and B is defined by $A \circ B = \{\alpha \circ \beta \mid \alpha \in A \text{ and } \beta \in B\}$.

Requiring the decomposable variance condition is equivalent to enforcing some constraints on the elements of $\Sigma_{\mathbf{X}}$. As a matter of fact, the decomposable variance condition is also a natural condition that can be generally satisfied in many tensor applications. This condition is especially useful in the high-order tensor analysis, as they can naturally save the number of parameters in variance and covariance estimation, which are in general very difficult for higher-order tensor analysis due to the high dimensionality of the data. Similar assumptions are often imposed in other statistical analysis such as the compound symmetric assumption in longitudinal data analysis. The coverage condition is the tensor counterpart of the coverage condition that was first proposed by Cook (2004) and is commonly used in the SDR literature; see Cook and Ni (2005). The coverage condition rules out the possibilities that span$\{\mathrm{E}(\mathbf{X} \mid Y)\}$ is a proper subspace of $\mathcal{S}_{Y|\mathbf{X}}$. Consequently, some awkward situations that exist in most of the SDR approaches, such as $\mathrm{E}(\mathbf{X}|Y) = 0$ for $Y = (\langle \mathbf{X}, \gamma \rangle)^2 + \varepsilon$, can be bypassed.

Remark 3.1: Condtion 3.1 is satisfied when the distribution of vec($\mathbf{X}$) is elliptically symmetric. As suggested in Rocke and Woodruff (1996), in pracitce, it is helpful to remove outliers and clusters if the emprirical distribution of vec($\mathbf{X}$) is diviated from elliptically symmetric distribution. In our revised manuscript, we add remark 3.1 to disscuss this issue.

Remark 3.2: Condition 3.3 is used in the theoretical analysis of our proposed method. Essentially, it is assumed that the covariance matrix of the tensor data can be decomposed as the Kronecker product of several smaller matrices. This decomposition dramatically reduces the number of parameters to be estimated, which is a great advantage in high-dimensional settings. In practice, the Kronecker product assumption has proven to be a useful approximation in many applications, like signal processing, image analysis, and other fields where tensor data are common. However, the assumption of a Kronecker-structured covariance matrix is indeed a strong one. Many real-world data might not naturally adhere to this structure, which can lead to potential model misspecification and bias in the estimated dimension reduction subspace. To check the robustness of our proposed method when this assumption is invalid, in our simulation studies (Case II and III), we show that our proposed method still maitains the best performance to recover the CTDS compared to other methods. Nevertheless, as with any model, one should be aware of these assumptions when applying Tensor SIR and be mindful of potential model checking or validation techniques. For instance, one might consider using diagnostic plots or goodness-of-fit tests to assess the suitability of the Kronecker structure for the covariance matrix of the data at hand.

3.2 Main results

Theorem 3.1. *(Parsimonious Property) Assume that model (3) holds, Condition 3 holds, and* $\mathbf{X}$ *satisfies Condition 1-2. We have*

$$\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{F}_K \subseteq \mathcal{S}_{Y|\mathbf{X}}^{(1)} \circ \cdots \circ \mathcal{S}_{Y|\mathbf{X}}^{(m)}.$$

The proof of Theorem 3.1 can be found in the Appendix. Theorem 3.1 states that $\mathcal{F}_K$ can be a proper subspace of $\mathcal{S}^{(1)}_{Y|\mathbf{X}} \circ \cdots \circ \mathcal{S}^{(m)}_{Y|\mathbf{X}}$. For example, consider a special case of model (3) with $K = 2$ and $\gamma_1 = \beta^{(1)}_1 \circ \beta^{(2)}_1$ and $\gamma_2 = \beta^{(1)}_2 \circ \beta^{(2)}_2$, there is a chance that $\mathcal{F}_K$ is the space spanned by γ_1 and γ_2 and has two dimensions, while the dimension folding subspace $\mathcal{S}^{(1)}_{Y|\mathbf{X}} \circ \mathcal{S}^{(2)}_{Y|\mathbf{X}}$ is spanned by $\{\beta^{(1)}_1 \circ \beta^{(2)}_1, \beta^{(1)}_1 \circ \beta^{(2)}_2, \beta^{(1)}_2 \circ \beta^{(2)}_1, \beta^{(1)}_2 \circ \beta^{(2)}_2\}$ and has four dimensions. Nevertheless, the largest possible space of $\mathcal{F}_K$ is the dimension folding subspace $\mathcal{S}^{(1)}_{Y|\mathbf{X}} \circ \mathcal{S}^{(2)}_{Y|\mathbf{X}}$. Thus, pTDR model has significant improvement over the dimension folding model in terms of model complexity, and further improves the estimation efficiency of g in the downstream nonparametric model fitting step. Another remark on Theorem 3.1 is that the pTDR estimates are still meaningful if its conditions are not satisfied, since they are still important directions that have the maximum squared correlations with the transformed response. However, the linear condition and the coverage condition are essential conditions for dimension folding, as the dimension folding subspace will be meaningless without these conditions.

Let $\beta^{(1)}_{1*}, \ldots, \beta^{(m)}_{1*}$ be the true parameters that maximize $L_1(\cdot)$ in (5), and λ_{1*} as the maximized value $L_1(\beta^{(1)}_{1*} \circ \cdots \circ \beta^{(m)}_{1*})$. Let $\widehat{\beta}^{(1)}_1, \ldots, \widehat{\beta}^{(m)}_1$ be the estimators that maximize L_1 in (6) and denote $\widehat{\lambda}_1$ as the estimated value of λ_{1*} at $\widehat{\beta}^{(1)}_1 \circ \cdots \circ \widehat{\beta}^{(m)}_1$. Next, we will show that the maximizer $\widehat{\beta}^{(1)}_1, \ldots, \widehat{\beta}^{(m)}_1$ and maximum value $\widehat{\lambda}_1$ are consistent estimators and enjoy the asymptotic normality. It is worth noting that the asymptotic results are valid regardless of the linear condition in Condition 1. Also our asymptotic results holds when the number of slices is an arbitrarily fixed number, which is consitant with the result in Zhu and Ng (1995).

Theorem 3.2. *(Asymptotic Property) Under conditions (A1) – (A3) listed in Appendix, $\widehat{\beta}_1^{(1)}$, ..., $\widehat{\beta}_1^{(m)}$, $\widehat{\lambda}_1$ jointly follows a multivariate normal distribution.*

$$\sqrt{n}\left\{\begin{pmatrix}\widehat{\beta}_1^{(1)}\\ \vdots\\ \widehat{\beta}_1^{(m)}\\ \widehat{\lambda}_1\end{pmatrix}-\begin{pmatrix}\beta_{1*}^{(1)}\\ \vdots\\ \beta_{1*}^{(m)}\\ \lambda_{1*}\end{pmatrix}\right\}\sim N(0, BE(A)^{-1}\Sigma_\psi E(A)^{-1}B),$$

where $B = diag(I_{p_1}, \ldots, I_{p_m}, J)$ and $J = (1/(m-1), \ldots, 1/(m-1), 1)$ is a $1\times m$ row vector. The expression of A and Σ_ψ are given in appendix due to their complexity.

Following the similar argument, we can also obtain the asymptotic normality of the estimates of $(\widehat{\beta}_k^{(1)}, \ldots, \widehat{\beta}_k^{(m)})$ for $k \geq 2$. Notice that this asymptotic distribution is essentially a conditional distribution given $\{\widehat{\beta}_j^{(i)}, \widehat{\lambda}_j, i = 1, \ldots, m; j = 1, \ldots, k-1\}$. Technically, we are able to write out the joint density of those $\widehat{\beta}_k^{(i)}$'s and $\widehat{\lambda}_k$'s using conditional density. We do not pursue here due to its complexity in notations.

One important question is how many rank-one tensors we should keep. In the following, we derive a sequential testing procedure to determine K, the number of rank-one tensors defined by SIDR. Specifically, the hypothesis testing is $H_0 : \lambda_{1*} = 0$. Although Theorem 3.2 derives the asymptotic distribution of $\widehat{\lambda}_1$, it is not applicable to the testing, because $\Sigma_\psi = 0$ under H_0 and thus $\widehat{\lambda}_1 = o_p(n^{-1/2})$. Let $M_{(1)} \triangleq \text{var}[\text{E}(\text{vec}(\mathbf{X}) \mid Y)]$ and $\widehat{M_{(1)}}$ be its estimator using the slicing strategy stated in Algorithm 1, i.e.,

$$\widehat{M}_{(1)} = \sum_{h=1}^{H} p_h \text{vec}(\bar{\mathbf{X}}_h - \bar{\mathbf{X}})\text{vec}(\bar{\mathbf{X}}_h - \bar{\mathbf{X}})^T$$

where $\bar{\mathbf{X}}_h = \frac{1}{n_h}\sum_{j:Y_j\in I_h} \mathbf{X}_j$ and $p_h = n_h/n$. Notice that $\lambda_{1*} = 0$ is equivalent to $M = 0$. Hence we propose to use the test statistic

$$\widehat{S}^2_{(1)} = n\text{tr}(\widehat{M}_{(1)}),$$

where a small value supports H_0 and a large value indicates $M \neq 0$ or $\lambda_{1*} > 0$. This test statistic is inspired by a testing procedure for SIR in Li (1991), where the test statistic can be written as $n\text{tr}(\widehat{\Sigma}^{-1/2}\widehat{M}_{(1)}\widehat{\Sigma}^{-1/2})$. In this paper, we drop $\widehat{\Sigma}^{-1/2}$ because when $p > n$, $\widehat{\Sigma}^{-1/2}$ is no longer a consistent estimate of $\Sigma_{\mathbf{X}}^{-1/2}$ and the computational cost for calculating the inverse is high.

Theorem 3.3. *(Sequential Hypothesis Testing) The asymptotic distribution of $\widehat{S}^2_{(1)}$ is a weighted chi-squared distribution. More precisely, $\widehat{S}^2_{(1)}$ asymptotically has the same distribution as $\sum_{l=1}^{\infty} z_l\chi^2_l(1)$, where $\chi^2_l(1)$ are independently chi-squared random variables with one degree of freedom and z_l are eigenvalues of kernel function $\Phi(\mathbf{X}_1, y_1, \mathbf{X}_2, y_2)$ where $\Phi(\mathbf{X}_1, y_1, \mathbf{X}_2, y_2) = \sum_{h=1}^{H}(\frac{1}{p_h}I_{1h}I_{2h} - p_h)vec(\mathbf{X}_1)^\top vec(\mathbf{X}_2)$ and $I_{jh} = I_{\{y_j \,\in\, the\ h^{th}\ slice\}}$ based on the slicing strategy stated in Algorithm 1, and $(\mathbf{X}_1, y_1)$ and $(\mathbf{X}_2, y_2)$ are i.i.d. copies of $(\mathbf{X}, y)$.*

Similar weighted chi-squared tests have been proposed in the literatures on dimension reduction, for example Bura and Cook (2001) and Zeng (2008). Although Theorem 3.3 gives the limiting distribution of $\widehat{S}^2_{(1)}$, it is difficult to calculate all z_l's explicitly. We further propose an approximation for practical usage; see section 4.1 for details. For $k \geq 1$, we denote $M_{(k+1)} \overset{\triangle}{=} \text{var}[\text{E}(\text{vec}(\mathbf{X}_{(k)}) \mid Y)]$ and $\widehat{M}_{(k+1)}$ be its estimator using the slicing strategy stated in Algorithm 1, i.e.,

$$\widehat{M}_{(k+1)} = \sum_{h=1}^{H} p_h \text{vec}(\mathbf{X}_{(k)h} - \bar{\mathbf{X}}_{(k)})\text{vec}(\mathbf{X}_{(k)h} - \bar{\mathbf{X}}_{(k)})^T$$

where $\mathbf{X}_{(k)h} = \frac{1}{n_h}\sum_{j:Y_j \in I_h} \mathbf{X}_{j(k)}$ and $p_h = n_h/n$. Similarly, we can derive the test statistics $\widehat{S}_{(k+1)} = n\text{tr}(\widehat{M}_{(k+1)})$ for the hypothesis testing $H_0 : \lambda^*_{k+1} = 0$, for $k \geq 0$. By replacing $\mathbf{X}$ by $\mathbf{X}_{(k)}$, we could apply Theorem 3.3 to obtain the asymptotic distribution of $\widehat{S}_{(k+1)}$. Then we have our estimate of K as

$$\widehat{K} := \arg\min_{k}\{\widehat{S}_{(k+1)} > \Psi_{1-\alpha}(\sum_{l=1}^{\infty} z_l \chi^2_l(1))\} \tag{9}$$

where $\Psi_{1-\alpha}(\cdot))$ is the $1-\alpha$ quantile of $\sum_{l=1}^{\infty} z_l \chi^2_l(1))$, and α is the predefined significant level. With a little abuse of notation, we note that z_ls in (9) are obtained by replace $\mathbf{X}$ by $\mathbf{X}_{(k)}$ in Theorem 3.3.

4 Numerical Study

In this section, we evaluate the performance of the proposed pTDR approach via Monte Carlo studies. We use SIR to denote the results obtained by vectorizing tensors, use pTDR to represent the results obtained by our proposed pTDR method, use foldedSIR to indicate the results obtained by dimension folding proposed in Li et al. (2010), and use tensorSIR to denote the results obtained by tensor regression proposed in Ding and Cook (2015). Additionally, each method may be followed by a number to indicate the number of directions extracted using this method. For example, pTDR(2) means that we extract two directions using pTDR, foldedSIR(2, 1) indicates that we have a two-dimensional central left-folding subspace and one-dimensional central right-folding subspace. We consider the input tensor with independent, spatial correlated, and locally correlated covariance structures. In the Supplementary, we show additional synthetic experiments verifying our theoretical contribution to sequential testing.

Here is the revision:

We evaluate the performance of SIR, folded-SIR, tensorSIR, and pTDR using four different settings that combine two data generative models with two coefficient configurations. The data generative models and coefficient settings are as follows:

Setting 1:

$$Y = f_1(\mathbf{X}) + \sigma\varepsilon = \frac{\alpha_1^T \mathbf{X} \beta_1}{2 + (\alpha_2^T \mathbf{X} \beta_2 + 3)^2} + \sigma\varepsilon,$$

Setting 2:

$$y = f_2(\mathbf{X}) + \sigma\varepsilon = \frac{\alpha_1^T \mathbf{X} \beta_1}{2 + (\alpha_1^T \mathbf{X} \beta_2 + 3)^2} + \sigma\varepsilon,$$

Setting 3:

$$y = f_3(\mathbf{X}) + \sigma\varepsilon = \sin(\alpha_1^T \mathbf{X} \beta_1) + (\alpha_2^T \mathbf{X} \beta_2)^3 + \sigma\epsilon,$$

Setting 4:

$$y = f_4(\mathbf{X}) + \sigma\varepsilon = \sin(\alpha_1^T \mathbf{X} \beta_1) + (\alpha_1^T \mathbf{X} \beta_2)^3 + \sigma\epsilon$$

where the random error ε is independent of $\mathbf{X}$ and is distributed as $N(0,1)$, $\sigma = 0.5$, $\alpha_1 = (1,1,0,0,0)^T \in \mathbb{R}^5$, $\alpha_2 = (0,0,0,1,1)^T \in \mathbb{R}^5$, $\beta_1 = (1,1,-1,0,\ldots,0)^T \in \mathbb{R}^9$, and $\beta_2 = (1,-1,1,0,\ldots,0)^T \in \mathbb{R}^9$. Also, in our Supplimentary, we show more settings with higher dimensions. For SIR, the central subspace is spanned by $\beta_1 \otimes \alpha_1$ and $\beta_2 \otimes \alpha_2$. For folded-SIR, the central left-folding subspace is spanned by α_1 and α_2, and the central right-folding subspace is spanned by β_1 and β_2. Hence the central folding subspace has four dimensions. If we want to find a subspace of two dimensions, we can use foldedSIR(1, 2), foldedSIR(2, 1), tensorSIR(1,2) or tensorSIR(2,1). We randomly generate 100 samples each with size $n = 500$ for Settings 1-4 and apply six methods, pTDR(2), foldedSIR(1, 2), foldedSIR(2, 1), tensorSIR(1,2), tensorSIR(2,1), and SIR(2) to estimate the directions. The number of slices is $H = 10$ for all four methods. The results are represented as a projection matrix $\widehat{P}$. Let P_0 be the projection matrix. The performance of the estimation is measured by the correlation distance introduced by Hooper (1959) and Ye and Weiss (2003). Let P_0 be the orthogonal basis corresponding to the space spanned by $\{\beta_1 \otimes \alpha_1, \beta_2 \otimes \alpha_2\}$, $\widehat{P}$ be the orthogonal basis corresponding to the estimated space, and $\rho_i^2, i = 1, \ldots, d$ are eigenvalues of matrix $\widehat{P}^T P P^T \widehat{P}$. The correlation distance is defined as $1 - |\Pi_{i=1}^d \rho_i|$ where smaller values indicate better performance. Note that, in both settings 2 and 4, the two directions ($\alpha_1 \otimes \beta_1$ and $\alpha_1 \otimes \beta_2$) include α_1. We assume that the left space is one-dimensional and the right space is two-dimensional. Under these conditions, foldedSIR and tensorSIR can theoretically recover the CTDR directions consistently.

Case I: independent covariance structure: we simulate the component of $\mathbf{X} \in \mathbb{R}^{5\times 9}$ independently from $N(0,1)$. Figure 2 displays the boxplot for the five methods side-by-side. For all settings, we simulate the components of $X \in \mathbb{R}^{5\times 9}$ independently from a standard normal distribution, N(0, 1). Our findings show that pTDR consistently outperforms the other methods across all settings. In settings 2 and 4, where the assumptions of foldedSIR and tensorSIR are satisfied, their performance is only slightly lower than pTDR, but with larger variances. However, when the direction is misspecified, as in foldedSIR(2,1) and tensorSIR(2,1), their performance degrades significantly. In settings 1 and 3, where foldedSIR and tensorSIR cannot obtain parsimonious solutions due to their limitations, pTDR demonstrates the best performance. The performance of SIR is not good because in this case SIR is solving a problem with $5 \times 9 = 45$ dimensions, while pTDR is solving a problem with $5 + 9 = 14$ dimensions. It is understandable that the latter is expected to get more accurate results. The inferior performance of foldedSIR and tensorSIR is also expected because in the population level foldedSIR and tensorSIR needs a space of dimension four to include both directions.

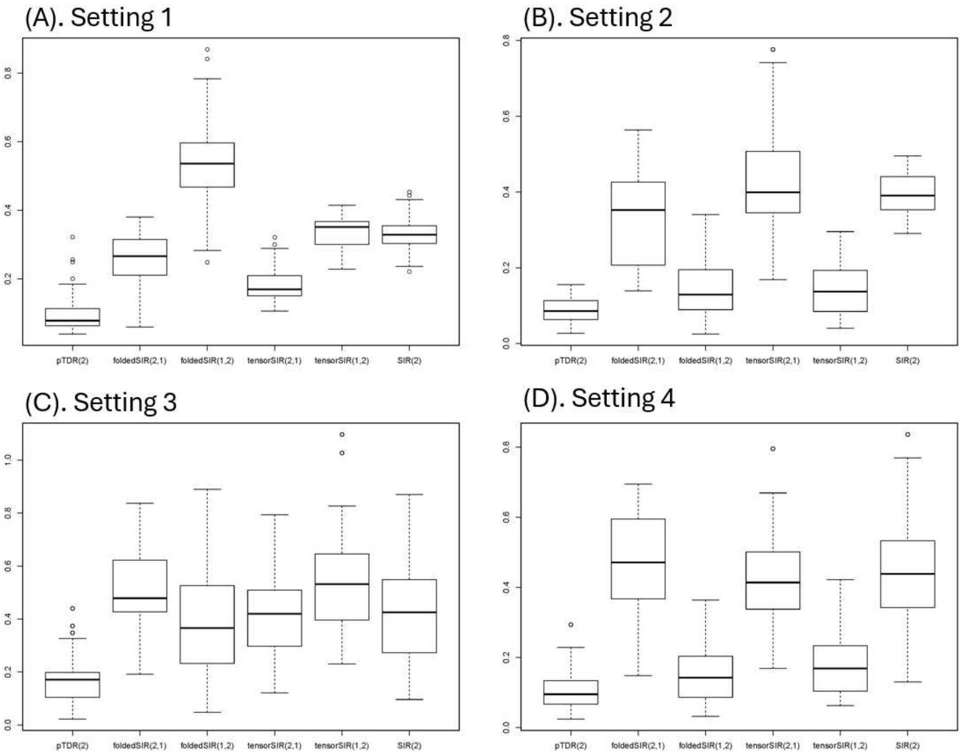


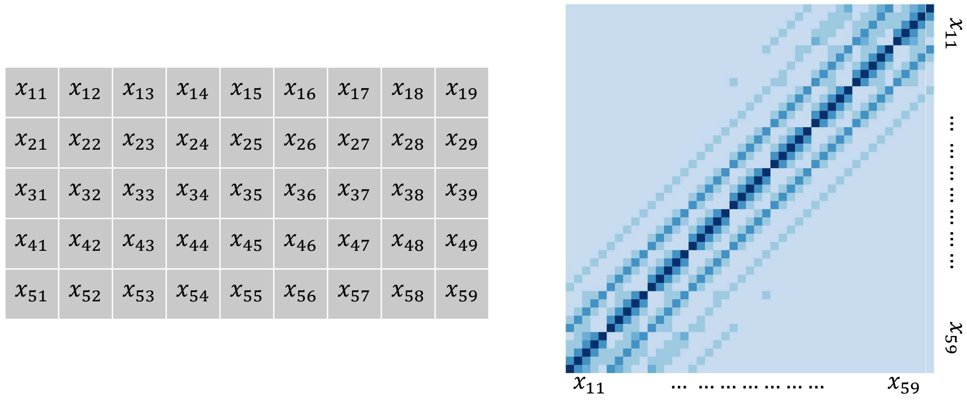
Figure 2: *Boxplot for the performance of pTDR(2), foldedSIR(2,1), foldedSIR(1,2), tensorSIR(2,1), tensorSIR(1,2), and SIR(2) in the Settings 1-4 where the input tensor with independent entries.*

Case II: spatial correlated covariance structure: In this setting, we consider the X as a 5×9 tensor with entries located on a two-dimensional grid shown in Figure 3(a). We set the correlation between x_{ij} and $x_{i'j'}$ as $\rho^{|i-j|+|i'-j'|}$ where $\rho = 0.5$. The correlation matrix of $vec(X)$ is shown in Figure 3(b). For fair comparison, we set the number of directions equal to two for all methods and calculate the correlation distance between the estimated space and the true space. As shown in Figure 4, the proposed pTDR(2) has most accurate estimate. The performance of SIR(2) is not as good as the others since the

n

te

ir



x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	x_{29}
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}	x_{39}
x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	x_{46}	x_{47}	x_{48}	x_{49}
x_{51}	x_{52}	x_{53}	x_{54}	x_{55}	x_{56}	x_{57}	x_{58}	x_{59}

(a). Spatial structure of input tensor

(b). Correlation structure of $vec(X)$

Figure 3: *The spatial correlation structure for* 5×9 *tensor* X *with entries located on two-dimensional grid.*

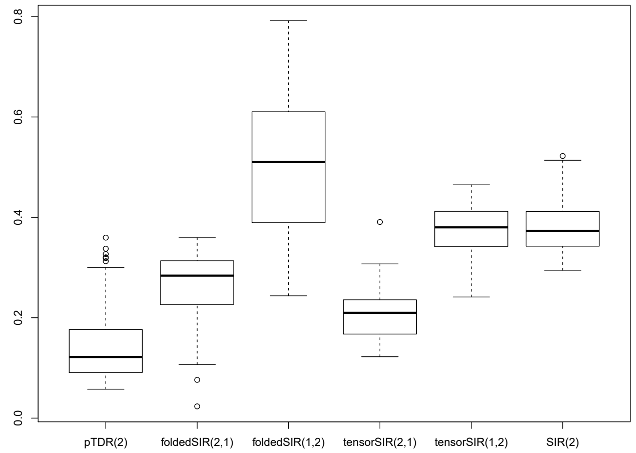
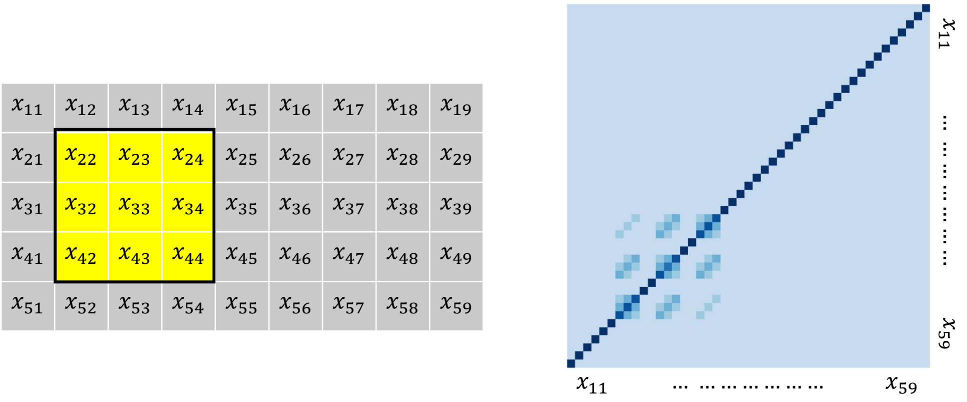


Figure 4: *Boxplot for the performance of pTDR(2), foldedSIR(2,1), foldedSIR(1,2), tensorSIR(2,1), tensorSIR(1,2), and SIR(2) in the setting where the input tensor has spatial correlation structure.*

Case III: locally correlated covariance structure: In this setting, we consider that the spatial correlation structure only exists in the highlighted region shown in Figure 5(a). If x_{ij} and $x_{i'j'}$ are both in the highlighted region, we set the correlation between x_{ij} and $x_{i'j'}$ as $\rho^{|i-j|+|i'-j'|}$. If either x_{ij} or $x_{i'j'}$ is not in the high lighted region, we randomly set their correlation from uniform distribution in $c(-0.1, 0.1)$. In this setting, the decomposable assumption is not valid. As shown in Figure 6, the proposed method still shows the smallest distance between the true and estimated space.



x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	x_{29}
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}	x_{39}
x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	x_{46}	x_{47}	x_{48}	x_{49}
x_{51}	x_{52}	x_{53}	x_{54}	x_{55}	x_{56}	x_{57}	x_{58}	x_{59}

(a). Spatial structure of input tensor

(b). Correlation structure of $vec(X)$

Figure 5: *The local correlation structure for* 5×9 *tensor* X *with entries located on two-dimensional grid.*

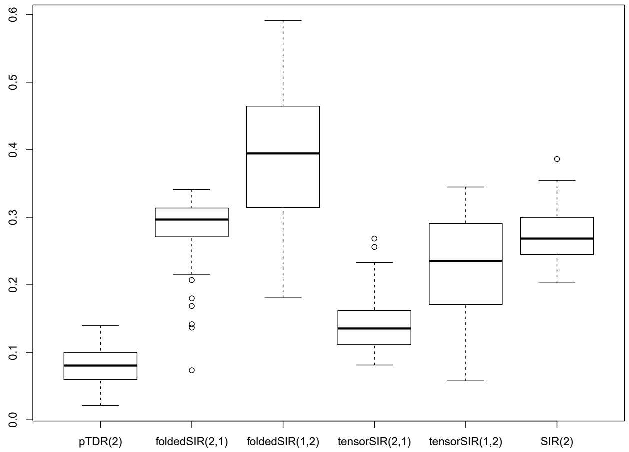


Figure 6: *Boxplots for the performance of pTDR(2), foldedSIR(2,1), foldedSIR(1,2), tensorSIR(2,1), tensorSIR(1,2), and SIR(2) in the setting where the input tensor has local correlation structure.*

4.1 Simulation on Sequential Test

This example is intended to check the performance of the proposed testing procedure. Consider the following model,

$$Y = \frac{c\alpha_1^T \mathbf{X}\beta_1}{2 + (c\alpha_2^T \mathbf{X}\beta_2 + 3)^2} + \sigma\varepsilon,$$

where $\sigma = 0.5$ and c is constant. When $c = 0$, Y is independent of $\mathbf{X}$, while when $c > 0$, Y depends on $\mathbf{X}$ via two directions of $\mathbf{X}$. As c increases, the signal becomes stronger.

We sample 500 samples each with size n from this model with a given value of c. Then we apply the proposed testing procedure to check if there is only one direction at significance level $\alpha = 0.05$. The number of slices is $H = 10$. Figure 7(a) shows the proportion of rejecting H_0 as c increases from 0 to 1. The two lines correspond to different sample size $n = 400$ and $n = 1000$, respectively. This plot illustrates the power of the test. As c increases, the power increases quickly, and a larger sample size leads to a higher power.

The p-values in the above simulations are calculated by approximating the sampling distribution by a single scaled chi-squared distribution. To verify the performance of this approximation, we also calculate the p-values from the weighted chi-squared distribution directly. Because the cumulative distribution function of a weighted chi-squared distribution is difficult to derive, we calculate the p-values using Monte Carlo. To evaluate $P(\sum_{k=1}^N z_k \chi_k(1) \geq a)$ for a scalar a, we randomly generate $10,000$ numbers independently from $\chi^2(1)$, denoted by $\chi^2_{i,j}$, $i = 1, \ldots, 10,000$, $j = 1, \ldots, N$. Calculate $a_i = \sum_{i=1}^N z_k \chi^2_{i,k}$ for $i = 1, \ldots, 10,000$. The p-values are the proportion of $a_i \geq a$.

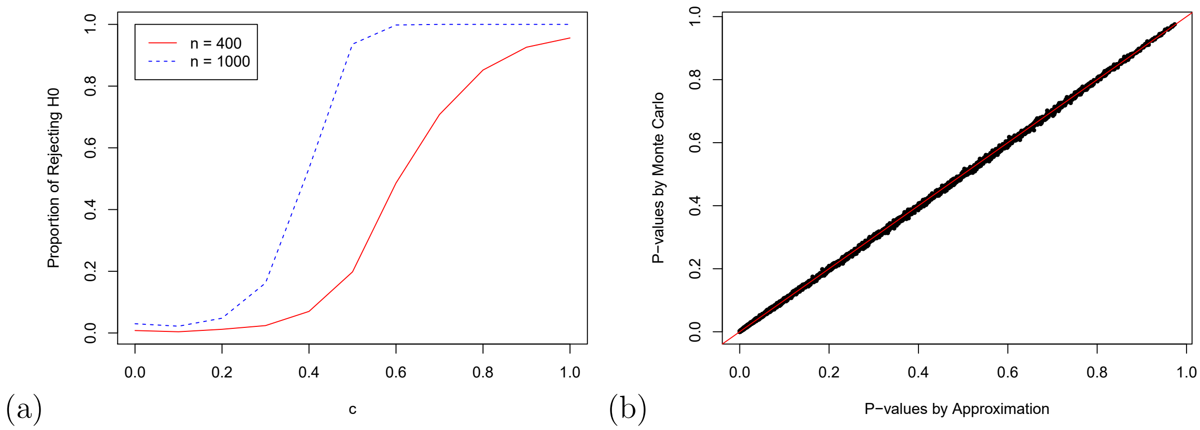


Figure 7: (a) The proportion of rejecting H_0 increases as c increases, where the solid line corresponds to $n = 400$ and the dash line corresponds to $n = 1000$. (c) The comparison of p-values by approximation and by Monte Carlo.

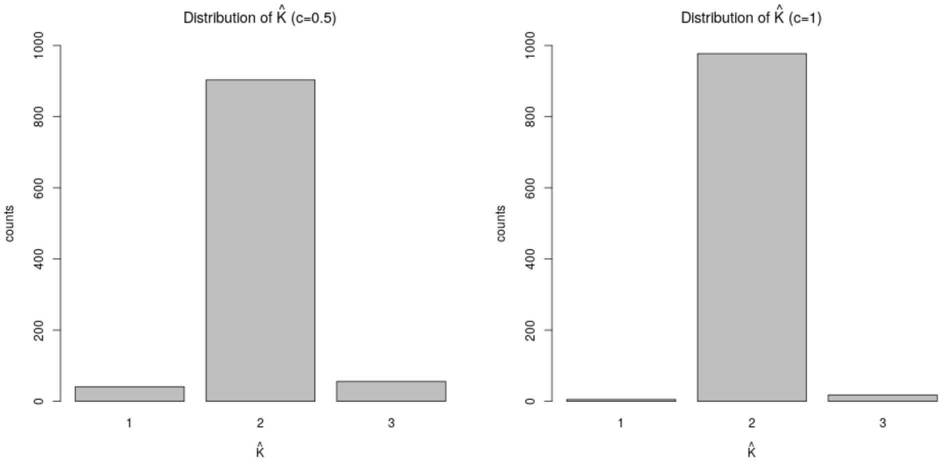


Figure 8: The empirical distribution of $\widehat{K}$ for $c = 0.5$ and $c = 1$.

Also, we show the empirical distribution of $\widehat{K}$. We set $c = 0.5$ and $c = 1$ to present the week and strong signal scenarios, the significance level is set as $\alpha = 0.05$, and sample size

$n = 1000$. We repeat the simulation for 1000 times. As shown in Figure 2, when the signal is weak $c = 0.5$, there are around 89% of $\widehat{K}$ equal to 2. When the signal is strong $c = 1$, there are around 97% equal to 2.

5 Real Data Analysis

In this part, we apply the pTDR model to the TwinsUK RNA-seq dataset and GTEx projected described in Section 1, to predict the age process and obesity using multiple tissues.

5.1 Application to the TwinsUK RNA-seq dataset

We have 262 individuals from the TwinsUK cohort (`http://www.twinsuk.ac.uk/`) with gene expression measured via RNA-seq analysis in blood, adipose, lymphoblastoid cell, and skin. We modeled the age as our response Y. The predictor variables formed 2-d tensor with one dimension as genes and the other dimension as tissues (see Figure 1 for the data structure illustration). We selected 884 age-related genes altered during aging according to GenAge database (`http://genomics.senescence.info/genes/`).

After applying pTDR on this data set, we obtained K rank one tensor as $\beta_1^{(1)} \otimes \beta_1^{(2)}, \ldots, \beta_K^{(1)} \otimes \beta_K^{(2)}$. By using our proposed sequential test, the estimated optimal number of direction $\widehat{K} = 7$. Then we mapped multiple-tissue gene expressions on the first and the second pTDR directions. As shown in the left panel of Figure 9, we observe a trend that the age is decreasing along the first pTDR direction.

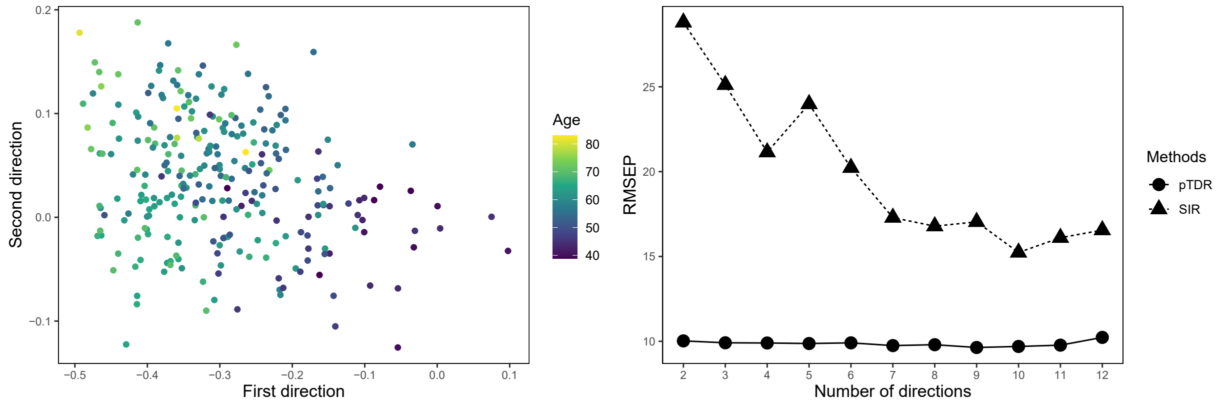


Figure 9: The left panel shows the projections of 262 individuals on the first two pTDR directions. The right panel is the average RMSEP of testing sets versus the number of directions.

To evaluate the sensitivity and specificity of age prediction, we randomly separated the data into 5 folds and choose one as the testing set and the others as the training set. We applied pTDR on the training data with K rank one tensors selected as the pTDR directions. We range K from 2 to 12. For comparison, we also applied SIR on the vectorized predictors and selected K vectors as SIR directions. Through projecting the predictors on the K selected directions, we obtained K new variables $Z_1, \ldots, Z_K$ and built a nonparametric additive model as $y = f_1(Z_1) + \cdots + f_K(Z_K)$ where $f_1, \ldots, f_K$ are unknown function estimated using the *gss* R package. Based on the well-trained model via the training set, we did the prediction on the testing set. To evaluate the prediction accuracy, we defined the root mean squared error of prediction (RMSEP) as RMSEP $= \sqrt{\sum_{i=1}^{n_t} (\widehat{y_i} - y_i)^2 / n_t}$ where $\widehat{y_i}, i = 1, \ldots, n_t$ are the predicted values on the testing set and n_t is the size of the testing set. As shown in the right panel of Figure 9, pTDR has the

RMSEP nearly around 10 and SIR has the RMSEP all above 15 with high fluctuations.

We show that $\beta_1^{(2)} = [-0.849, -0.298, -0.217, -0.379]^T$ corresponding to coefficients for the adipose, blood, lymphoblastoid cell, and skin . We rewrite the coefficients for $vec(X)$ as

$$\beta_1^{(1)} \otimes \beta_1^{(2)} = [-0.849\beta_1^{(1)T}, -0.298\beta_1^{(1)T}, -0.217\beta_1^{(1)T}, -0.379\beta_1^{(1)T}]^T$$

where the $-0.849\beta_1^{(1)T}$ are coefficients for genes in adipose tissues, $-0.298\beta_1^{(1)T}$ are coefficients for genes in blood tissue, $-0.217\beta_1^{(1)T}$ are coefficient for genes in the lymphoblastoid cell, and $-0.379\beta_1^{(1)T}$ are coefficients for genes in skin tissues. These similar coefficients between multiple tissues show that one tissue appears to be young, the other tissue tends to be young too. These results supported the biological assumption that aging is a complex biological process related to the interaction of multiple tissues, which is also observed in (Yang et al., 2015) that there is cross-tissue synchronization of age-related gene expression changes in multiple tissues.

5.2 GTEx project dataset

We next analyze the GTEx project dataset using pTDR and explore the relationship between multiple tissues gene expression and obesity. Obesity has prevailed in the United States in recent decades. There is a lot of medical research studying deeply into the relationship between genes, tissues, and obesity, from which we found that adipose, skeletal muscle and thyroid are closely related to it (Valenzuela et al., 2020; Sanyal and Raychaudhuri, 2016). We conventionally take Body Mass Index (BMI) as the measurement of obesity.

In this study, we include 159 individuals available from the GTEx project (`https://www.gtexportal.org/home/datasets`) with gene expressions measured in three tissues: adipose, skeletal muscle, and thyroid. The predictor variables X formed a 2-d tensor with one dimension as genes and the other dimension as tissues. We select 50 obesity-related genes in (Herrera et al., 2011). Thus, the predictor for each individual is a 50×3 matrix (2-d tensor). Then we applied pTDR model on the pre-processed data set and obtained the first K rank-one tensors as $\beta_1^{(1)} \otimes \beta_1^{(1)}, \ldots, \beta_K^{(1)} \otimes \beta_K^{(1)}$. By using our proposed sequential test, the estimated optimal number of direction $\widehat{K} = 6$.

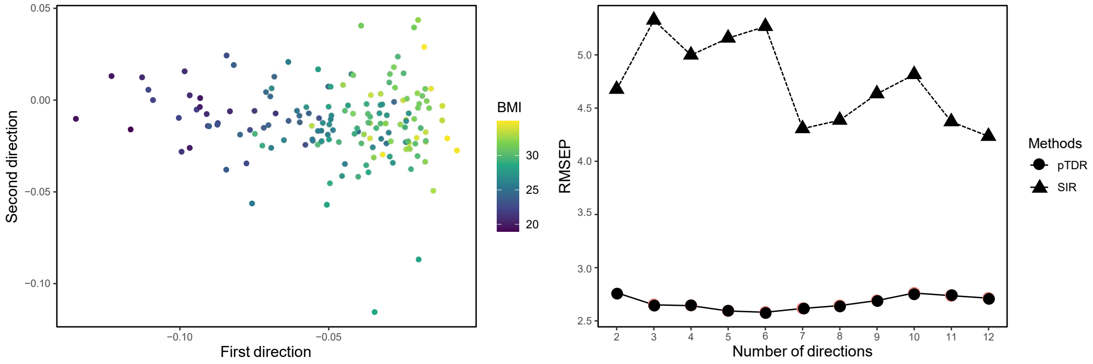


Figure 10: The left panel shows the projections of 159 individuals on the first two pTDR directions. The right panel is the average RMSEP of testing sets versus the number of directions.

We extracted the first two directions from the results of our proposed pTDR algorithm, drawing every individual on the plot. It is worth noting that the BMIs have a strong increasing trend along the first pTDR direction. Similarly, we conducted SIR on the vectorized predictors and also selected K vectors as SIR directions with K ranging from 2 to 12. Following the same procedure in Section 5.1, we obtained the RMSEP for both models. pTDR has a smaller RMSEP of nearly 2.58 and SIR has a much higher RMSEP of 4.23 which supported that including interactions among genes in multiple tissues could improve the prediction performance of BMIs. We show that $\beta_1^{(2)} = [-0.611, 0.078, -0.787]^T$ corresponding to coefficients adipose, skeletal muscle, and thyroid. The coefficients of adipose and thyroid are both negative and have a much larger magnitude than the coefficient of skeletal muscle which indicates adipose and thyroid might play more important roles in obesity (Song et al., 2019; Sam and Mazzone, 2014). Also, the similarity of the coefficient in adipose and thyroid shows a positively correlated effect of gene expression to obesity. These results are also consistent with recent studies where the interaction effects across multiple tissues plays important role in obesity (Grundberg et al., 2012; Glastonbury et al., 2016).

6 Discussions

As science and technology advance swiftly, tensor observations become increasingly prevalent in our everyday lives, thus creating a significant demand for efficient tools to analyze tensor data. In this study, we introduced a pTDR model in the context of the SDR framework. Like all other sufficient dimension reduction methods, our pTDR model does not presume a specific link between the response variable and the explanatory variables, and it provides a parsimonious solution. Furthermore, the tensor's dimensionality is not limited to 2. For instance, sophisticated neural imaging data are typically 4-dimensional, incorporating spatial and temporal aspects. Therefore, this model is versatile and applicable to a broad range of scientific pursuits, such as the one discussed in this paper.

Also, in our real data analysis using gene expression data from the TwinsUK and GTEx projects, we acknowledge the limitation of insufficient replicates for each individual, which may have affected our ability to fully capture the individule-level variation. To address this in future studies, we recommend prioritizing the collection of more replicates per individual and exploring methods such as including a random effect term in the model to account for individual-level variability. While this limitation does not invalidate our overall findings, addressing individual-level variability in future analyses will provide a more comprehensive understanding of gene expression patterns and their role in biological systems.

7 Appendix: proofs of main theorems

A.1 Proof of Theorem 3.1

The following lemma is the tensor counterpart of Theorem 3.1 in Li (1991), and hence its proof is omitted. It is needed in the proof of Theorem 1.

Lemma A.1. *Assume that model (3) holds and* $\mathbf{X}$ *satisfies the linear condition with respect to tensors* $\beta_1^{\circ(1\to m)}, \ldots, \beta_K^{\circ(1\to m)}$. *If* $E(\mathbf{X}) = 0$*, the inverse regression curve* $E(\mathbf{X} \mid Y)$ *satisfies* $\langle u, E(\mathbf{X} \mid Y)\rangle = 0$ *for any tensor* u *that is orthogonal to* $\beta_j^{\circ(1\to m)}$ *with respect to* $\Sigma_{\mathbf{X}}$*,* $j = 1, \ldots, K$*, that is,* $\langle u, \beta_j^{\circ(1\to m)}\rangle_{\Sigma_{\mathbf{X}}} = 0$.

Next we prove Theorem 3.1.

Proof. Note that $\mathcal{F}_K$ is spanned by $\beta_1^{\circ(1\to m)}$, ..., $\beta_K^{\circ(1\to m)}$, which are the K rank-one tensor parameters obtained by SIRD algorithm. Let us first consider the first part, the relationship between $\mathcal{S}_{Y|\mathbf{X}}$ and $\mathcal{F}_K$. It is enough to show that for any tensor (not necessary rank-one) u, if u is orthogonal to $\mathcal{F}_K$, then $\mathrm{vec}(u)$ is also orthogonal to $\Sigma_{\mathbf{X}}^{-1}\mathrm{E}(\mathrm{vec}(\mathbf{X}) \mid Y)$.

The fact that u is orthogonal to $\mathcal{F}_K$ implies that $\mathrm{vec}(u)^\top\Gamma_{(K)} = 0$, which further implies

$$\mathrm{vec}(u)^\top\Sigma_{\mathbf{X}}^{-1}\mathrm{vec}(\mathbf{X}_{(K)}) = \mathrm{vec}(u)^\top\Sigma_{\mathbf{X}}^{-1}\mathrm{vec}(\mathbf{X}).$$

Since λ_{K+1}^* equal to zero where λ_{K+1}^* is the maximum value of L_{K+1}, we have $\mathrm{var}[\mathrm{E}(\langle\mathbf{X}_{(k)}, \beta^{\circ(1\to m)}))] = 0$ for any rank one tensor. We can write any tensor γ as a linear combination of rank-one tensors, which shows that $\mathrm{var}[\mathrm{E}(\langle\mathbf{X}_{(k)}, \gamma) \mid Y] = 0$. Also, since we assume $\mathrm{E}(X) = 0$, we have $\mathrm{E}(\langle\mathbf{X}_{(k)}, \gamma\rangle) = 0$, and $\mathrm{E}\{[\mathrm{E}(\langle\mathbf{X}_{(k)}, \gamma) \mid Y)]^2\} = 0$, i.e., $\mathrm{E}(\langle\mathbf{X}_{(k)}, \gamma) \mid Y) = 0$. Then,

$$\mathrm{vec}(u)^\top\Sigma_{\mathbf{X}}^{-1}\mathrm{E}(\mathrm{vec}(\mathbf{X}) \mid Y) = \mathrm{E}[(\Sigma_{\mathbf{X}}^{-1}\mathrm{vec}(u))^\top\mathrm{vec}(\mathbf{X}_{(K)}) \mid Y] = 0.$$

Here we prove the first part of Theorem 1.

Now let us prove the second part, the relationship between $\mathcal{F}_K$ and $\mathcal{S}_{Y|\mathbf{X}}^{(1)}\circ\cdots\circ\mathcal{S}_{Y|\mathbf{X}}^{(m)}$, using mathematical induction. We begin with the case $K = 0$. Let $\beta_0^{(i)} = 0$, for $i = 1, \ldots, m$. We have $\beta_0^{(i)} \in \mathcal{S}_{Y|X}^{(i)}$ and

$$\mathcal{F}_0 \subseteq \mathcal{S}_{Y|X}^{(1)} \circ \cdots \circ \mathcal{S}_{Y|X}^{(m)}.$$

Then, it is enough to show that for any $k = 0, 1, \ldots, K-1$, if $\beta_j^{(i)} \in \mathcal{S}_{Y|\mathbf{X}}^{(i)}$, for $i = 1, \ldots, m$ and $j = 1, \ldots, k$, then $\beta_{k+1}^{(i)} \in \mathcal{S}_{Y|\mathbf{X}}^{(i)}$. To prove this result, it is enough to show that for an arbitrary rank-one tensor $\beta^{\circ(1\to m)}$ there exists a rank-one tensor $\beta_{\|}^{\circ(1\to m)}$, which yields a larger value of L_{k+1} and also satisfies $\beta_{\|}^{(i)} \in \mathcal{S}_{Y|\mathbf{X}}^{(i)}$ for $i = 1, \ldots, m$.

Because the covariance matrix $\Sigma_{\mathbf{X}}$ is decomposable. We assume that $\Sigma_{\mathbf{X}} = \Sigma_m \otimes \cdots \otimes \Sigma_1$. For $\beta^{(i)}$, the ith component of $\beta^{\circ(1\to m)}$, we can uniquely decompose it as the sum of two terms,

$$\beta^{(i)} = \beta_{\|}^{(i)} + \beta_{\perp}^{(i)},$$

where $\beta_{\|}^{(i)} \in \mathcal{S}_{Y|\mathbf{X}}^{(i)}$, and $\beta_{\perp}^{(i)} \in {\mathcal{S}_{Y|\mathbf{X}}^{(i)}}^{\perp}$ and ${\beta_{\|}^{(i)}}^{\top} \Sigma_i \beta_{\perp}^{(i)} = 0$. Therefore, $\beta^{\circ(1\to m)}$ can be written as

$$\beta^{\circ(1\to m)} = \beta_{\|}^{\circ(1\to m)} + \gamma_{\perp},$$

where

$$\gamma_{\perp} = \sum_{\zeta_1,\ldots,\zeta_m \in \{\|,\perp\}, \zeta_i = \perp \text{ for at least one } i} \beta_{\zeta_1}^{(1)} \circ \cdots \circ \beta_{\zeta_m}^{(m)}.$$

Hence, $\beta_{\|}^{\circ(1\to m)}$ is in the space of $\mathcal{S}_{Y|\mathbf{X}}^{(1)} \circ \cdots \circ \mathcal{S}_{Y|\mathbf{X}}^{(m)}$ and $\gamma_{\perp}$ is in the complementary space. One can verify that $(\beta_{\|}^{\circ(1\to m)})^{\top} \Sigma_{\mathbf{X}} \gamma_{\perp} = 0$.

Notice that $(I - P_{(k)})^{\top} \mathrm{vec}(\gamma_{\perp}) = \mathrm{vec}(\gamma_{\perp})$ because $P_{(k)}$ is the projection matrix onto $\mathcal{F}_k$. Therefore,

$$\mathrm{E}(\langle \mathbf{X}_{(k)}, \gamma_{\perp} \rangle \mid Y) = \mathrm{E}(\langle \mathbf{X}, \gamma_{\perp} \rangle \mid Y) = \langle \mathrm{E}(\mathbf{X} \mid Y), \gamma_{\perp} \rangle = 0,$$

where the last equality holds because of Lemma A.1. Hence the numerator of L_{k+1} is

$$\begin{aligned}\operatorname{var}[\mathrm{E}(\langle \mathbf{X}_{(k)}, \gamma\rangle \mid Y)] &= \operatorname{var}[\mathrm{E}(\langle \mathbf{X}_{(k)}, \beta_{\parallel}^{\circ(1\to m)}\rangle \mid Y) + \mathrm{E}(\langle \mathbf{X}_{(k)}, \gamma_{\perp}\rangle \mid Y)] \\ &= \operatorname{var}[\mathrm{E}(\langle \mathbf{X}_{(k)}, \beta_{\parallel}^{\circ(1\to m)}\rangle \mid Y)].\end{aligned}$$

Noticing that

$$\operatorname{var}(\mathbf{X}_{(k)}) = (I - P_{(k)})\Sigma_{\mathbf{X}}(I - P_{(k)})^{\top} = \Sigma_{\mathbf{X}} - \Sigma_{\mathbf{X}}\Gamma_{(k)}(\Gamma_{(k)}^{\top}\Sigma_{\mathbf{X}}\Gamma_{\mathbf{X}})^{-1}\Gamma_{(k)}^{\top}\Sigma_{\mathbf{X}},$$

we have

$$(\beta_{\parallel}^{\circ(1\to m)})^{\top}(I - P_{(k)})\Sigma_{\mathbf{X}}(I - P_{(k)})^{\top}\gamma_{\perp} = (\beta_{\parallel}^{\circ(1\to m)})^{\top}\Sigma_{\mathbf{X}}\gamma_{\perp} = 0,$$

where $\Gamma_{(k)}^{\top}\Sigma_{\mathbf{X}}\gamma_{\perp} = 0$ due to the definition of $\gamma_{\perp}$. Hence the denominator of L_{k+1} is

$$\operatorname{var}[\langle \mathbf{X}_{(k)}, \gamma\rangle] = \operatorname{var}[\langle \mathbf{X}_{(k)}, \beta_{\parallel}^{\circ(1\to m)}\rangle] + \gamma_{\perp}^{\top}\Sigma_{\mathbf{X}}\gamma_{\perp},$$

Therefore, we have

$$L_{k+1}(\beta_{\parallel}^{(1)}, \ldots, \beta_{\parallel}^{(m)}) \leq L_{k+1}(\beta^{(1)}, \ldots, \beta^{(m)}),$$

where the equality holds if and only if $\gamma_{\perp} = 0$. It is clear that $\beta_{\parallel}^{\circ(1\to m)} \notin \mathcal{F}_k$ when $L_{k+1}(\beta_{\parallel}^{(1)}, \ldots, \beta_{\parallel}^{(m)}) > 0$, because $L_{k+1} = 0$ otherwise. This completes the proof.

□

A.2 Proof of Theorem 3.2

Because the objective function in (9) contains a fraction, which makes it difficult to discuss the asymptotic properties of the estimators, for $k = 0$, we consider an equivalent but easier problem instead.

$$\max_{\beta^{(1)},\ldots,\beta^{(m)}} (\beta^{(m)} \otimes \cdots \otimes \beta^{(1)})^{\top} \widehat{M} (\beta^{(m)} \otimes \cdots \otimes \beta^{(1)}), \tag{A.1}$$

subject to

$$\beta^{(1)^{\top}} \beta^{(1)} = \cdots = \beta^{(m-1)^{\top}} \beta^{(m-1)} = (\beta^{\otimes(m \to 1)})^{\top} \widehat{\Sigma} (\beta^{\otimes(m \to 1)}) = 1.$$

The optimization of (A.1) is easier than (9), because $\beta^{(1)}$, ..., $\beta^{(m)}$ does not appear in the denominator. Notice that (9) and (A.1) have exactly the same maximum and also have exactly the same maximizer of $\beta^{(1)}$, ..., $\beta^{(m-1)}$. The maximizer of $\beta^{(m)}$, however, is different only by a multiplier. With a slight abuse of notation, we still use $\widehat{\beta}_1^{(1)}$, ..., $\widehat{\beta}_1^{(m)}$ to denote the maximizer of (A.1) and $\widehat{\lambda}_1$ to denote its maximum.

Write the Lagrangian function for (A.1) as

$$\begin{aligned} & Q(\beta^{(1)}, \ldots, \beta^{(m)}, \xi_1, \ldots, \xi_m) \\ & = (\beta^{\otimes(m \to 1)})^{\top} \widehat{M} (\beta^{\otimes(m \to 1)}) - \sum_{k=1}^{m-1} \xi_k (\beta^{(k)^{\top}} \beta^{(k)} - 1) \\ & - \xi_m [(\beta^{\otimes(m \to 1)})^{\top} \widehat{\Sigma} (\beta^{\otimes(m \to 1)}) - 1], \end{aligned}$$

where $\xi_1, \ldots, \xi_m$ are scalar Lagrangian multipliers. When finding the maximizer of $\beta^{(i)}$s, the values of ξ_is are also found together. Hence we define an augmented vector of parameters

$$\theta = (\beta^{(1)^\top}, \ldots, \beta^{(m)^\top}, \xi_1, \ldots, \xi_m)^\top$$

to include all of them. Denote the true parameter as

$$\theta_* = (\beta_{1*}^{(1)^\top}, \ldots, \beta_{1*}^{(m)^\top}, \xi_{1*}, \ldots, \xi_{m*})^\top.$$

Let $\Psi(\theta)$ be the Jacobian of function $Q(\theta)$. When the maximizer of $\widehat{\theta}$ is in the interior of its domain, it is a root of

$$\Psi(\theta) = \frac{\partial Q(\theta)}{\partial \theta} = 0.$$

The explicit expression of $\Psi(\theta)$ is in equation (A.2). Notice that $\Psi(\theta)$ may have multiple roots and we assume that θ_* is the root that corresponds to the maximum λ_{1*}, and $\widehat{\theta}$ is the root of $\Psi(\theta)$ corresponding to the maximum $\widehat{\lambda}_1$.

Let us outline the derivation of the asymptotic normality of $\widehat{\theta}$. Under mild conditions (see Condition 3.1-3.3 for details), expanding $\Psi(\widehat{\theta})$ in a neighborhood of θ_* yields

$$0 = \Psi(\widehat{\theta}) = \Psi(\theta_*) + \frac{\partial \Psi(\theta_*)}{\partial \theta^\top}(\widehat{\theta} - \theta_*) + o_p(\|\widehat{\theta} - \theta_*\|),$$

where $\|\cdot\|$ is the ℓ_2-norm of a vector, which further implies

$$\sqrt{n}(\widehat{\theta}-\theta_*) = \Big[\frac{\partial\Psi(\theta_*)}{\partial\theta^\top}\Big]^{-1}\Big(-\sqrt{n}\Psi(\theta_*)\Big) + o_p(\|\sqrt{n}(\widehat{\theta}-\theta_*)\|),$$

provided that $\partial\Psi(\theta_*)/\partial\theta^\top$ is invertible. In order to show that $\sqrt{n}(\widehat{\theta}-\theta_*)$ asymptotically follows a normal distribution, it is enough to show that $\partial\Psi(\theta_*)/\partial\theta^\top$ converges to a nonsingular constant matrix in probability and $\sqrt{n}\Psi(\theta_*)$ asymptotically follows a normal distribution.

In fact, it is shown in equation (A.2) that $\Psi(\theta)$ can be written as

$$\Psi(\theta) = \iint \psi(\theta;\mathbf{X}_1,y_1,\mathbf{X}_2,y_2)dF_n(\mathbf{X}_1,y_1)dF_n(\mathbf{X}_2,y_2),$$

where F_n is the empirical cumulative distribution function and the exact expression of $\psi(\theta;\mathbf{X}_1,y_1,\mathbf{X}_2,y_2)$ is in equation (A.3). Hence Ψ is a V-statistic, and it asymptotically follows a normal distribution Serfling (1980). Because $\mathrm{E}[\psi(\theta;\mathbf{X}_1,Y_1,\mathbf{X}_2,Y_2)] = 0$, we can have

$$\begin{aligned}\Psi(\theta) =& \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}_{\mathbf{X}_2,Y_2}[\psi(\theta;\mathbf{X}_1,Y_1,\mathbf{X}_2,Y_2)\\ &+ \psi(\theta;\mathbf{X}_2,Y_2,\mathbf{X}_1,Y_1) \mid (\mathbf{X}_1,Y_1),(\mathbf{X}_i,y_i)] + o_p(n^{-1/2}),\end{aligned}$$

where the expectation is taken with respect to $(\mathbf{X}_2,Y_2)$. By the Central Limit Theorem, we know that

$$\sqrt{n}\,\Psi(\theta_*) \to N(0, \Sigma_\psi) \text{ in distribution, as } n \to \infty,$$

where Σ_ψ is the covariance matrix of $\mathrm{E}_{\mathbf{X}_2,Y_2}[\psi(\theta; \mathbf{X}_1, Y_1, \mathbf{X}_2, Y_2) + \psi(\theta; \mathbf{X}_2, Y_2, \mathbf{X}_1, Y_1) \mid (\mathbf{X}_1, Y_1)]$. If we further have that

$$\frac{\partial \Psi(\theta)}{\partial \theta^\top} \to \mathrm{E}(A) \text{ in probability, as } n \to \infty,$$

where $A = \partial\psi(\theta)/\partial\theta^\top$, assuming $\mathrm{E}(A)$ is nonsingular, then by Slutsky's Theorem,

$$\sqrt{n}(\widehat{\theta} - \theta) \to N(0, \mathrm{E}(A)^{-1}\Sigma_\psi \mathrm{E}(A)^{-1}) \text{ in distribution, as } n \to \infty.$$

The above argument outlines the proof of the asymptotic normality of $\widehat{\theta}$. Notice that the above argument essentially treats $\widehat{\theta}$ as an M-estimator. See Serfling (1980) for more discussions on M-estimators. The following theorem is the consequence of the asymptotic normality of $\widehat{\theta}$ and the fact that $\widehat{\lambda} = (\widehat{\xi}_1 + \cdots + \widehat{\xi}_{m-1})/(m-1) + \widehat{\xi}_m$.

The detailed proof is presented below. Let us work out an explicit expression for $\Psi(\theta)$, where

$$\Psi(\theta) = \frac{\partial Q(\theta)}{\partial \theta} = \begin{pmatrix} \frac{\partial Q(\theta)}{\partial \beta^{(1)}} \\ \vdots \\ \frac{\partial Q(\theta)}{\partial \beta^{(m)}} \\ \frac{\partial Q(\theta)}{\partial \xi_1} \\ \vdots \\ \frac{\partial Q(\theta)}{\partial \xi_m} \end{pmatrix}. \tag{A.2}$$

Recall that $\beta^{\otimes(m\to 1)} = \beta^{(m)} \otimes \cdots \otimes \beta^{(1)}$. For ease of presentation, we define $B^{(k)}$ as replacing $\beta^{(k)}$ in $\beta^{\otimes(m\to 1)}$ by I_{p_k}. Noticing that $\beta^{\otimes(m\to 1)} = B^{(k)}\beta^{(k)}$, we have

$$\frac{\partial Q(\theta)}{\partial \beta^{(k)}} = \begin{cases} 2(B^{(k)})^\top(\widehat{M} - \xi_m\widehat{\Sigma})\beta^{\otimes(m\to 1)} - 2\xi_k\beta^{(k)}, & k = 1, \ldots, m-1 \\ 2(B^{(m)})^\top(\widehat{M} - \xi_m\widehat{\Sigma})\beta^{\otimes(m\to 1)}, & k = m \end{cases}$$

and

$$\frac{\partial Q(\theta)}{\partial \xi_k} = \begin{cases} 1 - {\beta^{(k)}}^\top\beta^{(k)}, & k = 1, \ldots, m-1 \\ 1 - (\beta^{\otimes(m\to 1)})^\top\widehat{\Sigma}(\beta^{\otimes(m\to 1)}), & k = m \end{cases}$$

We can derive a relationship between $\widehat{\lambda}$ and $\widehat{\xi}_k$'s. Noticing that $\partial Q(\widehat{\theta})/\partial\beta^{(k)} = 0$, we have

$$0 = \sum_{k=1}^{m-1} {\beta^{(k)}}^\top \frac{\partial Q(\widehat{\theta})}{\partial \beta^{(k)}} = 2(m-1)\widehat{\lambda} - 2(m-1)\widehat{\xi}_m - 2(\widehat{\xi}_1 + \cdots + \widehat{\xi}_{m-1}),$$

which further yields

$$\widehat{\lambda} = (\widehat{\xi}_1 + \cdots + \widehat{\xi}_{m-1})/(m-1) + \widehat{\xi}_m.$$

Before proving Theorem 3.2, we first give the explicit expressions of ψ and A. Noticing that

$$\widehat{\Sigma} = \iint (\text{vec}(\mathbf{X}_1)\text{vec}(\mathbf{X}_1)^\top - \text{vec}(\mathbf{X}_1)\text{vec}(\mathbf{X}_2)^\top) dF_n(\mathbf{X}_1) dF_n(\mathbf{X}_2),$$

$$\begin{aligned}\widehat{M} = \iint \Big(\sum_{h=1}^{H} \frac{1}{p_h} \text{vec}(\mathbf{X}_1)\text{vec}(\mathbf{X}_2)^\top I_{1h} I_{2h} \\ - \text{vec}(\mathbf{X}_1)\text{vec}(\mathbf{X}_2)^\top \Big) dF_n(\mathbf{X}_1, y_1) dF_n(\mathbf{X}_2, y_2),\end{aligned}$$

where $I_{ih} = 1$ if y_i is in the hth slice and 0 otherwise. Hence

$$\widehat{M} - \xi_m \widehat{\Sigma} = \iint Z_{12}\, dF_n(\mathbf{X}_1, y_1) dF_n(\mathbf{X}_2, y_2),$$

where

$$\begin{aligned}Z_{12} = \sum_{h=1}^{H} \frac{1}{p_h} \text{vec}(\mathbf{X}_1)\text{vec}(\mathbf{X}_2)^\top I_{1h} I_{2h} - \text{vec}(\mathbf{X}_1)\text{vec}(\mathbf{X}_2)^\top \\ - \xi_m \text{vec}(\mathbf{X}_1)\text{vec}(\mathbf{X}_1)^\top + \xi_m \text{vec}(\mathbf{X}_1)\text{vec}(\mathbf{X}_2)^\top.\end{aligned}$$

Therefore, we can write $\Psi(\theta)$ as

$$\Psi(\theta) = \iint \psi(\theta; \mathbf{X}_1, y_1, \mathbf{X}_2, y_2)\ dF_n(\mathbf{X}_1, y_1) dF_n(\mathbf{X}_2, y_2)$$

where

$$\psi(\theta; \mathbf{X}_1, y_1, \mathbf{X}_2, y_2) = \begin{pmatrix} \psi_{\beta_1}(\theta) \\ \vdots \\ \psi_{\beta_m}(\theta) \\ \psi_{\xi_1}(\theta) \\ \vdots \\ \psi_{\xi_m}(\theta) \end{pmatrix} \tag{A.3}$$

and

$$\psi_{\beta_k}(\theta) = \begin{cases} 2(B^{(k)})^\top Z_{12}(\beta^{\otimes(m\to 1)}) - 2\xi_k \beta^{(k)}, & k = 1, \dots, m-1, \\ 2(B^{(m)})^\top Z_{12}(\beta^{\otimes(m\to 1)}), & k = m \end{cases}$$

and

$$\psi_{\xi_k}(\theta) = \begin{cases} 1 - {\beta^{(k)}}^\top \beta^{(k)}, \quad k = 1, \dots, m-1 \\ 1 - (\beta^{\otimes(m\to 1)})^\top (\mathrm{vec}(\mathbf{X}_1)\mathrm{vec}(\mathbf{X}_1)^\top \\ -\mathrm{vec}(\mathbf{X}_1)\mathrm{vec}(\mathbf{X}_2)^\top)(\beta^{\otimes(m\to 1)}),\ k = m \end{cases}$$

Note that A is a symmetric matrix and it can be expressed as follows.

$$A = \frac{\partial \psi(\theta)}{\partial \theta^\top} = \begin{pmatrix} \frac{\partial \psi_{\beta_1}(\theta)}{\partial \beta_1^\top} & \cdots & \frac{\partial \psi_{\beta_1}(\theta)}{\partial \beta_m^\top} & \cdots & \frac{\partial \psi_{\beta_1}(\theta)}{\partial \xi_1} & \cdots & \frac{\partial \psi_{\beta_1}(\theta)}{\partial \xi_m} \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \\ \frac{\partial \psi_{\beta_m}(\theta)}{\partial \beta_1^\top} & \cdots & \frac{\partial \psi_{\beta_m}(\theta)}{\partial \beta_m^\top} & \cdots & \frac{\partial \psi_{\beta_m}(\theta)}{\partial \xi_1} & \cdots & \frac{\partial \psi_{\beta_m}(\theta)}{\partial \xi_m} \\ \frac{\partial \psi_{\xi_1}(\theta)}{\partial \beta_1^\top} & \cdots & \frac{\partial \psi_{\xi_1}(\theta)}{\partial \beta_m^\top} & \cdots & \frac{\partial \psi_{\xi_1}(\theta)}{\partial \xi_1} & \cdots & \frac{\partial \psi_{\xi_1}(\theta)}{\partial \xi_m} \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \\ \frac{\partial \psi_{\xi_m}(\theta)}{\partial \beta_1^\top} & \cdots & \frac{\partial \psi_{\xi_m}(\theta)}{\partial \beta_m^\top} & & \frac{\partial \psi_{\xi_m}(\theta)}{\partial \xi_1} & \cdots & \frac{\partial \psi_{\xi_m}(\theta)}{\partial \xi_m} \end{pmatrix}$$

where for diagonal elements in the first block,

$$\frac{\partial \psi_{\beta_k}(\theta)}{\partial \beta_k^\top} = \begin{cases} 2(B^{(k)})^\top Z_{12}(B^{(k)}) - 2\xi_k I_{p_k}, & k = 1, \ldots, m-1 \\ 2(B^{(m)})^\top Z_{12}(B^{(m)}), & k = m \end{cases}$$

for off-diagonal elements in the first block, $k_1, k_2 = 1, \ldots, m$, and $k_1 \neq k_2$,

$$\begin{aligned} \frac{\partial \psi_{\beta_{k_1}}(\theta)}{\partial \beta_{k_2}^\top} &= 2[(\beta^{(m)} \otimes \cdots \otimes I_{p_{k_1}} \otimes I_{p_{k_2}} \otimes \cdots \otimes \beta^{(1)})^\top Z_{12}(\beta^{\otimes(m \to 1)})]_{p_{k_2} \times p_{k_1}}^\top \\ &\quad + 2(B^{(k_1)})^\top Z_{12}(B^{(k_2)}) \end{aligned}$$

where $[\cdot]_{p_1 \times p_2}$ means converting this vector to a $p_1 \times p_2$ matrix.

We also have

$$\frac{\partial \psi_{\beta_{k_1}}(\theta)}{\partial \xi_{k_2}} = \begin{cases} -2\beta^{(k_1)}, & k_1 = k_2, k_2 \neq m \\ 0, & k_1 \neq k_2, k_2 \neq m \end{cases}$$

For $k = 1, \dots, m$,

$$\frac{\partial \psi_{\beta_k}(\theta)}{\partial \xi_m} = -2(B^{(k)})^\top (\mathrm{vec}(\mathbf{X}_1)\mathrm{vec}(\mathbf{X}_1)^\top - \mathrm{vec}(\mathbf{X}_1)\mathrm{vec}(\mathbf{X}_2)^\top)(\beta^{\otimes(m\to 1)}),$$

$$\frac{\partial \psi_{\xi_{k_1}}(\theta)}{\partial \xi_{k_2}} = 0, \quad k_1, k_2 = 1, \dots, m.$$

Therefore, the matrix A can be written as

$$\begin{pmatrix}
\frac{\partial \psi_{\beta_1}(\theta)}{\partial \beta_1^\top} & \cdots & \frac{\partial \psi_{\beta_1}(\theta)}{\partial \beta_{m-1}^\top} & \frac{\partial \psi_{\beta_1}(\theta)}{\partial \beta_m^\top} & \cdots & -2\beta^{(1)} & \cdots & 0 & \frac{\partial \psi_{\beta_1}(\theta)}{\partial \xi_m} \\
\vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \ddots & & \vdots \\
\frac{\partial \psi_{\beta_{m-1}}(\theta)}{\partial \beta_1^\top} & \cdots & \frac{\partial \psi_{\beta_{m-1}}(\theta)}{\partial \beta_{m-1}^\top} & \frac{\partial \psi_{\beta_{m-1}}(\theta)}{\partial \beta_m^\top} & \cdots & 0 & \cdots & -2\beta^{(m-1)} & \frac{\partial \psi_{\beta_{m-1}}(\theta)}{\partial \xi_m} \\
\frac{\partial \psi_{\beta_m}(\theta)}{\partial \beta_1^\top} & \cdots & \frac{\partial \psi_{\beta_m}(\theta)}{\partial \beta_{m-1}^\top} & \frac{\partial \psi_{\beta_m}(\theta)}{\partial \beta_m^\top} & \cdots & 0 & \cdots & 0 & \frac{\partial \psi_{\beta_m}(\theta)}{\partial \xi_m} \\
-2{\beta^{(1)}}^\top & \cdots & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 \\
\vdots & \cdots & \vdots & \vdots & \cdots & \vdots & & & \\
0 & \cdots & -2{\beta^{(m-1)}}^\top & 0 & & 0 & \cdots & 0 & 0 \\
\frac{\partial \psi_{\xi_m}(\theta)}{\partial \beta_1^\top} & \cdots & \frac{\partial \psi_{\xi_m}(\theta)}{\partial \beta_{m-1}^\top} & \frac{\partial \psi_{\xi_m}(\theta)}{\partial \beta_m^\top} & & 0 & \cdots & 0 & 0
\end{pmatrix}$$

Next, we state the assumptions for Theorem 3.2 as follows.

(A1) The true parameter θ_0 is an interior point of Θ, where Θ is the domain of θ. It is the unique root of $\mathrm{E}[\psi(\theta; \mathbf{X}_1, y_1, \mathbf{X}_2, y_2)] = 0$ in its neighborhood.

(A2) The second moment of $\psi(\theta; \mathbf{X}_1, y_1, \mathbf{X}_2, y_2)$ are bounded.

(A3) The matrix $\mathrm{E}(A)$ is nonsingular.

According to the discussion in the sketch of proof, it is enough to show two issues: (1) $\sqrt{n}\Psi(\theta)$ asymptotically follows a normal distribution and (2) $\partial\Psi(\theta)/\partial\theta^T$ converges to A in probability.

We first show that $\sqrt{n}\Psi(\theta)$ asymptotically follows a normal distribution. In fact, $\sqrt{n}\Psi(\theta)$ can be expressed as a V-statistic. Notice that according to the theory of V-statistic Serfling (1980), we can have the following expansion,

$$\begin{aligned}\Psi(\theta) =& \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{\mathbf{X}_2, Y_2}[\psi(\mathbf{X}_1, Y_1, \mathbf{X}_2, Y_2) + \psi(\mathbf{X}_2, Y_2, \mathbf{X}_1, y_1) \mid (\mathbf{X}_1, Y_1) \\ =& (\mathbf{X}_i, y_i)] + o_p(n^{-1/2}).\end{aligned}$$

Therefore, $\sqrt{n}\Psi(\theta)$ asymptotically follows a normal distribution.

We now show that $\partial\Psi(\theta)/\partial\theta^\top$ converges to A in probability. Because

$$\frac{\partial\Psi(\theta)}{\partial\theta^\top} = \iint A dF_n(\mathbf{X}_1, y_1) dF_n(\mathbf{X}_2, y_2),$$

by the Law of Large Numbers, $\partial\Psi(\theta)/\partial\theta^\top$ converges to $\mathrm{E}(A)$.

Noticing that $({\beta_1^{(1)}}^\top, \ldots, {\beta_1^{(m)}}^\top, \lambda_1)^\top = B\theta$, this theorem holds.

A.3 Proof of Theorem 3.3

Sequential test for the number of tensors: Based on the asymptotic distribution derived in Theorem 2, we derive a sequential testing procedure to determine K, the number of rank-one tensors in pTDR model. Recall that in the population version, the rank-one tensors maximize (6) and (7). When Y is independent of $\mathbf{X}_{(k)}$, any index of $\mathbf{X}_{(k)}$ is uncorrelated with any transformed Y, and then the maximum $\lambda^*_{k+1} = 0$. Hence the null hypothesis is $H_0 : \ \lambda^*_{k+1} = 0$. We will sequentially test $H_0 : \lambda^*_{k+1} = 0$ for $k = 0, 1, 2, \ldots$. When testing $H_0 : \lambda^*_{k+1} = 0$, we treat the previously identified k rank-one tensors are given, or equivalently, the test is conditional on the first k rank-one tensors. Suppose that the test does not first reject for an integer $k + 1$. Then we should keep k rank-one tensors, or $\widehat{K} = k$. Because the asymptotic results for estimated rank-one tensors are similar, the test procedure is also similar at each step of the sequential procedure. We preset the result in terms of $k = 0$ for the first rank-one tensor.

Proof. Let us outline the derivation of the asymptotic distribution of $\widehat{\Lambda}$ under H_0. First, $\widehat{S}^2_{(1)}$ can be expressed as a V-statistic. In fact

$$\begin{aligned}
\widehat{S}^2_{(1)} &= n \sum_{h=1}^{H} p_h \mathrm{vec}(\bar{\mathbf{X}}_h - \bar{\mathbf{X}})^\top \mathrm{vec}(\bar{\mathbf{X}}_h - \bar{\mathbf{X}}) \\
&= n \iint \Phi(\mathbf{X}_1, y_1, \mathbf{X}_2, y_2) dF_n(\mathbf{X}_1, y_1) dF_n(\mathbf{X}_2, y_2),
\end{aligned}$$

where

$$\Phi(\mathbf{X}_1, y_1, \mathbf{X}_2, y_2) = \sum_{h=1}^{H}(\frac{1}{p_h} I_{1h} I_{2h} - p_h)\mathrm{vec}(\mathbf{X}_1)^\top \mathrm{vec}(\mathbf{X}_2),$$

where $I_{ih} = I(y_i \text{ in the } h\text{th slice})$. We also know that $I_{ih} I_{jh} = 0$ if $i \neq j$ because one observation can only belong to one slice. Without loss of generality, we assume that $\mathrm{E}(\mathbf{X}) = 0$. Under H_0, $\mu_h = \mathrm{E}(\mathbf{X} \mid Y \in I_h) = 0$ for $h = 1, \ldots, H$. Because

$$\mathrm{E}_{\mathbf{X}_2, Y_2}[\Phi(\mathbf{X}_1, Y_1, \mathbf{X}_2, Y_2) \mid (\mathbf{X}_1, Y_1)] = \sum_{h=1}^{H} \mathrm{vec}(\mathbf{X}_1)^\top I_{1h} \mathrm{vec}(\mu_h) = 0,$$

we know that $\widehat{S}^2_{(1)}$ is a first-order degenerated V-statistic. Due to Serfling (1980), the asymptotic distribution of $\widehat{S}^2_{(1)}$ is a weighted chi-squared distribution with weights as eigenvalues of $\Phi(\mathbf{X}_1, y_1, \mathbf{X}_2, y_2)$. More formally, a constant z_k is called an eigenvalue of Φ and a function $\varphi_k(\mathbf{X}, y)$ is called the associated eigenfunction if

$$\mathrm{E}_{\mathbf{X}_1, Y_1}[\Phi(\mathbf{X}_1, Y_1, \mathbf{X}_2, Y_2)\varphi_k(\mathbf{X}_1, Y_1)] = z_k \varphi_k(\mathbf{X}_2, Y_2),$$

where the expectation is taken with respect to $(\mathbf{X}_1, Y_1)$. Following the Fredholm theory of integral equations, there exists sequences of eigenvalues and eigenfunctions, z_k and $\varphi_k(\mathbf{X})$, $k = 1, 2, \ldots$, such that the function Φ admits the expansion

$$\Phi(\mathbf{X}_1, y_1, \mathbf{X}_2, y_2) = \sum_{k=1}^{\infty} z_k \varphi_k(\mathbf{X}_1, y_1)\varphi_k(\mathbf{X}_2, y_2),$$

where $\mathrm{E}[\varphi_i(\mathbf{X}, Y)\varphi_j(\mathbf{X}, Y)] = 1$ if $i = j$ and $= 0$ if $i \neq j$. Since the constant 1 is an eigenfunction corresponding to the eigenvalue zero, we have $\mathrm{E}[\varphi_k(\mathbf{X}, Y)] = 0$ for $k = 1, 2, \ldots$. The following theorem asserts that under H_0, $\widehat{S}^2$ asymptotically follows a weighted chi-squared distribution with weights being exactly the eigenvalues of $\Phi(\mathbf{X}_1, y_1, \mathbf{X}_2, y_2)$. □

Supplementary Materials

The Supplementary contains additional simulation results.

References

Aguet, F. and Muñoz Aguirre, M. (2017). Genetic effects on gene expression across human tissues. *Nature*, 550:204–213.

Bura, E. and Cook, R. D. (2001). Extending sliced inverse regression: the weighted chi-squared test. *Journal of the American Statistical Association*, 96:996–1003.

Chen, C.-H. and Li, K.-C. (1998). Can SIR be as popular as multiple linear regression? *Statist. Sin.*, 8:289–316.

Cook, R. D. (1996). Graphics for regressions with a binary response. *J. Amer. Statist. Assoc.*, 91:983–992.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics.* John Wiley & Sons.

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *The Annals of Statistics*, 32(3):1062–1092.

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression. *Journal of the American Statistical Association*, 100:410–428.

Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics.* John Wiley & Sons.

Ding, S. and Cook, R. D. (2015). Tensor sliced inverse regression. *Journal of Multivariate Analysis*, 133:216–231.

Erola, P., Björkegren, J. L., and Michoel, T. (2020). Model-based clustering of multi-tissue gene expression data. *Bioinformatics*, 36(6):1807–1813.

Fung, W. K., He, X., Liu, L., and Shi, P. (2002). Dimension reduction based on canonical correlation. *Statistica Sinica*, 12:1093–1113.

Gamazon, E. R., Zwinderman, A. H., Cox, N. J., Denys, D., and Derks, E. M. (2019). Multi-tissue transcriptome analyses identify genetic mechanisms underlying neuropsychiatric traits. *Nature genetics*, 51(6):933–940.

Glastonbury, C. A., Viñuela, A., Buil, A., Halldorsson, G. H., Thorleifsson, G., Helgason, H., Thorsteinsdottir, U., Stefansson, K., Dermitzakis, E. T., Spector, T. D., et al. (2016). Adiposity-dependent regulatory effects on multi-tissue transcriptomes. *The American Journal of Human Genetics*, 99(3):567–579.

Grundberg, E., Small, K. S., Hedman, Å. K., Nica, A. C., Buil, A., Keildson, S., Bell, J. T., Yang, T.-P., Meduri, E., Barrett, A., et al. (2012). Mapping cis-and trans-regulatory effects across multiple tissues in twins. *Nature genetics*, 44(10):1084–1089.

Hall, P. and Li, K.-C. (1993). On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, 21:867–889.

Harshman, R. and Lundy, M. (1984). The PARAFAC model for three-way factor analysis and multidimensional scaling. In *Research Methods for Multimode Data Analysis*, volume 46, pages 122–215.

Herrera, B. M., Keildson, S., and Lindgren, C. M. (2011). Genetics and epigenetics of obesity. *Maturitas*, 69(1):41–49.

Hooper, J. W. (1959). Simultaneous equations and canonical correlation theory. *Econometrica: Journal of the Econometric Society*, 27:245–256.

Hore, V., Viñuela, A., Buil, A., Knight, J., McCarthy, M. I., Small, K., and Marchini, J. (2016). Tensor decomposition for multiple-tissue gene expression experiments. *Nature genetics*, 48(9):1094–1100.

Kaminsky, Z., Tochigi, M., Jia, P., Pal, M., Mill, J., Kwan, A., Ioshikhes, I., Vincent, J., Kennedy, J., Strauss, J., et al. (2012). A multi-tissue analysis identifies hla complex group 9 gene methylation differences in bipolar disorder. *Molecular psychiatry*, 17(7):728–740.

Kessler, D. C., Taylor, J. A., and Dunson, D. B. (2014). Learning phenotype densities conditional on many interacting predictors. *Bioinformatics*, 30(11):1562–1568.

Kolda, T. and Bader, B. (2009). Tensor decompositions and applications. *SIAM Review*, 51(3):455–500.

Kolda, T. G. (2001). Orthogonal tensor decompositions. *SIAM J. Matrix Anal. Appl.*, 23:243–255.

Li, B., Kim, M. K., and Altman, N. (2010). On dimension folding of matrix or array valued statistical objects. *Ann. Statist.*, 38(8):1094–1121.

Li, B. and Wang, S. (2007). On directional regresssion for dimension reduction. *jasa*, 102:997–1008.

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

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585.

Merris, R. (1997). *Multilinear Algebra*. CRC Press.

Moody, H. R. and Sasser, J. R. (2020). *Aging: Concepts and controversies*. Sage publications.

Nilsson, J., Sha, F., and Jordan, M. I. (2007). Regression on manifolds using kernel dimension reduction. In *Proceedings of the 24th international conference on Machine learning*, pages 697–704.

Rocke, D. M. and Woodruff, D. L. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 91(435):1047–1061.

Sam, S. and Mazzone, T. (2014). Adipose tissue changes in obesity and the impact on metabolic function. *Translational Research*, 164(4):284–292.

Sanyal, D. and Raychaudhuri, M. (2016). Hypothyroidism and obesity: An intriguing link. *Indian Journal of*

Endocrinology and Metabolism, 20(4):554–557.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons.

Smilde, A., Bro, R., and Geladi, P. (2004). *Multi-way Analysis with Applications in the Chemical Sciences.* John Wiley & Sons.

Song, R.-h., Wang, B., Yao, Q.-m., Li, Q., Jia, X., and Zhang, J.-a. (2019). The impact of obesity on thyroid autoimmunity and dysfunction: a systematic review and meta-analysis. *Frontiers in immunology*, 10:2349.

Szilard, L. (1959). On the nature of the aging process. *Proceedings of the National Academy of Sciences of the United States of America*, 45(1):30.

Talukdar, H. A., Asl, H. F., Jain, R. K., Ermel, R., Ruusalepp, A., Franzén, O., Kidd, B. A., Readhead, B., Giannarelli, C., Kovacic, J. C., et al. (2016). Cross-tissue regulatory gene networks in coronary artery disease. *Cell systems*, 2(3):196–208.

Tucker, L. (1951). A method for synthesis of factor analysis studies. *Personnel Research Section,*, Report 984(Dept. of Army).

Tucker, L. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311.

Valenzuela, P. L., Maffiuletti, N. A., Tringali, G., De Col, A., and Sartorio, A. (2020). Obesity-associated poor muscle quality: prevalence and association with age, sex, and body mass index. *BMC Musculoskelet Disord*, 21:1–8.

Walker, R. F., Pakula, L. C., Sutcliffe, M. J., Kruk, P. A., Graakjaer, J., and Shay, J. W. (2009). A case study of "disorganized development" and its possible relevance to genetic determinants of aging. *Mechanisms of ageing and development*, 130(5):350–356.

Xia, Y. (2008). A multiple-index model and dimension reduction. *Journal of the American Statistical Association*, 103(484):1631–1640.

Yang, J., Huang, T., Petralia, F., Long, Q., Zhang, B., Argmann, C., Zhao, Y., Mobbs, C. V., Schadt, E. E., Zhu, J., et al. (2015). Synchronized age-related gene expression changes across multiple tissues in human and the link to complex diseases. *Scientific reports*, 5(1):1–16.

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, 98(464):968–979.

Zeng, P. (2008). Determining the dimension of the central subspace and central mean subspace. *Biometrika*, 95:469–479.

Zhong, W., Zeng, P., Ma, P., Liu, J. S., and Zhu, Y. (2005). Rsir: regularized sliced inverse regression for motif discovery. *Bioinformatics*, 21(22):4169–4175.

Zhong, W., Zhang, T., Zhu, Y., and Liu, J. (2012). Correlation pursuit: forward stepwise variable selection for index model. *J. Roy. Statist. Soc. Ser. B*, 74:849–870.

Zhou, H. and Li, L. (2014). Regularized matrix regression. *J. Roy. Statist. Soc. Ser. B*, 76:463–483.

Zhou, H., Li, L., and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *J. Amer. Statist. Assoc.*, 108:540–552.

Zhu, L.-X. and Ng, K. W. (1995). Asymptotics of sliced inverse regression. *Statistica Sinica*, pages 727–736.