Statistica Sinica Preprint No: SS-2022-0075	
Title	Mean Dimension Reduction And Testing For Nonparametric Tensor Response Regression
Manuscript ID	SS-2022-0075
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0075
Complete List of Authors	Chung Eun Lee, Xin Zhang and Lexin Li
Corresponding Authors	Chung Eun Lee
E-mails	chungeun.lee@baruch.cuny.edu
Notice: Accepted version subject to English editing.	

MEAN DIMENSION REDUCTION AND TESTING FOR NONPARAMETRIC TENSOR RESPONSE REGRESSION

Chung Eun Lee, Xin Zhang, and Lexin Li

Baruch College, Florida State University, and University of California at Berkeley

Abstract: In this article, we propose a flexible model-free approach to the regression analysis of a tensor response and a vector predictor. Without specifying the specific form of the regression mean function, we consider two closely related statistical problems: (i) estimation of the dimension reduction subspace that captures all the variations in the regression mean function, and (ii) hypothesis testing of whether the conditional expectation of a linear dimension reduction of the response given the predictor is invariant to the changes in the predictor. We propose a new nonparametric metric called tensor martingale difference divergence, and study its statistical properties. Built on this new metric, we develop computationally efficient estimation and asymptotically valid testing procedures. We demonstrate the efficacy of our method through both simulations and two real data applications for macroeconomics and e-commerce.

Key words and phrases: Martingale difference divergence; Nonlinearity; Sufficient dimension reduction; Tensor decomposition; Tensor regression; Wild bootstrap.

1. Introduction

Tensor data is now becoming ubiquitous in a wide range of scientific and business applications. For instance, in economics, multiple macroeconomic indices at varying time points across different countries are assembled as a three-way tensor. In neuroscience, images obtained by anatomical or functional magnetic resonance imaging take the form of three-way or four-way tensors. Tensor data analysis is thriving in statistics and machine learning in recent years. See Bi et al. (2021), Sun et al. (2021) for reviews.

A central question in tensor analysis is finding meaningful low-dimensional tensor structures given the complex and high-dimensional tensor data. There is a line of research modeling the tensor as predictor or response in a regression setting. Early tensor predictor regression solutions focus on parametric and usually linear or generalized linear type models (Zhou et al. 2013, Li et al. 2018, Zhang & Li 2017, Chen et al. 2019). More recently, Hao et al. (2019), Zhou et al. (2020) extended tensor predictor regression to nonparametric models through basis expansion. Relatedly, Rabusseau & Kadri (2016), Li & Zhang (2017), Sun & Li (2017) studied tensor response regression under different low-dimensional structures, but all assumed linear association models and often assumed the normality distribution.

There is another line of research focusing on sufficient dimension reduction (SDR) without losing any regression information and without imposing any specific model form; see Li (2018) for a review. While most SDR solutions consider vector-valued regression, there have been extensions to matrix or tensor-valued regression (Li et al. 2010, Xue & Yin 2014,

Ding & Cook 2015, Sheng & Yuan 2020, Wang et al. 2022). Nevertheless, they all targeted tensor predictor instead of tensor response.

In this article, we propose a flexible and assumption-lean approach for tensor response regressions. Particularly, we seek linear subspaces that transform the tensor response into two parts: a low-dimensional part that contains all relevant mean function information, and the orthogonal part that is invariant to the change of the predictor values. Consequently, it effectively reduces the number of free parameters while retaining both the tensor structure and full regression mean information. It also lends naturally to sparse estimation and prediction. Furthermore, we develop a mean independence test using wild bootstrap to assess whether a linear reduction of tensor is mean independent of another random vector. We achieve our goals by generalizing the martingale difference divergence metric of Lee & Shao (2018) from vector response to tensor response, which fully quantifies the mean dependence between a tensor response and a vector predictor in a model-free fashion. Our proposal is nonparametric in nature, and does not impose any specific model forms or data distributions. This differentiates our solution from most of existing tensor analyses. Meanwhile, the extension from the vector case SDR and martingale difference divergence to the tensor case is far from trivial, and requires utterly new techniques. On the application side, our approach provides a useful tool for numerous types of applications. One example is neuroscience, where the scientific interest is to identify brain regions that exhibit different patterns based on magnetic resonance imaging between groups of patients with a neurological disorder and

healthy controls. Another example is educational study, where the interest is to understand the effect of trainings towards students' grades of series of testings in various courses. A third example is economics, where the interest is to predict different types of housing prices from different locations based on the macroeconomic indices such as interest rate, unemployment rate, and stock prices. All these questions can be formulated as regressions with a tensor response and a vector predictor.

The rest of the article is organized as follows. Section 2 presents our mean dimension reduction approach, and Section 3 develops a mean independence test. Section 4 presents the simulations, and Section 5 shows data applications. Section 6 concludes with a discussion, and the Supplementary Materials collect proofs and additional numerical results.

2. Mean Dimension Reduction

2.1 Flexible tensor response regression model

We consider an order-m response tensor $\mathcal{Y} \in \mathbb{R}^{r_1 \times \cdots \times r_m}$, and a vector predictor $X \in \mathbb{R}^p$. We impose a mild finite moment condition that $\mathbb{E}\left(\|\mathcal{Y}\|_F^2 + \|X\|^2\right) < \infty$, where $\|\cdot\|_F$ is the Frobenius norm. We begin with a quick review of the linear tensor response regression model of Li & Zhang (2017),

$$\mathcal{Y} = \widetilde{\mathcal{B}} \times_{(m+1)} X + \epsilon, \tag{2.1}$$

where the error tensor $\epsilon \in \mathbb{R}^{r_1 \times r_2 \times \ldots \times r_m}$ is independent of X and normally distributed, with $\mathbb{E}(\epsilon) = 0$ and $\mathrm{cov}\{\mathrm{vec}(\epsilon)\}$ having a separable Kronecker covariance structure, and $\widetilde{\mathcal{B}} \times_{(m+1)}$

X is the $(m+1)$-mode product where each element is the product of mode-$(m+1)$ fiber of $\widetilde{\mathcal{B}}$ multiplied by X. The coefficient tensor $\widetilde{\mathcal{B}} \in \mathbb{R}^{r_1 \times \cdots \times r_m \times p}$ is of a Tucker decomposition structure (Kolda & Bader 2009), $\widetilde{\mathcal{B}} = [\![\Phi; B_1, B_2, \ldots, B_m, I_p]\!] = \sum_{j_1=1}^{u_1} \cdots \sum_{j_m=1}^{u_m} b_{1,j_1} \circ \cdots \circ b_{m,j_m} \circ \phi_{j_1,\ldots,j_m}$, where $\Phi = (\phi_{k_1,\ldots,k_m}) \in \mathbb{R}^{u_1 \times \ldots \times u_m \times p}$, $B_k = (b_{k,1}, \cdots, b_{k,u_k}) \in \mathbb{R}^{r_k \times u_k}$, $u_k < r_k$, $k = 1, \ldots, m$, and $\circ$ is the vector outer product. Their goal was to uncover the subspaces $\mathrm{span}(B_k), k = 1, \ldots, m$ to reduce the dimension of $\mathcal{Y}$.

Next, we generalize the parametric model of (2.1), by considering a more flexible, non-parametric mean function, while imposing no distributional assumption on the error term. Specifically, we consider

$$\mathcal{Y} = \mathbb{E}(\mathcal{Y} \mid X) + \varepsilon = \mathcal{B} \times_{(m+1)} f(X) + \varepsilon, \tag{2.2}$$

where the error tensor $\varepsilon \in \mathbb{R}^{r_1 \times r_2 \times \ldots \times r_m}$ satisfies that $\mathbb{E}(\varepsilon \mid X) = 0$, and $f(\cdot) : \mathbb{R}^p \to \mathbb{R}^v$ is an arbitrary function. We further assume the coefficient tensor $\mathcal{B} = [\![\Theta; \beta_1, \beta_2, \ldots, \beta_m, I_v]\!] \in \mathbb{R}^{r_1 \times \ldots \times r_m \times v}$, where $\Theta \in \mathbb{R}^{d_1 \times \ldots \times d_m \times v}$, $\beta_k \in \mathbb{R}^{r_k \times d_k}$, $d_k < r_k$, $k = 1, \ldots, m$. Our goal is to uncover $\{\mathrm{span}(\beta_k)\}_{k=1}^m$ that captures all regression mean information of $\mathcal{Y}|X$ under (2.2).

We make a few remarks. First, the coefficient tensor $\mathcal{B}$ and the function f in (2.2) are not fully identifiable, but their product $\mathbb{E}(\mathcal{Y} \mid X) = \mathcal{B} \times_{(m+1)} f(X)$ is unique. Relatedly, each basis matrix β_k is not identifiable, but the corresponding subspace $\mathrm{span}(\beta_k)$ is unique in (2.2). Therefore, in our dimension reduction inquiry, we seek to uncover $\{\mathrm{span}(\beta_k)\}_{k=1}^m$. Second, our goal differs from that of Li & Zhang (2017). Both aim to reduce the dimension of a tensor response. However, Li & Zhang (2017) targeted the conditional mean while

accounting for the variance of $\mathcal{Y}$ given X under the linear model (2.1), whereas we target the conditional mean without imposing any distributional or linear model assumptions.

2.2 Tensor martingale difference divergence

We first briefly review the notion of martingale difference divergence proposed by Lee & Shao (2018), which quantifies the dependence between two random vectors. Specifically, for $U \in \mathbb{R}^u$ and $V \in \mathbb{R}^v$ satisfying $\mathbb{E}(\|U\|^2 + \|V\|^2) < \infty$, define the martingale difference divergence matrix (MDDM) as,

$$M(V \mid U) = -\mathbb{E}\left[\{V - \mathbb{E}(V)\}\{V' - \mathbb{E}(V')\}^{\mathrm{T}}\|U - U'\|\right] \in \mathbb{R}^{v\times v}, \tag{2.3}$$

where (U', V') is an independent copy of (U, V), and $\|\cdot\|$ is the Euclidean norm. By definition, $M(V \mid U)$ is a symmetric and positive semi-definite matrix. It characterizes the dependence of the conditional mean function $\mathbb{E}(V|U)$ on U, in that, when $\mathbb{E}(V|U) - \mathbb{E}(V)$ lies within a lower-dimensional subspace, the eigenvectors of $M(V \mid U)$ span the same subspace (Lee & Shao 2018).

Next, we extend this notion to the tensor case, and develop the concept of *tensor martingale difference divergence*, which is a *set* of symmetric positive semi-definite matrices. For the tensor $\mathcal{Y}$, its mode-k matricization, $\mathcal{Y}_{(k)}$, maps $\mathcal{Y}$ into an $r_k \times \prod_{j\neq k} r_j$ matrix so that the $(i_1, \ldots, i_m)$th element of $\mathcal{Y}$ maps to the (i_k, j)th element of $\mathcal{Y}_{(k)}$, with $j = 1 + \sum_{k'\neq k}(i_{k'} - 1)\prod_{k''<k', k''\neq k} r_{k''}$.

Definition 1 *Suppose* $\mathbb{E}(\|X\|^2 + \|\mathcal{Y}\|_F^2) < \infty$. *Let* $\mu_{(k)} = \mathbb{E}(\mathcal{Y}_{(k)})$. *Define the mode-k*

tensor martingale difference divergence between $\mathcal{Y}$ and X as,

$$M^{(k)}(\mathcal{Y} \mid X) = -\mathbb{E}\left\{\left(\mathcal{Y}_{(k)} - \mu_{(k)}\right)\left(\mathcal{Y}'_{(k)} - \mu_{(k)}\right)^{\mathrm{T}} \left\|X - X'\right\|\right\} \in \mathbb{R}^{r_k \times r_k},$$

Collectively, define the tensor martingale difference divergence as the set,

$$\mathcal{M}(\mathcal{Y} \mid X) = \left\{M^{(1)}(\mathcal{Y} \mid X), \ldots, M^{(m)}(\mathcal{Y} \mid X)\right\}.$$

Let $\mathcal{E}_k$ denote the column space spanned by $M^{(k)}(\mathcal{Y} \mid X)$, i.e., $\mathcal{E}_k = \operatorname{span}\{M^{(k)}(\mathcal{Y} \mid X)\} \subseteq \mathbb{R}^{r_k}$, and let $\mathcal{E}_k^{\perp}$ denote its complement space. We next establish some properties of $\mathcal{M}(\mathcal{Y} \mid X)$ through $\mathcal{E}_k$, and also the connection between $\mathcal{M}(\mathcal{Y} \mid X)$ and its MDDM counterpart for the vector case.

Proposition 1 *The following statements are true about $\mathcal{E}_k$.*

(i) $\mathbb{E}(\mathcal{Y} \times_{(k)} Q_k \mid X) = \mathbb{E}(\mathcal{Y} \times_{(k)} Q_k)$ *almost surely, where Q_k is the projection onto $\mathcal{E}_k^{\perp}$.*

(ii) $\mathcal{E}_k = \sum_j \operatorname{span}\left\{M(\mathcal{Y}_{(k),j} \mid X)\right\}$, *where $\mathcal{Y}_{(k),j} \in \mathbb{R}^{r_k}$ is the jth column of $\mathcal{Y}_{(k)}$.*

(iii) $d_k = \dim(\mathcal{E}_k) = \operatorname{rank}\{M^{(k)}(\mathcal{Y} \mid X)\}$.

Proposition 1 suggests a way to decompose the conditional mean $\mathbb{E}(\mathcal{Y} \mid X)$ into two parts: a part that contains all relevant information, plus an orthogonal part that is totally irrelevant. That is,

$$\mathbb{E}(\mathcal{Y} \mid X) = \mathbb{E}(\mathcal{Y} \times_{(k)} P_k \mid X) + \mathbb{E}(\mathcal{Y} \times_{(k)} Q_k \mid X) = \mathbb{E}(\mathcal{Y} \times_{(k)} P_k \mid X) + \mathbb{E}(\mathcal{Y} \times_{(k)} Q_k),$$

for $k = 1, \ldots, m$, where P_k is the projection onto $\mathcal{E}_k$. In other words, $\{\mathcal{E}_k\}_{k=1}^m$ fully captures all the changes in the mean function, and $\mathcal{M}(\mathcal{Y} \mid X)$ captures all necessary information regarding $\mathbb{E}(\mathcal{Y} \mid X)$. Combined with model (2.2), Proposition 1 implies that

$$\mathcal{E}_k = \text{span}(\beta_k), \qquad d_k = \dim\{\text{span}(\beta_k)\}.$$

We thus estimate $\{\text{span}(\beta_k)\}_{k=1}^m$ through $\mathcal{M}(\mathcal{Y} \mid X)$. Furthermore, Proposition 1 establishes the connection between $\mathcal{M}(\mathcal{Y} \mid X)$ and the individual columns of the mode-k matricization of $\mathcal{Y}$, based on which we develop our estimation and testing methods.

2.3 Subspace and dimension estimation

Given the data observations $\{(X_i, \mathcal{Y}_i)\}_{i=1}^n$, we obtain the sample estimate of $M^{(k)}(\mathcal{Y} \mid X)$,

$$\widehat{M}^{(k)}(\mathcal{Y} \mid X) = -\frac{1}{n^2} \sum_{i,i'} \left\{(\mathcal{Y}_i)_{(k)} - \overline{\mathcal{Y}}_{(k)}\right\} \left\{(\mathcal{Y}_{i'})_{(k)} - \overline{\mathcal{Y}}_{(k)}\right\}^{\mathrm{T}} \|X_i - X_{i'}\|, \tag{2.4}$$

where $\overline{\mathcal{Y}}_{(k)}$ is the sample mean of $\mathcal{Y}_{(k)}$, $k = 1, \ldots, m$. Let $\{\widehat{\lambda}_j^{(k)}\}_{j=1}^{r_k}$ denote the eigenvalues in the descending order, and $\{\widehat{\gamma}_j^{(k)}\}_{j=1}^{r_k}$ the corresponding eigenvectors of the matrix $\widehat{M}^{(k)}(\mathcal{Y} \mid X)$. We propose to estimate $\mathcal{E}_k = \text{span}(\beta_k)$ by the space spanned by $\{\widehat{\gamma}_j^{(k)}\}_{j=1}^{\widehat{d}_k}$, where $\widehat{d}_k$ is the estimated dimension that we discuss later. We note that the computation is fast, as it only involves matrix spectral decompositions.

Let $\{\lambda_j^{(k)}\}_{j=1}^{r_k}$ denote the eigenvalues in the descending order, and $\{\gamma_j^{(k)}\}_{j=1}^{r_k}$ the corresponding eigenvectors of $M^{(k)}(\mathcal{Y} \mid X)$. From Section 2.2, we have that $\lambda_1^{(k)} > \lambda_2^{(k)} > \ldots > \lambda_{d_k}^{(k)} > \lambda_{d_k+1}^{(k)} = \ldots = \lambda_{r_k}^{(k)} = 0$. The next two theorems justify our proposed estimator. The

first is derived under the setting where the dimension r_k is fixed, while the second is derived when r_k diverges and other dimensions $d_k,\ p$ are all fixed.

Theorem 1 *Suppose* $\mathbb{E}\left(\|X\|^2 + \|\mathcal{Y}\|_F^2\right) < \infty$ *and* $\mathbb{E}(\|X - \mu_X\|^2\|\mathcal{Y} - \mu\|_F^2) < \infty$, *where* $\mu_X = \mathbb{E}(X), \mu = \mathbb{E}(\mathcal{Y})$. *Suppose* r_k *is fixed,* $k = 1, \ldots, m$. *Then,*

$$\left\|\widehat{\gamma}_j^{(k)} - \gamma_j^{(k)}\right\| = O_p(n^{-1/2}), \quad j = 1, \ldots, d_k,\ k = 1, \ldots, m.$$

Theorem 2 *Suppose* $\mathbb{E}\left(\|X\|^2 + \|\mathcal{Y}\|_F^2\right) < \infty$, $\mathbb{E}(\|X - \mu_X\|^2\|\mathcal{Y} - \mu\|_F^2) < \infty$, $\mathbb{E}(\|X - \mu_X\|\|\mathcal{Y} - \mu\|_F^2) < \infty$, *and* $r_k^2/n \to 0$ *as* $n \to \infty$ *for* $k = 1, \ldots, m$. *Then,*

$$\left\|\widehat{\gamma}_j^{(k)} - \gamma_j^{(k)}\right\| \to^p 0, \quad j = 1, \ldots, d_k,\ k = 1, \ldots, m.$$

Next, we propose to estimate the subspace dimension d_k using the approach of Zhu et al. (2020), i.e.,

$$\widehat{d}_k = \text{argmax}_{1 \le j \le r_k} \left\{ j : \frac{\widehat{S}_{j+1}^* + c_{2n}}{\widehat{S}_j^* + c_{2n}} \le \tau \right\}, \text{ where } \widehat{S}_j^* = \frac{\widehat{S}_j^2 + c_{1n}}{\widehat{S}_{j+1}^2 + c_{1n}} - 1,\ \widehat{S}_j = \frac{\widehat{\lambda}_j^{(k)}}{\widehat{\lambda}_j^{(k)} + 1}, \tag{2.5}$$

τ is a thresholding parameter, and $\widehat{d}_k = 0$ if $\frac{\widehat{S}_{j+1}^* + c_{2n}}{\widehat{S}_j^* + c_{2n}} > \tau$ for all $j = 1, \ldots, r_k$. The next theorem shows that the estimated dimension is consistent.

Theorem 3 *Suppose* $\mathbb{E}\left(\|X\|^2 + \|\mathcal{Y}\|_F^2\right) < \infty$, $\mathbb{E}(\|X - \mu_X\|^2\|\mathcal{Y} - \mu\|_F^2) < \infty$, *and* $c_{1n} \to 0$, $c_{2n} \to 0$, $c_{1n}c_{2n}n \to \infty$, *and* $0 < \tau < 1$. *Then,* $\mathbb{P}(\widehat{d}_k = d_k) \to 1$, $k = 1, \ldots, m$, *as* $n \to \infty$.

Following the recommendation of Zhu et al. (2020), we choose $c_{1n} = 0.1\log(n)/\sqrt{n}$, $c_{2n} = 0.2\log(n)/\sqrt{n}$, and $\tau = 0.8$. We also comment that, there are alternative ways to estimate the dimension d_k, e.g., Zhu et al. (2006), Luo et al. (2009), Xia et al. (2015).

2.4 Sparse subspace estimation

To further improve the interpretability of our dimension reduction method, we introduce sparsity on the elements of the basis $\{\mathrm{span}(\beta_k)\}_{k=1}^m$, and propose to apply the sparse principal component analysis (Zou et al. 2006) to obtain a sparse estimate. The algorithm iterates between two key steps. In the first step, given an estimate $\widehat{\alpha}_k \in \mathbb{R}^{r_k \times d_k}$, we seek $\widehat{\beta}_k \in \mathbb{R}^{r_k \times d_k}$, whose lth column $\widehat{\beta}_{kl}$ is obtained from

$$\widehat{\beta}_{kl} = \mathrm{argmin}_{\beta \in \mathbb{R}^{r_k}} (\widehat{\alpha}_{kl} - \beta)^{\mathrm{T}} \widehat{M}^{(k)}(\mathcal{Y} \mid X)(\widehat{\alpha}_{kl} - \beta) + \lambda \|\beta\|^2 + \lambda_1 \|\beta\|_1, \tag{2.6}$$

where $\widehat{\alpha}_{kl}$ is the lth column of $\widehat{\alpha}_k$, $l = 1, \ldots, d_k$, $\|\cdot\|_1$ is the L_1 norm, and λ, λ_1 are the tuning parameters. After obtaining $\widehat{\beta}_{kl}$, we normalize the vector. In the second step, given the estimate $\widehat{\beta}_k$, we update $\widehat{\alpha}_k = \widehat{U}_k \widehat{V}_k^{\mathrm{T}}$, where $\widehat{U}_k, \widehat{V}_k$ are the left and right singular vector matrices of $\widehat{M}^{(k)}(\mathcal{Y} \mid X)\widehat{\beta}_k$. We stop the algorithm when some convergence criterion is met, e.g., when the difference of two consecutive estimates is below a threshold, e.g., 10^{-3}. We remark that (2.6) is a direct modification of the sparse principal component analysis method of Zou et al. (2006), where we replace the sample covariance matrix of X in the objective function with the tensor MDD $\widehat{M}^{(k)}$. This way, it takes into account the mean dependence between $\mathcal{Y}$ and X, and also allows us to search for sparse $\{\beta_k\}_{k=1}^m$ under model (2.2). One may adopt other methods similar to sparse principal component analysis, e.g., Yuan & Zhang (2013). We choose Zou et al. (2006) for its simplicity and competitive empirical performance. For the tuning parameters, we follow a similar strategy as the im-

plementation of (Zou et al. 2006). That is, we choose the number of nonzero elements s using the BIC criterion of Sun & Li (2017), while we fix the ridge penalty $\lambda = 10^{-6}$. We further carry out a sensitivity analysis in Section B.2 of the Supplementary Materials.

3. Testing Mean Independence

In addition to dimension reduction estimation, another related but crucial question is to test if the mean of some linear dimension reduction of the response is invariant to the changes in the predictor. Toward that end, we study the mean independence testing problem. That is, for a given matrix $\alpha_k \in \mathbb{R}^{r_k \times q_k}$, $q_k < r_k$, we aim to test the null hypothesis, without imposing any parametric assumptions, that

$$H_0 : \mathbb{E}\left(\mathcal{Y} \times_{(k)} \alpha_k^{\mathrm{T}} \mid X\right) = \mathbb{E}\left(\mathcal{Y} \times_{(k)} \alpha_k^{\mathrm{T}}\right) \text{ almost surely}, \tag{3.7}$$

We test (3.7) again using the tensor MDD $\mathcal{M}(\mathcal{Y} \mid X)$, based on the following observation. When the null (3.7) holds, $\mathrm{trace}\left\{\alpha_k^{\mathrm{T}} M^{(k)}(\mathcal{Y} \mid X)\alpha_k\right\} = 0$, and when the null does not hold, $\mathrm{trace}\left\{\alpha_k^{\mathrm{T}} M^{(k)}(\mathcal{Y} \mid X)\alpha_k\right\} > 0$. This suggests the following test statistic,

$$T_n = n\mathrm{trace}\left\{\alpha_k^{\mathrm{T}} \widetilde{M}^{(k)}(\mathcal{Y} \mid X)\alpha_k\right\},$$

where $\widetilde{M}^{(k)} = \frac{1}{n(n-3)} \sum_{h \neq l} \widetilde{A}_{hl}\widetilde{B}_{hl}$, $\widetilde{A}_{hl} = a_{hl} - a_{\cdot l} - a_{h\cdot} + a_{\cdot\cdot}$, $a_{hl} = \|X_h - X_l\|$, $a_{\cdot l} = \frac{1}{(n-2)} \sum_{h=1}^{n} a_{hl}$, $a_{h\cdot} = \frac{1}{(n-2)} \sum_{l=1}^{n} a_{hl}$, $a_{\cdot\cdot} = \frac{1}{(n-1)(n-2)} \sum_{h \neq l} a_{hl}$, and $\widetilde{B}_{hl}$ is defined similarly as $\widetilde{A}_{hl}$ with $b_{hl} = \frac{1}{2}\left\{(\mathcal{Y}_h)_{(k)} - (\mathcal{Y}_l)_{(k)}\right\}\left\{(\mathcal{Y}_h)_{(k)} - (\mathcal{Y}_l)_{(k)}\right\}^{\mathrm{T}}$. Here, we use a different estimator $\widetilde{M}^{(k)}$ of $M^{(k)}$ than the estimator $\widehat{M}^{(k)}$ in (2.4). This is because $\widehat{M}^{(k)}$ is a biased es-

timator with an asymptotically negligible bias, but $\widetilde{M}^{(k)}$ is an unbiased estimator, which can be shown following Székely & Rizzo (2014). This is crucial for our subsequent bootstrap-based testing procedure. Relatedly, it is possible to use $\widetilde{M}^{(k)}$ for the subspace estimation in Section 2.3. Nevertheless, we choose $\widehat{M}^{(k)}$ for subspace estimation, mainly because it is positive semi-definite, which allows us to apply the thresholding double ridge ratio approach of Zhu et al. (2020) to estimate the dimensions $\{d_k\}_{k=1}^m$.

Theorem 4 *Suppose* $\mathbb{E}\left(\|X\|^2 + \|\mathcal{Y}\|_F^2\right) < \infty$, *and* $\mathbb{E}(\|X - \mu_X\|^2 \|\mathcal{Y} - \mu\|_F^2) < \infty$.

(i) When H_0 *holds, we have,*

$$T_n \xrightarrow{d} \sum_{l=1}^{\infty} \nu_l (G_l^2 - 1),$$

where $\{G_l\}_{l=1}^{+\infty}$ *are a sequence of independent standard normal variables,* $\{\nu_l\}_{l=1}^{+\infty}$, $\{\phi_l(x)\}_{l=1}^{+\infty}$ *are eigenvalues and eigenfunctions, such that* $J(z, z') = \sum_{l=1}^{\infty} \nu_l \phi_l(z) \phi_l(z')$, $z = \{x, y\}$ *is a sample from the joint distribution of* X *and* $\mathcal{Y}$, $J(z, z') = U(x, x')$ $V(y, y')$, $U(x, x') = \|x - x'\| + \mathbb{E}(\|X - X'\|) - \mathbb{E}(\|x - X'\|) - \mathbb{E}(\|X - x'\|)$, $V(y, y') = -\langle \alpha_k^{\mathrm{T}}(y_{(k)} - \mu_{(k)}), \alpha_k^{\mathrm{T}}(y'_{(k)} - \mu_{(k)}) \rangle_F$, $\langle \cdot, \cdot \rangle_F$ *is the Frobenius inner product of two tensors, and* ϕ_k *is an orthonormal sequence in that* $\mathbb{E}\{\phi_h(\mathcal{Z})\phi_l(\mathcal{Z})\} = \mathbb{I}\{h = l\}$, $\mathcal{Z} = \{X, \mathcal{Y}\}$ *is a random variable from the joint distribution of* X *and* $\mathcal{Y}$.

(ii) When H_0 *does not hold, we have,*

$$\sqrt{n}\left[n^{-1}T_n - \operatorname{trace}\left\{\alpha_k^{\mathrm{T}} M^{(k)}(\mathcal{Y} \mid X)\alpha_k\right\}\right] \xrightarrow{d} \operatorname{Normal}(0, 4\sigma^2),$$

where $\sigma^2 = \operatorname{var}(K(\mathcal{Z}))$, $K(z) = \mathbb{E}[U(x, X)V(y, \mathcal{Y})] \in \mathbb{R}$.

Noting that the limiting null distribution of T_n involves an infinite sum, we next develop a wild bootstrap procedure following Lee et al. (2020).

Step 1: Generate the bootstrap statistic, $T^*_{n,b} = \text{trace}\left\{\frac{1}{(n-3)}\sum_{h\neq l}\eta_h^{(b)}\left(\alpha_k^{\mathrm{T}}\widetilde{A}_{hl}\widetilde{B}_{hl}\alpha_k\right)\eta_l^{(b)}\right\}$, where $\eta_h^{(b)}, h = 1,\ldots,n$, are i.i.d. random variables with zero mean and unit variance, e.g., standard normal random variables.

Step 2: Repeat Step 1 for B times, collect $\{T^*_{n,b}\}_{b=1}^B$, and obtain the $(1-\alpha)$th quantile $Q^*_{(1-\alpha),n}$ of $\{T^*_{n,b}\}_{b=1}^B$ under the significance level α.

Step 3: Reject the null hypothesis H_0, if T_n is greater than the critical value $Q^*_{(1-\alpha),n}$.

The next theorem establishes the asymptotic validity of our bootstrap test, and shows that it provides a consistent approximation of the limiting null distribution of the test statistic. Recall the definition of the bootstrap consistency (Chang & Park 2003, Li et al. 2003); i.e., for a bootstrap statistic T^*_n that depends on the random samples $\{\mathcal{Z}_i\}_{i=1}^n$, $\mathcal{Z}_i = \{X_i, \mathcal{Y}_i\}$, we say T^*_n converges to T in distribution almost surely (a.s.), if T^*_n converges to T in distribution for almost every sequence $(\mathcal{Z}_1, \mathcal{Z}_2, \ldots)$. We denote it as $T^*_n \xrightarrow{d^*} T$ a.s.

Theorem 5 *Suppose* $\mathbb{E}(\|X\|^4 + \|\mathcal{Y}\|_F^8) < \infty$, $\mathbb{E}(\|X-\mu_X\|^4\|\mathcal{Y}-\mu\|_F^4) < \infty$, *and* $\mathbb{E}(\eta^4) < \infty$. *Then, under the null hypothesis* H_0, *we have,*

$$T^*_n \xrightarrow{d^*} \sum_{l=1}^{\infty}\nu_l(G_l^2 - 1) \text{ a.s.},$$

where $\{\nu_l, G_l\}_{l=1}^{+\infty}$ *are as defined in Theorem 4.*

We make a few remarks. First, our mean independence test can be applied sequentially to select the dimension d_k of the dimension reduction subspace $\mathcal{E}_k = \text{span}(\beta_k)$. Second, it is possible to extend the test to assess the mean independence between X and $\mathcal{Y}$ by simultaneously considering more than one fixed α_k. That is, for a set of $\alpha_{k_j} \in \mathbb{R}^{r_{k_j} \times q_{k_j}}$, $H_0 : \mathbb{E}(\mathcal{Y} \times_{(k_1)} \alpha_{k_1}^{\mathrm{T}} \times_{(k_2)} \ldots \times_{(k_w)} \alpha_{k_w}^{\mathrm{T}} \mid X) = \mathbb{E}(\mathcal{Y} \times_{(k_1)} \alpha_{k_1}^{\mathrm{T}} \times_{(k_2)} \ldots \times_{(k_w)} \alpha_{k_w}^{\mathrm{T}})$ a.s, for $j = 1, \ldots, w, w \leq m, q_{k_j} < r_{k_j}$. Third, our test is built upon and extends Park et al. (2015), Lee et al. (2020) from vector-valued data to tensor data. This extension, however, is highly nontrivial. Park et al. (2015) focused on the partial mean independence test between two vectors controlling for the third vector and developed a permutation test, while we assess the mean independence between a tensor and a vector and employ bootstrap. Lee et al. (2020) studied the mean independence for functional data and used a metric defined specifically for the functional data, whereas we develop a new tensor MDD metric designed for the tensor data. Besides, our theoretical analysis is utterly new. Considering that there is a relative paucity of testing methods for tensor data, we view our test a useful addition to the toolbox of tensor data inference.

4. Simulations

4.1 Non-sparse coefficient tensor

We first examine the finite-sample performance of our dimension reduction method when the coefficient tensor $\mathcal{B}$ is non-sparse. We consider three models,

$$\begin{aligned}
&\text{Linear model:} && \mathcal{Y} = \mathcal{B} \times_{(m+1)} X + 0.1\varepsilon, \\
&\text{Nonlinear model I:} && \mathcal{Y} = \mathcal{B} \times_{(m+1)} \exp|X| + 0.1\varepsilon, \\
&\text{Nonlinear model II:} && \mathcal{Y} = \mathcal{B} \times_{(m+1)} X^2 + 0.1\varepsilon,
\end{aligned}$$

where $m = 2$, $\mathcal{B} = [\![\Theta; \beta_1, \beta_2, I_p]\!]$, the entries of $\Theta \in \mathbb{R}^{d_1 \times d_2 \times p}$ are randomly generated from Uniform$(0, 1)$, $\beta_1 \in \mathbb{R}^{r_1 \times d_1}$ and $\beta_2 \in \mathbb{R}^{r_2 \times d_2}$ are randomly generated from Uniform$(-1, 1)$ and orthogonalized. The predictors $X \in \mathbb{R}^p$ are generated from a standard normal distribution, and the errors $\text{vec}(\varepsilon)$ are generated from Normal$(0, I_{r_2} \circ I_{r_1})$. We set $r_1 = r_2 = 100$, $(d_1, d_2) = (5, 5)$, $p = 5$, and vary the sample size $n = \{10, 50, 100\}$.

We apply the proposed method, where we set the reduced dimension at the truth first, then study the dimension estimation by (2.5) later. We also compare with two alternative tensor response regression methods, the tensor envelope method of Li & Zhang (2017), and the sparse tensor response regression method of Sun & Li (2017), both of which have assumed the tensor response linear model. We evaluate the dimension reduction estimation accuracy by $\mathcal{D}(\beta_k, \widehat{\beta}_k) = \|P_k - \widehat{P}_k\|_F = \|\gamma^{(k)}(\gamma^{(k)})^{\mathrm{T}} - \widehat{\gamma}^{(k)}(\widehat{\gamma}^{(k)})^{\mathrm{T}}\|_F$, $\gamma^{(k)} = (\gamma_1^{(k)}, \ldots, \gamma_{d_k}^{(k)})$, $\widehat{\gamma}^{(k)} = (\widehat{\gamma}_1^{(k)}, \ldots, \widehat{\gamma}_{d_k}^{(k)})$, where a smaller $\mathcal{D}$ indicates a more accurate result.

Table 1 reports the results based on 1000 data replications. For the linear model, the tensor envelope method performs slightly better than our method when $n = 50$ and 100, which is not surprising since the data indeed follows their assumed model. However, it is interesting to see that our method is superior when the sample size is really small $n = 10$. For the nonlinear models, our method consistently outperforms the two alternative solutions

Table 1: Dimension reduction estimation. Reported are the average and standard deviation of $\mathcal{D}(\beta_k, \widehat{\beta}_k)$ based on 1000 replications. Three methods are compared: our tensor martingale difference divergence method (TMDDM), Li & Zhang (2017, T-Envelope), and Sun & Li (2017, STORE).

			TMDDM	T-Envelope	STORE
Linear model	$\mathcal{D}(\beta_1, \widehat{\beta}_1)$	$n = 10$	0.974 (0.222)	1.978 (0.294)	1.735 (0.295)
		$n = 50$	0.335 (0.037)	0.333 (0.038)	1.127 (0.340)
		$n = 100$	0.229 (0.022)	0.226 (0.022)	0.904 (0.375)
	$\mathcal{D}(\beta_2, \widehat{\beta}_2)$	$n = 10$	0.975 (0.217)	1.967 (0.300)	1.729 (0.304)
		$n = 50$	0.335 (0.039)	0.332 (0.038)	1.108 (0.348)
		$n = 100$	0.228 (0.022)	0.225 (0.022)	0.875 (0.371)
Nonlinear model I	$\mathcal{D}(\beta_1, \widehat{\beta}_1)$	$n = 10$	0.513 (0.217)	3.022 (0.280)	1.346 (0.438)
		$n = 50$	0.199 (0.047)	2.595 (0.663)	1.096 (0.420)
		$n = 100$	0.143 (0.027)	2.847 (0.482)	1.066 (0.426)
	$\mathcal{D}(\beta_2, \widehat{\beta}_2)$	$n = 10$	0.516 (0.226)	2.994 (0.305)	1.357 (0.428)
		$n = 50$	0.198 (0.047)	1.955 (0.878)	1.106 (0.426)
		$n = 100$	0.143 (0.027)	1.575 (1.018)	1.046 (0.408)
Nonlinear model II	$\mathcal{D}(\beta_1, \widehat{\beta}_1)$	$n = 10$	0.912 (0.345)	3.056 (0.199)	1.619 (0.364)
		$n = 50$	0.370 (0.079)	2.588 (0.610)	1.418 (0.383)
		$n = 100$	0.266 (0.044)	2.646 (0.579)	1.382 (0.389)
	$\mathcal{D}(\beta_2, \widehat{\beta}_2)$	$n = 10$	0.913 (0.348)	3.045 (0.201)	1.626 (0.352)
		$n = 50$	0.369 (0.078)	2.315 (0.706)	1.422 (0.378)
		$n = 100$	0.265 (0.044)	2.045 (0.849)	1.390 (0.389)

across all sample sizes. This example illustrates the advantage of our method when there is no clear indication that the data actually follows a linear model.

Table 2: Dimension selection. Reported is the percentage of times of that the dimension is correctly selected, under-selected, and over-selected, out of 1000 data replications.

		$\widehat{d}_1 < d_1$	$\widehat{d}_1 = d_1$	$\widehat{d}_1 > d_1$	$\widehat{d}_2 < d_2$	$\widehat{d}_2 = d_2$	$\widehat{d}_2 > d_2$
Linear model	$n = 10$	10.1	77.1	12.8	9.8	77.1	13.1
	$n = 50$	0.0	100.0	0.0	0.0	100.0	0.0
	$n = 100$	0.0	100.0	0.0	0.0	100.0	0.0
		$\widehat{d}_1 < d_1$	$\widehat{d}_1 = d_1$	$\widehat{d}_1 > d_1$	$\widehat{d}_2 < d_2$	$\widehat{d}_2 = d_2$	$\widehat{d}_2 > d_2$
Nonlinear model I	$n = 10$	1.0	85.4	13.6	1.3	83.8	14.9
	$n = 50$	0.0	100.0	0.0	0.0	100.0	0.0
	$n = 100$	0.0	100.0	0.0	0.0	100.0	0.0
		$\widehat{d}_1 < d_1$	$\widehat{d}_1 = d_1$	$\widehat{d}_1 > d_1$	$\widehat{d}_2 < d_2$	$\widehat{d}_2 = d_2$	$\widehat{d}_2 > d_2$
Nonlinear model II	$n = 10$	15.3	71.8	12.9	14.0	71.4	14.6
	$n = 50$	0.2	99.8	0.0	0.0	100.0	0.0
	$n = 100$	0.0	100.0	0.0	0.0	100.0	0.0

We next examine the performance of dimension selection via (2.5). Table 2 reports the percentage of times of that the dimension is correctly selected, under-selected, and over-selected, out of 1000 data replications. We see that (2.5) selects the true dimension fairly accurately, especially when the sample size is reasonably large.

4.2 Sparse coefficient tensor

We next examine the finite-sample performance when the coefficient tensor $\mathcal{B}$ is sparse. We consider the same models as in Section 4.1, plus another nonlinear model with $m = 3$,

Nonlinear model III: $\mathcal{Y} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ is generated from nonlinear model I, but with $\mathcal{B}$ following a CP decomposition.

For the first three models with $m = 2$, we set $r_1 = r_2 = 100$, and generate $\mathcal{B} = [\![\Theta; \beta_1, \beta_2, I_p]\!]$, the entries of $\Theta \in \mathbb{R}^{d_1 \times d_2 \times p}$ are randomly generated from Uniform$(0, 1)$, $\beta_k = (b_k, 0_{d_k \times r_k/2})^{\mathrm{T}}$, $b_k \in \mathbb{R}^{d_k \times r_k/2}$, $k = 1, 2$, are generated from Uniform$(-1, 1)$ and orthogonalized. Correspondingly, there are $r_k/2$ nonzero elements in b_k, $k = 1, 2$. For the last model with $m = 3$, we set $r_1 = r_2 = r_3 = 50$, and generate $\mathcal{B} = \sum_{k=1}^{5} w_k \beta_{1,k} \circ \beta_{2,k} \circ \ldots \circ \beta_{4,k}$, where $\beta_k = (b_k, 0_{5 \times r_k/2})^{\mathrm{T}}$, $b_k \in \mathbb{R}^{5 \times r_k/2}$, $k = 1, 2, 3$, are generated from a standard normal distribution and orthogonalized, β_4 is a $p \times 5$ matrix with all entries equal to one, and $(w_1, \ldots, w_5)$ are generated from a standard normal distribution. We also note that, for this last model, although B follows a CP decomposition, the β's are semi-orthogonal, so it is still of a Tucker form. The semi-orthogonality is to avoid the CP degeneracy; see Zhou et al. (2021) for more discussion.

We apply the proposed sparse version of our tensor MDD method, where we first fix $\lambda = 10^{-6}$ and $s = 50$ in this study. We again compare with the tensor envelope method of Li & Zhang (2017), and the sparse tensor response regression method of Sun & Li (2017). In addition to the estimation accuracy, we select s adaptively using the BIC criterion of Sun & Li (2017) and evaluate the selection accuracy by the true and false positive rates,

$$\text{TPR}_k = \frac{\sum_{l=1}^{d_k} \sum_{j=1}^{r_k} \mathbf{1}\{\beta_{klj} \neq 0, \widehat{\beta}_{klj} \neq 0\}}{\sum_{l=1}^{d_k} \sum_{j=1}^{r_k} \mathbf{1}\{\beta_{klj} \neq 0\}}, \quad \text{FPR}_k = \frac{\sum_{l=1}^{d_k} \sum_{j=1}^{r_k} \mathbf{1}\{\beta_{klj} = 0, \widehat{\beta}_{klj} \neq 0\}}{\sum_{l=1}^{d_k} \sum_{j=1}^{r_k} \mathbf{1}\{\beta_{klj} = 0\}},$$

where β_{klj} and $\widehat{\beta}_{klj}$ are the (l, j)th element of β_k and $\widehat{\beta}_k$, respectively. We also note that,

although $\{\beta_k\}_{k=1}^m$ are not identifiable in terms of nonsingular transformation, the rows in which all elements are zero would remain all zero after the nonsingular transformation. This allows us to evaluate the selection accuracy.

Table 3 reports the estimation results and Table 4 reports the selection results, based on 1000 replications. We see that, compared to the alternative solutions, our method produces a more accurate dimension reduction estimation with a smaller $\mathcal{D}$ measure. For the selection accuracy, our approach also produces an accurate selection with a high TPR and a low FPR.

4.3 Mean independence test

Finally, we study the finite-sample performance of our proposed mean independence test. We consider the models in Section 4.1, and set $r_1 = r_2 = \{50, 100\}$, and $n = \{10, 50, 100\}$. We test three different null hypotheses,

$$\begin{aligned} H_{0,1}: &\quad \mathbb{E}\left(\mathcal{Y} \times_{(1)} \alpha_1^{\mathrm{T}} \mid X\right) = \mathbb{E}\left(\mathcal{Y} \times_{(1)} \alpha_1^{\mathrm{T}}\right); \\ H_{0,2}: &\quad \mathbb{E}\left(\mathcal{Y} \times_{(2)} \alpha_2^{\mathrm{T}} \mid X\right) = \mathbb{E}\left(\mathcal{Y} \times_{(2)} \alpha_2^{\mathrm{T}}\right); \\ H_{0,3}: &\quad \mathbb{E}\left(\mathcal{Y} \times_{(1)} \alpha_1^{\mathrm{T}} \times_{(2)} \alpha_2^{\mathrm{T}} \mid X\right) = \mathbb{E}\left(\mathcal{Y} \times_{(1)} \alpha_1^{\mathrm{T}} \times_{(2)} \alpha_2^{\mathrm{T}}\right), \end{aligned} \tag{4.8}$$

where, for the null, we set $\alpha_1 = \beta_{1,0}$ and $\alpha_2 = \beta_{2,0}$, with $\beta_{1,0} \in \mathbb{R}^{r_1 \times (r_1 - d_1)}$ and $\beta_{2,0} \in \mathbb{R}^{r_2 \times (r_2 - d_2)}$ being the matrices orthogonal to β_1 and β_2, respectively, and for the alternative, we set $(q_1, q_2) = (3, 3)$, $\alpha_1 = (a_1, 0_{q_1 \times r_1/2})^{\mathrm{T}} \in \mathbb{R}^{r_1 \times q_1}$, and $\alpha_2 = (a_2, 0_{q_2 \times r_2/2})^{\mathrm{T}} \in \mathbb{R}^{r_2 \times q_2}$, with $a_1 \in \mathbb{R}^{q_1 \times r_1/2}$ and $a_2 \in \mathbb{R}^{q_2 \times r_2/2}$ being randomly generated from Uniform$(-1, 1)$ and orthogonalized. We set the nominal level at 5%. We set the bootstrap sample size as

Table 3: Sparse dimension reduction estimation. Reported are the average and standard deviation of $\mathcal{D}(\beta_k, \widehat{\beta}_k)$ based on 1000 data replications. Three methods are compared: our sparse tensor martingale difference divergence method (s-TMDDM), Li & Zhang (2017, T-Envelope), and Sun & Li (2017, STORE).

			s-TMDDM	T-Envelope	STORE
Linear model	$\mathcal{D}(\beta_1, \widehat{\beta}_1)$	$n = 10$	0.902 (0.214)	1.969 (0.292)	1.866 (0.372)
		$n = 50$	0.332 (0.037)	0.332 (0.038)	1.118 (0.428)
		$n = 100$	0.236 (0.024)	0.226 (0.022)	0.836 (0.405)
	$\mathcal{D}(\beta_2, \widehat{\beta}_2)$	$n = 10$	0.909 (0.213)	1.968 (0.299)	1.850 (0.371)
		$n = 50$	0.334 (0.037)	0.333 (0.038)	1.073 (0.426)
		$n = 100$	0.235 (0.024)	0.225 (0.022)	0.800 (0.400)
Nonlinear model I	$\mathcal{D}(\beta_1, \widehat{\beta}_1)$	$n = 10$	0.495 (0.198)	3.013 (0.301)	1.859 (0.609)
		$n = 50$	0.209 (0.045)	2.597 (0.665)	1.124 (0.471)
		$n = 100$	0.155 (0.028)	2.846 (0.489)	1.004 (0.434)
	$\mathcal{D}(\beta_2, \widehat{\beta}_2)$	$n = 10$	0.500 (0.205)	2.997 (0.301)	1.830 (0.600)
		$n = 50$	0.210 (0.046)	1.959 (0.876)	1.102 (0.454)
		$n = 100$	0.156 (0.028)	1.588 (1.008)	0.989 (0.430)
Nonlinear model II	$\mathcal{D}(\beta_1, \widehat{\beta}_1)$	$n = 10$	0.854 (0.333)	3.052 (0.200)	1.995 (0.523)
		$n = 50$	0.358 (0.068)	2.605 (0.598)	1.430 (0.442)
		$n = 100$	0.268 (0.041)	2.648 (0.582)	1.346 (0.448)
	$\mathcal{D}(\beta_2, \widehat{\beta}_2)$	$n = 10$	0.868 (0.347)	3.044 (0.203)	1.986 (0.513)
		$n = 50$	0.361 (0.068)	2.317 (0.691)	1.418 (0.427)
		$n = 100$	0.271 (0.041)	2.039 (0.843)	1.342 (0.433)
Nonlinear model III	$\mathcal{D}(\beta_1, \widehat{\beta}_1)$	$n = 10$	1.284 (0.681)	2.301 (0.818)	2.093 (0.546)
		$n = 50$	0.836 (0.634)	1.793 (0.987)	1.836 (0.657)
		$n = 100$	0.688 (0.609)	1.687 (1.003)	1.776 (0.673)
	$\mathcal{D}(\beta_2, \widehat{\beta}_2)$	$n = 10$	1.273 (0.680)	2.300 (0.820)	2.099 (0.544)
		$n = 50$	0.827 (0.633)	1.816 (0.997)	1.851 (0.661)
		$n = 100$	0.684 (0.608)	1.734 (1.003)	1.786 (0.678)
	$\mathcal{D}(\beta_3, \widehat{\beta}_3)$	$n = 10$	1.284 (0.681)	2.300 (0.822)	2.124 (0.562)
		$n = 50$	0.828 (0.631)	1.799 (0.994)	1.866 (0.674)
		$n = 100$	0.687 (0.609)	1.682 (1.026)	1.794 (0.685)

Table 4: Sparse dimension reduction selection. Reported are the average of the true positive rate (TPR), and false positive rate (FPR), based on 1000 data replications. Two methods are compared: our sparse tensor martingale difference divergence method (s-TMDDM) and Sun & Li (2017, STORE).

		s-TMDDM		STORE			s-TMDDM		STORE	
		TPR	FPR	TPR	FPR		TPR	FPR	TPR	FPR
Linear model	$n = 10$	0.685	0.122	0.561	0.073	Nonlinear model II	0.752	0.151	0.671	0.047
	$n = 50$	0.913	0.087	0.660	0.060		0.897	0.103	0.620	0.029
	$n = 100$	0.939	0.061	0.740	0.054		0.927	0.073	0.629	0.029
	$n = 10$	0.682	0.105	0.563	0.071		0.749	0.134	0.671	0.046
	$n = 50$	0.916	0.083	0.662	0.058		0.904	0.096	0.621	0.028
	$n = 100$	0.940	0.060	0.743	0.051		0.932	0.068	0.629	0.028
Nonlinear model I	$n = 10$	0.866	0.117	0.695	0.036	Nonlinear model III	0.677	0.251	0.832	0.093
	$n = 50$	0.945	0.055	0.640	0.034		0.816	0.195	0.826	0.102
	$n = 100$	0.959	0.041	0.642	0.025		0.854	0.163	0.845	0.113
	$n = 10$	0.870	0.110	0.695	0.036		0.690	0.253	0.838	0.087
	$n = 50$	0.948	0.052	0.641	0.034		0.830	0.196	0.832	0.096
	$n = 100$	0.960	0.041	0.642	0.025		0.867	0.164	0.852	0.106
							0.696	0.254	0.832	0.092
							0.829	0.192	0.827	0.101
							0.863	0.159	0.846	0.112

$B = 499$, and the external variables $\eta_j = -(\sqrt{5}-1)/2$, with probability $(\sqrt{5}+1)/(2\times\sqrt{5})$, and $\eta_j = (\sqrt{5}+1)/2$, with probability $1-(\sqrt{5}+1)/(2\times\sqrt{5})$, following Mammen (1993).

Table 5 summarizes the empirical size and power of our test based on 1000 data replications. We see that, as the sample size n increases, the empirical size of the test is close to the nominal level, and the empirical power is high, demonstrating the efficacy of the test.

Table 5: Empirical size and power of the mean independence test.

Size		$H_{0,1}$			$H_{0,2}$			$H_{0,3}$		
		$n=10$	$n=50$	$n=100$	$n=10$	$n=50$	$n=100$	$n=10$	$n=50$	$n=100$
	$r_1=r_2=50$	9.9	7.3	4.2	9.3	7.0	5.3	9.3	7.5	5.0
	$r_1=r_2=100$	8.5	5.7	5.8	8.5	6.0	5.7	8.1	5.3	5.1
Power		$H_{0,1}$			$H_{0,2}$			$H_{0,3}$		
		$n=10$	$n=50$	$n=100$	$n=10$	$n=50$	$n=100$	$n=10$	$n=50$	$n=100$
Linear model	$r_1=r_2=50$	91.8	100.0	100.0	95.5	100.0	100.0	78.9	100.0	100.0
	$r_1=r_2=100$	95.0	100.0	100.0	95.3	100	100	42.7	99.8	100.0
Nonlinear model I	$r_1=r_2=50$	20.1	81.8	99.7	19.3	81.1	99.8	17.0	78.2	99.7
	$r_1=r_2=100$	19.9	81.2	99.7	17.8	75.8	99.5	12.8	54.4	94.4
Nonlinear model II	$r_1=r_2=50$	20.3	89.1	100.0	19.8	89.8	100.0	17.8	83.3	100.0
	$r_1=r_2=100$	20.3	88.4	100.0	17.9	82.4	100.0	11.6	43.0	88.5

5. Data Applications

5.1 OECD data

We first illustrate our method with a macroeconomic study from the Organization for Economic Co-operation and Development (OECD). It is important to understand how the macroeconomic indices of foreign countries interact with the economic indices of the United States to produce accurate and meaningful forecasting. We analyze the transformed quarterly data (Chen et al. 2020, Liu & Chen 2019) with $n = 105$ observations. We choose 8 macroeconomic indices of the US as the predictor vector, and 8 macroeconomic indices of 13 countries as the response matrix, which results in $X \in \mathbb{R}^8$ and $\mathcal{Y} \in \mathbb{R}^{13\times 8}$. Since the response

has a moderate dimension, we apply the non-sparse version of our tensor MDD method.

First, we access the dimension of mean dimension reduction. We apply the dimension estimation criterion (2.5) to the entire data, which yields $\widehat{d}_1 = 1, \widehat{d}_2 = 2$. We then further carry out the mean independence testing. Toward that end, we randomly divide the data into a training set of 85 data observations and a testing set of the remaining 20 observations. We obtain the estimates $\widehat{\beta}_1, \widehat{\beta}_2$ from the training samples, and based on those estimates, we further test the three null hypotheses in (4.8), with $\alpha_1 = \widehat{\beta}_{1,0} \in \mathbb{R}^{13\times 12}, \alpha_2 = \widehat{\beta}_{2,0} \in \mathbb{R}^{8\times 6}$ using the testing samples. Table 6, top section, reports the testing results. We see that, the test rejects the null $H_{0,2}$ when $d_2 = 2$. We thus further test $H_{0,2}$ with $d_2 = 3, 4$ and $\alpha_2 = \widehat{\beta}_{2,0} \in \mathbb{R}^{8\times(8-d_2)}$, where the test rejects $H_{0,2}$ with $d_2 = 3$, but not $d_2 = 4$, which in turn suggests that $d_2 = 4$ seems sufficient for the OECD data. We also remark that, the difference between the dimension selected by (2.5) and the mean independence test is likely due to the sample splitting.

Next, we investigate the prediction performance based on $(d_1, d_2) = (1, 2), (1, 3), (1, 4)$. We also compare with the tensor envelope method of Li & Zhang (2017), and the method of Sun & Li (2017). Recall that Sun & Li (2017) only considered a special version of the Tucker decomposition, i.e., the CP decomposition, which in effect requires $d_1 = d_2$. As such, we set $d_1 = d_2 = 1$ for Sun & Li (2017). We again randomly divide the data into a training set of 85 observations and a testing set of 20 observations. We first obtain the estimated $\widehat{\beta}_1, \widehat{\beta}_2$ based on the training data by the three methods, respectively. We then build

Table 6: Mean independence test and prediction accuracy for the OECD data.

$d_1 = 1, d_2 = 2$			$d_1 = 1, d_2 = 3$		$d_1 = 1, d_2 = 4$	
$H_{0,1}$	$H_{0,2}$	$H_{0,3}$	$H_{0,2}$	$H_{0,3}$	$H_{0,2}$	$H_{0,3}$
0.417	0.018	0.231	0.018	0.096	0.962	0.980

	TMDDM	Tensor Envelope	STORE
$d_1 = 1, d_2 = 2$	0.0761 (0.0049)	0.1020 (0.0049)	0.0829 (0.0049)
$d_1 = 1, d_2 = 3$	0.0760 (0.0049)	0.1021 (0.0049)	0.0829 (0.0049)
$d_1 = 1, d_2 = 4$	0.0760 (0.0049)	0.1022 (0.0049)	0.0829 (0.0049)

a nonparametric regression estimator for the reduced-dimensional response $[\![\mathcal{Y}; \widehat{\beta}_1^{\mathrm{T}}, \widehat{\beta}_2^{\mathrm{T}}]\!] \in \mathbb{R}^{\widehat{d}_1 \times \widehat{d}_2}$ using a Gaussian smoothing kernel, i.e., $\sum_i K_h(x - x_i)[\![\mathcal{Y}_i; \widehat{\beta}_1^{\mathrm{T}}, \widehat{\beta}_2^{\mathrm{T}}]\!] / \sum_i K_h(x - x_i)$, where K_h is a Gaussian kernel, and h is the bandwidth chosen by the Akaike information criterion. We transform the predicted response into the original scale, by multiplying $\widehat{\beta}_1, \widehat{\beta}_2$ to the prediction of $[\![\mathcal{Y}; \widehat{\beta}_1^{\mathrm{T}}, \widehat{\beta}_2^{\mathrm{T}}]\!]$ to compute $\widehat{\mathcal{Y}}$. We evaluate the prediction accuracy by the mean squared prediction error, $\mathrm{MSPE} = (n_{test} r_1 r_2)^{-1} \sum_{i=1}^{n_{test}} \|\mathcal{Y}_i - \widehat{\mathcal{Y}}_i\|_F^2$ based on the testing data. We repeat this process 100 times, and report the average results. Table 6, bottom section, reports the average and standard error of MSPE. We see that our method produces a much more accurate prediction, indicating that the underlying regression relation is likely nonlinear. We note that the improvement in prediction accuracy of our method compared to the two alternative solutions is mainly due to the proposed tensor MDD method, because after dimension reduction, we apply the same nonparametric regression procedure for the

reduced-dimensional response.

5.2 Bike sharing data

We next illustrate our method with an e-commerce data from the capital bike share website. In recent years, bikes are emerging as an encouraging transportation due to traffic, environmental and health concerns, and the bike sharing system is being established. It is useful to examine how the demand for bike rental is affected by the weather condition, so the system can predict the demand more accurately and avoid potentially bike rental shortage (Subbaswamy et al. 2019). We analyze the data from $n = 24$ months. We choose the monthly averages of normalized temperature and windspeed as the predictor vector, and the seven-day average of the number of casual and registered customers as the response matrix, which results in $X \in \mathbb{R}^2$ and $\mathcal{Y} \in \mathbb{R}^{7\times 2}$. The averages cannot completely remove the temporal dependency but can alleviate it. Besides, even though the dimension of the response matrix is only 7×2, the sample size is also limited with $n = 24$, and thus dimension reduction can still be helpful. We again apply the non-sparse version of our tensor MDD method.

We follow the same procedures as in Section 5.1 to first access the dimension of mean dimension reduction, then the prediction performance. We apply the dimension estimation criterion (2.5) to the entire data, which yields $\widehat{d}_1 = 5, \widehat{d}_2 = 1$. We then carry out the mean independence testing, with a randomly split of a training set of 19 data observations and a testing set of the remaining 5 observations. Table 7, top section, shows that the selected

dimensions are appropriate for this data. We compute the proportions explained by the first five components for mode 1 and the first component for mode 2, and both account for over 97% of the mean dependence between $\mathcal{Y}$ and X, suggesting that they capture most of the regression mean information. In addition, for mode 2, the estimate of β_2 is $(0.42, 0.91)^{\mathrm{T}}$, which indicates that this component is a weighted average of the number of casual and registered users, with more weight loaded to the registered customers. For mode 1, the estimate of the first two columns of β_1 is

$$\begin{pmatrix} 0.38 & 0.36 & 0.36 & 0.41 & 0.39 & 0.36 & 0.38 \\ -0.56 & 0.09 & 0.28 & 0.36 & 0.29 & 0.14 & -0.60 \end{pmatrix},$$

where the seven columns above correspond to the bike usage from Sunday to Saturday. As such, the first component is a weighted average of the users across seven days of the week, and the second component corresponds to the difference between the weekday usage and the weekend usage. We next compare the prediction performance in terms of MSPE. Table 7, bottom section, reports the average and standard error of MSPE. Again we see that our method yields a smaller MSPE and a more accurate forecasting.

6. Conclusion

We conclude this article with a discussion. First of all, it is generally beneficial to estimate $f(X)$ after the dimension reduction on tensor response is achieved. One possible solution is to fit a nonparametric regression with a reduced-dimensional response, as shown in our real

Table 7: Mean independence test and prediction accuracy for the bike-sharing data.

$d_1 = 5$	$d_2 = 1$	$d_1 = 5, d_2 = 1$
$H_{0,1}$	$H_{0,2}$	$H_{0,3}$
0.944	0.932	0.876
TMDDM	Tensor Envelope	STORE
214.698 (7.844)	251.739 (16.329)	220.253 (7.747)

data applications. The dimension reduction step is expected to help improve the estimation accuracy of $f(X)$. Moreover, it is possible to utilize our test statistic to analyze the form of $f(X)$, by computing and comparing $\mathrm{trace}\left\{M^{(k)}(\mathcal{Y} \mid f(X))\right\}$ for different forms of $f(X)$ under any mode k. Specifically, consider a set of candidate functions, $\{f_i(X)\}_{i=1}^q$, with the unit variance so there is no scale difference for different functions. Recognizing that the tensor MDD measures the mean dependence, we choose the function $f_i(X)$ that produces the largest $\mathrm{trace}\left\{M^{(k)}(\mathcal{Y} \mid f_i(X))\right\}$. Moreover, we can test the hypothesis, $H_0 : \mathbb{E}(\varepsilon \mid f(X)) = \mathbb{E}(\varepsilon)\ a.s$, where ε is the error term that is mean independent of $f(X)$ under the true model. We can construct a test statistic of the form, $\widetilde{T}_n = n\mathrm{trace}\left\{\widetilde{M}^{(k)}\left(\widehat{\varepsilon} \mid f(X)\right)\right\}$, where $\widehat{\varepsilon} = \mathcal{Y} - \widehat{\mathcal{B}} \times_{(m+1)} f(X)$ is the estimated residual. Deriving the limiting distribution of $\widetilde{T}_n$ is important but highly nontrivial. We report some preliminary simulation results about choosing $f(X)$ using our test statistic in Section B.1 of the Supplementary Materials.

Next, we discuss a number of potential extensions. First, it is possible to extend our method to tensor-on-tensor regression. This requires an extension of the metric that mea-

sures the dependence between two tensors, which can be achieved by replacing the Euclidean distance between two vectors with the Frobenius norm of the difference between two tensors. It also requires an updated objective function, along with an the estimation procedure and its theoretical investigation. Second, while we have established the asymptotic properties of our dimension reduction estimator, it is of equal interest to derive the properties of the sparse estimator. Third, there has been recent development on the post dimension reduction inference (Kim et al. 2020). Along this line, it would be interesting to extend the bootstrap test to perform diagnostics check, where the estimation effect from $\{\widehat{\beta}_{k,0}\}_{k=1}^{m}$ is expected to affect the corresponding limiting distribution. Finally, we have focused on the i.i.d. situation, while it would be useful to extend our dimension reduction approach to the tensor time series data (Wang et al. 2019, Chen et al. 2022). Research along some of these directions are underway.

Supplementary Materials

Supplementary materials available online include technical proofs of all the theoretical results and additional numerical results.

Acknowledgements

The authors thank the Editor, the Associate Editor, and two referees for their constructive comments and suggestions. Dr. Li's research is partially supported by NSF grant CIF-

2102227 and NIH grants R01AG061303 and R01AG062542. Dr. Zhang's research is supported partly by NSF grants DMS-2053697 and DMS-2113590, and NIH grant R03DE030509. Dr. Lee's research is supported by Eugene M. Lang grant.

References

Bi, X., Tang, X., Yuan, Y., Zhang, Y. & Qu, A. (2021), 'Tensor in statistics', *Annual Review of Statistics and Its Application* **8**, 345–368.

Chang, Y. & Park, J. Y. (2003), 'A sieve bootstrap for the test of a unit root', *Journal of Time Series Analysis* **24**(4), 379–400.

Chen, E. Y., Tsay, R. S. & Chen, R. (2020), 'Constrained factor models for high-dimensional matrix-variate time series', *Journal of the American Statistical Association* **115**(530), 775–793.

Chen, H., Raskutti, G. & Yuan, M. (2019), 'Non-convex projected gradient descent for generalized low-rank tensor regression', *The Journal of Machine Learning Research* **20**(1), 172–208.

Chen, R., Yang, D. & Zhang, C.-H. (2022), 'Factor models for high-dimensional tensor time series', *Journal of the American Statistical Association* **117**(537), 94–116.

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

Hao, B., Wang, B., Wang, P., Zhang, J., Yang, J. & Sun, W. W. (2019), 'Sparse tensor additive regression', *arXiv preprint arXiv:1904.00479* .

Kim, K., Li, B., Yu, Z., Li, L. et al. (2020), 'On post dimension reduction statistical inference', *Annals of Statistics* **48**(3), 1567–1592.

Kolda, T. G. & Bader, B. W. (2009), 'Tensor decompositions and applications', *SIAM review* **51**(3), 455–500.

Lee, C. E. & Shao, X. (2018), 'Martingale difference divergence matrix and its application to dimension reduction for stationary multivariate time series', *Journal of the American Statistical Association* **113**(521), 216–229.

Lee, C., Zhang, X. & Shao, X. (2020), 'Testing conditional mean independence for functional data', *Biometrika* **107**(2), 331–346.

Li, B. (2018), *Sufficient Dimension Reduction: Methods and Applications with R*, Chapman & Hall/CRC Monographs on Statistics and Applied Probability, CRC Press.

Li, B., Kim, M. K., Altman, N. et al. (2010), 'On dimension folding of matrix-or array-valued statistical objects', *The Annals of Statistics* **38**(2), 1094–1121.

Li, L. & Zhang, X. (2017), 'Parsimonious tensor response regression', *Journal of the American Statistical Association* **112**(519), 1131–1146.

Li, Q., Hsiao, C. & Zinn, J. (2003), 'Consistent specification tests for semiparametric/nonparametric models based on series estimation methods', *Journal of Econometrics* **112**(2), 295–325.

Li, X., Xu, D., Zhou, H. & Li, L. (2018), 'Tucker tensor regression and neuroimaging analysis', *Statistics in Biosciences* **10**(3), 520–545.

Liu, X. & Chen, E. (2019), 'Helping effects against curse of dimensionality in threshold factor models for matrix time series', *arXiv preprint arXiv:1904.07383* .

Luo, R., Wang, H., Tsai, C.-L. et al. (2009), 'Contour projected dimension reduction', *The Annals of Statistics* **37**(6B), 3743–3778.

Mammen, E. (1993), 'Bootstrap and wild bootstrap for high dimensional linear models', *The annals of statistics* pp. 255–285.

Park, T., Shao, X. & Yao, S. (2015), 'Partial martingale difference correlation', *Electronic Journal of Statistics* **9**(1), 1492–1517.

Rabusseau, G. & Kadri, H. (2016), Low-rank regression with tensor responses, *in* 'Proceedings of the 30th International Conference on Neural Information Processing Systems', pp. 1875–1883.

Sheng, W. & Yuan, Q. (2020), 'Sufficient dimension folding in regression via distance covariance for matrix-valued predictors', *Statistical Analysis and Data Mining: The ASA Data Science Journal* **13**(1), 71–82.

Subbaswamy, A., Schulam, P. & Saria, S. (2019), Preventing failures due to dataset shift: Learning predictive models that transport, *in* 'The 22nd International Conference on Artificial Intelligence and Statistics', PMLR, pp. 3118–3127.

Sun, W., Hao, B. & Li, L. (2021), 'Tensor data analysis', *Wiley StatsRef: Statistics Reference Online* pp. 1–26.

Sun, W. W. & Li, L. (2017), 'Store: sparse tensor response regression and neuroimaging analysis', *The Journal of Machine Learning Research* **18**(1), 4908–4944.

Székely, G. J. & Rizzo, M. L. (2014), 'Partial distance correlation with methods for dissimilarities', *The Annals of Statistics* **42**(6), 2382–2412.

Wang, D., Liu, X. & Chen, R. (2019), 'Factor models for matrix-valued high-dimensional time series', *Journal of econometrics* **208**(1), 231–248.

Wang, N., Zhang, X. & Li, B. (2022), 'Likelihood-based dimension folding on tensor data', *Statistica Sinica* **32**, 2405–2429.

Xia, Q., Xu, W. & Zhu, L. (2015), 'Consistently determining the number of factors in multivariate volatility modelling', *Statistica Sinica* pp. 1025–1044.

Xue, Y. & Yin, X. (2014), 'Sufficient dimension folding for regression mean function', *Journal of Computational and Graphical Statistics* **23**(4), 1028–1043.

Yuan, X.-T. & Zhang, T. (2013), 'Truncated power method for sparse eigenvalue problems', *Journal of Machine Learning Research* **14**(Apr), 899–925.

Zhang, X. & Li, L. (2017), 'Tensor envelope partial least-squares regression', *Technometrics* **59**(4), 426–436.

Zhou, H., Li, L. & Zhu, H. (2013), 'Tensor regression with applications in neuroimaging data analysis', *Journal of the American Statistical Association* **108**(502), 540–552.

Zhou, Y., Wong, R. K. & He, K. (2020), 'Broadcasted nonparametric tensor regression', *arXiv preprint arXiv:2008.12927* .

Zhou, Y., Wong, R. K. & He, K. (2021), 'Tensor linear regression: Degeneracy and solution', *IEEE Access* **9**, 7775–7788.

Zhu, L., Miao, B. & Peng, H. (2006), 'On sliced inverse regression with high-dimensional covariates', *Journal of the American Statistical Association* **101**(474), 630–643.

Zhu, X., Guo, X., Wang, T. & Zhu, L. (2020), 'Dimensionality determination: A thresholding double ridge ratio approach', *Computational Statistics & Data Analysis* **146**, 106910.

Zou, H., Hastie, T. & Tibshirani, R. (2006), 'Sparse principal component analysis', *Journal of computational and graphical statistics* **15**(2), 265–286.