Group SLOPE Penalized Low-Rank Tensor Regression

Yang Chen* YANGCHEN@BJTU.EDU.CN

Ziyan Luo   ZYLUO@BJTU.EDU.CN

School of Mathematics and Statistics
Beijing Jiaotong University
Beijing, P. R. China

Editor: Zhihua Zhang

Abstract

This article aims to seek a selection and estimation procedure for a class of tensor regression problems with multivariate covariates and matrix responses, which can provide theoretical guarantees for model selection in finite samples. Considering the frontal slice sparsity and low-rankness inherited in the coefficient tensor, we formulate the regression procedure as a group SLOPE penalized low-rank tensor optimization problem based on an orthogonal decomposition, namely TgSLOPE. This procedure provably controls the newly introduced tensor group false discovery rate (TgFDR), provided that the predictor matrix is column-orthogonal. Moreover, we establish the asymptotically minimax convergence with respect to the TgSLOPE estimate risk. For efficient problem resolution, we equivalently transform the TgSLOPE problem into a difference-of-convex (DC) program with the level-coercive objective function. This allows us to solve the reformulation problem of TgSLOPE by an efficient proximal DC algorithm (DCA) with global convergence. Numerical studies conducted on synthetic data and a real human brain connection data illustrate the efficacy of the proposed TgSLOPE estimation procedure.

Keywords: difference-of-convex, false discovery rate, group sparsity, low-rankness, tensor regression

1. Introduction

Tensor regression modeling, in which the regression coefficients take the form of a multi-way array or tensor, is an important and prevalent technique for coefficient estimation and/or feature selection in the high-dimensional statistical learning theory, with wide applications in many modern data science problems such as neuroimaging analysis, see, e.g., Zhou et al. (2013); Sun and Li (2017); Raskutti et al. (2019); Ahmed et al. (2020); Han et al. (2022). In this article, we focus on the tensor regression with multivariate covariates and matrix responses. Given n independent and identically distributed (i.i.d.) observations $\{(\boldsymbol{x}_i, \boldsymbol{Y}_i)\}_{i=1}^n$ with $\boldsymbol{x}_i \in \mathbb{R}^p$ the vector of predictors and $\boldsymbol{Y}_i \in \mathbb{R}^{p_1 \times p_2}$ the matrix of responses, the regression model can be expressed as follows

$$\boldsymbol{Y}_i = \mathcal{B}^* \times_3 \boldsymbol{x}_i + \boldsymbol{E}_i, i \in [n] := \{1, 2, \ldots, n\}, \tag{1}$$

where $\mathcal{B}^* \in \mathbb{R}^{p_1 \times p_2 \times p}$ is the unknown coefficient tensor with some inheritance structures like sparsity and/or low-rankness, $\boldsymbol{E}_i, i = 1, \ldots, n$, are i.i.d. noise matrices whose entries are

---

*. Corresponding author.

i.i.d. drawn from the Gaussian distribution $N(0, \sigma^2)$. Our goal is to seek a feature selection and tensor estimation procedure for model (1). A straightforward idea to estimate $\mathcal{B}^*$ is via optimization problems

$$\hat{\mathcal{B}} = \arg \min_{\mathcal{B} \in \Omega} f(\mathcal{B}; \mathcal{D}),$$

where Ω is any constraint set of sparse and/or low-rank tensors, $f(\mathcal{B}; \mathcal{D})$ can be taken as the least squares loss or any more general loss function with $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{Y}_i)\}_{i=1}^n$ the random sample set.

1.1 Related Work

One simple approach to estimate the sparse and/or low-rank coefficient tensor is to use matricization techniques such that the model (1) is degenerated into a linear matrix regression model, in which the estimated tensor $\mathcal{B}^*$ and the response matrix $\boldsymbol{Y}_i$ are unfolded into a $p \times p_1 p_2$ matrix and a $p_1 p_2$ dimensional vector, respectively. Therefore, a huge body of relevant work on sparse and low-rank matrix methods can be applied to deal with the tensor regression problems, see, e.g., variable selection by sparsity penalized methods (Obozinski et al., 2011; Raskutti et al., 2019; Chen et al., 2021); low-rank matrix estimation by reduced-rank regression methods (Bura et al., 2018; Fan et al., 2019; Wei et al., 2021) and nuclear norm penalized methods (Hu et al., 2020, 2021); joint penalized methods for selection and low-rank estimation (Yu et al., 2022); regularized covariance estimation for matrix-valued data (Zhang et al., 2022). However, the use of matricization techniques will not only break the sparse and low-rank structures of tensors, making resulting estimators difficult to interpret, but also lead to a dramatic increase in dimensionality, which is prone to over-fitting phenomenon.

Keeping the tensor format for regression models, the existing work can be divided into two categories from the perspective of different characterizations for the sparsity and low-rankness of coefficient tensors. One is based on the tensor decomposition, including CP decomposition (CPD) (Zhou et al., 2013; Sun and Li, 2017; Hao et al., 2020) and Tucker decomposition (Li et al., 2018; Zhang et al., 2020; Han et al., 2022; Hu et al., 2022). The other is the method without tensor decomposition, such as imposing sparsity on elements, fiber vectors, and slice matrices (Raskutti et al., 2019), assuming low-rankness of slice matrices (Chen et al., 2019; Raskutti et al., 2019; Kong et al., 2020), and considering Tucker low-rankness of coefficient tensors (Chen et al., 2019).

Focusing on the matrix response tensor regression model (1), limited work has been done on statistical property analysis and algorithm design. Sun and Li (2017) have considered an element-wise sparse tensor regression model based on the CPD, and established a non-asymptotic estimation error bound for the estimator obtained from the proposed alternating updating algorithm, which compounds the truncation-based sparse tensor decomposition procedure. Considering the frontal slice sparsity and low-rankness, Kong et al. (2020) have proposed a two-step screening and estimation procedure and shown that it enjoys estimation consistency and rank consistency. Hu et al. (2022) have developed a generalized Tucker decomposition model with features on multiple modes, and investigated the statistical convergence for the proposed supervised tensor decomposition algorithm with side information. Nevertheless, these methods may not work well for finite samples, since

their feature selection results are usually achieved in infinite samples, or unfortunately some of these are not capable of feature selection.

To seek a mechanism that enables to make inference about the validity of the selected model in finite samples, Bogdan et al. (2015) have introduced a new convex penalized method for classical linear regressions inspired by the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), namely Sorted L-One Penalized Estimation (SLOPE). They have shown that SLOPE controls the false discovery rate (FDR) under certain conditions. Following Bogdan et al. (2015), a recent line of research (Su and Candès, 2016; Bellec et al., 2018; Brzyski et al., 2019; Luo et al., 2019; Wei et al., 2021) has studied the SLOPE based methods, including statistical property analysis, algorithm design, etc. One of these work (Brzyski et al., 2019) has extended SLOPE to group SLOPE (gSLOPE) to deal with the situation when one aims to select whole groups of regressors instead of single one. This motivates us to investigate the use of gSLOPE penalty for feature selection and estimation in the tensor regression framework with finite samples.

1.2 Our Contributions

We consider the matrix response tensor regression model (1) and embed the frontal slice sparse and low-rank structures for the estimated tensor. Unlike the sparse and low-rank settings directly on the frontal slices (Kong et al., 2020), we characterize the inherited structures based on the low-rank, orthogonal decomposition (LROD). Similar to CPD, the orthogonal decomposition can also reduce the model dimensionality, thereby reducing the computational complexity of the regression procedure. In addition, the encouraging sparsity on factor matrices produced by decomposition has been shown not only to yield asymptotically consistent estimators in high-dimensional settings (Johnstone and Lu, 2009), but also to simplify visualization and interpretation of data analysis results (Allen, 2012). It is worth mentioning that there is always an optimal LROD approximation (Sørensen et al., 2012) for a given tensor, while the rank-k approximation problem for fitting a CPD tensor is ill-posed for many rank values of k in general (Silva and Lim, 2008). This makes the LROD tensor regression procedure more stable than those CPD based methods.

To investigate the selection and estimation properties, we develop a sparse and low-rank tensor regression procedure by formulating it as a gSLOPE penalized LROD tensor optimization problem, namely TgSLOPE. Then, to measure its feature selection performance in finite samples, we define the notion of the tensor group FDR (TgFDR). Under the column-orthogonality assumption on the predictor matrix, TgSLOPE is shown to control TgFDR at any given level $0 < q < 1$ with appropriate choice of the regularization parameters. Moreover, the tensor estimator produced by TgSLOPE provably achieves the asymptotically minimax rate. Overall, our proposed TgFDRT controlling method can also be minimax optimal with respect to the estimation risk.

To well resolve our proposed TgSLOPE model with Stiefel manifold constraints, we constructively reformulate this manifold optimization problem as a difference-of-convex (DC) program whose objective function is shown to be level-coercive. This allows us to adopt some globally convergent DC-type algorithm in which the decision variables are updated by the proximal operator of the gSLOPE penalty. Simulations on synthetic data verify the TgFDR control, and test the effects of sparsity, model size and LROD rank on TgSLOPE

(a) Horizontal slices $\boldsymbol{A}_{j_1::}$ (b) Lateral slices $\boldsymbol{A}_{:j_2:}$ (c) Frontal slices $\boldsymbol{A}_{::j_3}$
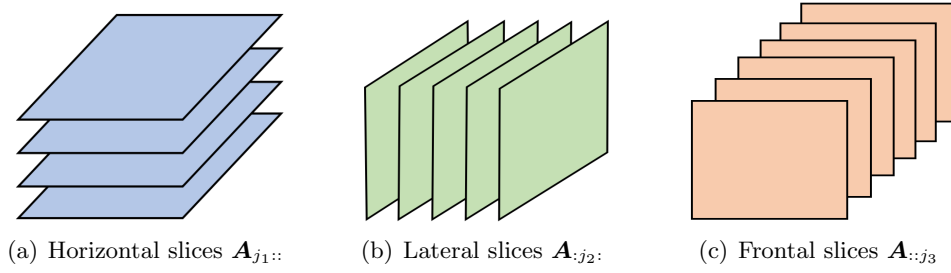
Figure 1: Slices of a order-3 tensor.

performance respectively. In addition, numerical results on both synthetic data and a real human brain connection data confirm the superiority of our proposed TgSLOPE procedure by comparing it with several state-of-the-art approaches.

1.3 Notation and Preliminaries

Throughout the article, we denote scalars, vectors, matrices and tensors by lowercase letters (e.g., a, b), boldface lowercase letters (e.g., $\boldsymbol{a}, \boldsymbol{b}$), boldface uppercase letters (e.g., $\boldsymbol{A}, \boldsymbol{B}$), and calligraphic letters (e.g., $\mathcal{A}, \mathcal{B}$), respectively. In addition, zero scalars, vectors, matrices and tensors are respectively denoted by $0, \boldsymbol{0}, \boldsymbol{O}$ and $\mathcal{O}$. For a vector $\boldsymbol{a} \in \mathbb{R}^n$, we denote $\boldsymbol{a}_{[1:s]} = (a_1, \ldots, a_s)^\top$ with $s \leq n$. For a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, the ith row and the jth column are denoted by $\boldsymbol{a}_{i:}$ and $\boldsymbol{a}_{:j}$, respectively, then write $\boldsymbol{A}_{[:,1:s]} = (\boldsymbol{a}_{:1}, \ldots, \boldsymbol{a}_{:s})$ with $s \leq n$. Figure 1 shows the horizontal, lateral, and frontal slices of the tensor $\mathcal{A} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, denoted by $\boldsymbol{A}_{j_1::} \in \mathbb{R}^{p_2 \times p_3}, \boldsymbol{A}_{:j_2:} \in \mathbb{R}^{p_1 \times p_3}$, and $\boldsymbol{A}_{::j_3} \in \mathbb{R}^{p_1 \times p_2}$, respectively. For convenience, we simply denote the ith row of $\boldsymbol{A}$ as $\boldsymbol{a}_i$ and the j_3th frontal slice of $\mathcal{A}$ as $\boldsymbol{A}_{j_3}$.

Given vectors $\boldsymbol{a} \in \mathbb{R}^m$ and $\boldsymbol{b} \in \mathbb{R}^n$, denote the outer product $\boldsymbol{a} \circ \boldsymbol{b} = \boldsymbol{a}\boldsymbol{b}^\top \in \mathbb{R}^{m \times n}$ and Kronecker product $\boldsymbol{a} \otimes \boldsymbol{b} = (a_1\boldsymbol{b}^\top, \ldots, a_m\boldsymbol{b}^\top)^\top \in \mathbb{R}^{mn}$. For matrices $\boldsymbol{A} \in \mathbb{R}^{m \times q}$ and $\boldsymbol{B} \in \mathbb{R}^{n \times q}$, the Khatri-Rao product is defined as $\boldsymbol{A} \odot \boldsymbol{B} = (\boldsymbol{a}_{:1} \otimes \boldsymbol{b}_{:1}, \ldots, \boldsymbol{a}_{:q} \otimes \boldsymbol{b}_{:q}) \in \mathbb{R}^{mn \times q}$; if $m = n$, the Hadamard product is defined as $\boldsymbol{A} * \boldsymbol{B} = [a_{ij}b_{ij}] \in \mathbb{R}^{m \times q}$, and the inner product is defined as $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \sum_{i,j} a_{ij}b_{ij}$. For an order-3 tensor $\mathcal{A} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, the mode-3 product with a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p_3}$ is denoted by $\mathcal{A} \times_3 \boldsymbol{X} \in \mathbb{R}^{p_1 \times p_2 \times n}$ with elements $(\mathcal{A} \times_3 \boldsymbol{X})_{i,j,l} = \sum_{k=1}^{p_3} \mathcal{A}_{i,j,k} \boldsymbol{X}_{l,k}$. We also denote the matricization operator as $\mathcal{M}_3(\cdot)$, which unfolds the tensor $\mathcal{A}$ along the third mode into the matrix $\mathcal{M}_3(\mathcal{A}) \in \mathbb{R}^{p_3 \times p_1 p_2}$. Specifically, $(\mathcal{M}_3(\mathcal{A}))_{k,l} = \mathcal{A}_{i,j,k}$ with $l = 1 + (i - 1) + (j - 1)p_1$. Then the inverse of mode-3 unfolding can be denoted as $\mathcal{M}_3^{-1}(\cdot)$. The tensor $\mathcal{A}$ is called rank-one if it can be written as the outer product of vectors $\mathcal{A} = \boldsymbol{u} \circ \boldsymbol{v} \circ \boldsymbol{w}$. More about tensor operations can be found in Kolda and Bader (2009).

Next we introduce some norms. For a vector $\boldsymbol{a} \in \mathbb{R}^n$, denote its ℓ_2-norm as $\|\boldsymbol{a}\| = \sqrt{\sum_i a_i^2}$ and ℓ_0-norm as $\|\boldsymbol{a}\|_0 = \sharp\{i : |a_i| \neq 0\}$. For a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, denote its Frobenius norm as $\|\boldsymbol{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$ and trace as $\operatorname{tr}(\boldsymbol{A}) = \sum_{i=j} a_{ij}$. Write singular values of $\boldsymbol{A}$ as $\sigma_1(\boldsymbol{A}) \geq \cdots \geq \sigma_p(\boldsymbol{A})$ with $p = \min\{m, n\}$. Then the nuclear norm is $\|\boldsymbol{A}\|_* = \sum_k \sigma_k(\boldsymbol{A})$ and the spectral norm is $\|\boldsymbol{A}\|_2 = \sigma_1(\boldsymbol{A})$. We also introduce the notation

$\|\boldsymbol{A}\|_r = (\|\boldsymbol{a}_1\|, \dots, \|\boldsymbol{a}_m\|)^\top$ for any matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\|\mathcal{A}\|_f = (\|\boldsymbol{A}_1\|_F, \dots, \|\boldsymbol{A}_{p_3}\|_F)^\top$ for any tensor $\mathcal{A} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$.

Moreover, we say that a random variable X has a chi distribution with n degrees of freedom, written by $X \sim \chi_n$, if it is the square root of a chi-squared random variable, i.e., $X^2 \sim \chi_n^2$. For a proper closed convex function $f : \mathbb{R}^{m \times n} \to (-\infty, \infty]$, the subdifferential of f at any given $\boldsymbol{X} \in dom(f)$, says $\partial f(\boldsymbol{X})$, is defined by

$$\partial f(\boldsymbol{X}) = \{\boldsymbol{G} \in \mathbb{R}^{m \times n} : f(\boldsymbol{Y}) \geq f(\boldsymbol{X}) + \langle \boldsymbol{G}, \boldsymbol{Y} - \boldsymbol{X} \rangle \; for \; all \; \boldsymbol{Y} \in \mathbb{R}^{m \times n}\},$$

where each matrix $\boldsymbol{G} \in \partial f(\boldsymbol{X})$ is called a subgradient of f at $\boldsymbol{X}$. The proximal operator of f is given by

$$\mathrm{Prox}_f(\boldsymbol{X}) = \arg\min_{\boldsymbol{Z}} \left\{ f(\boldsymbol{Z}) + \frac{1}{2}\|\boldsymbol{Z} - \boldsymbol{X}\|_F^2 \right\}, \quad \forall \boldsymbol{X} \in \mathbb{R}^{m \times n}.$$

Given positive scalars a and b, denote $a \sim b$ if $a/b \to 1$, and $a \asymp b$ if there exist uniform constants $c, C > 0$ such that $ca \leq b \leq Ca$.

The remainder of this article is organized as follows. In Section 2, we introduce a frontal slice sparse and low-rank tensor regression procedure, which optimizes the gSLOPE penalized LROD tensor optimization problem, and then the identifiability of parameters is analyzed. Section 3 makes statistical theory analysis for TgSLOPE procedure, including the TgFDR control at the prespecified level and the asymptotically minimax convergence with respect to the ℓ_2-loss. An efficient and globally convergent pDCAe algorithm is proposed in Section 4. Section 5 reports some numerical studies to verify the performance of our proposed TgSLOPE approach. Concluding remarks are drawn in Section 6. All proofs are deferred to the Appendix.

2. Model

In this section, we propose a feature selection and tensor estimate method for the tensor regression model (1), which can also be rewritten as

$$\mathcal{Y} = \mathcal{B}^* \times_3 \boldsymbol{X} + \mathcal{E}, \tag{2}$$

where $\mathcal{Y}, \mathcal{E} \in \mathbb{R}^{p_1 \times p_2 \times n}, \boldsymbol{X} \in \mathbb{R}^{n \times p}$. As discussed above, it is crucial to introduce some sparse and low-rank structures in order to facilitate estimation of the ultrahigh dimensional unknown parameters in finite samples. Thus, for the efficient feature selection, we consider the frontal slice sparsity of the coefficient tensor $\mathcal{B}^*$ based on the low-rank decomposition in the tensor regression framework (2).

2.1 TgSLOPE Estimate

Assume that the coefficient tensor is LROD, that is, the tensor $\mathcal{B}^*$ admits a rank-K CPD with column-orthogonal factor matrices. Specifically, the rank-K CPD models a tensor as a sum of K rank-one tensors (Kolda and Bader, 2009), i.e., $\mathcal{B}^* = \sum_{k=1}^K \boldsymbol{u}_k^* \circ \boldsymbol{v}_k^* \circ \boldsymbol{w}_k^*$, where the factor vectors $\boldsymbol{u}_k^* \in \mathbb{R}^{p_1}, \boldsymbol{v}_k^* \in \mathbb{R}^{p_2}, \boldsymbol{w}_k^* \in \mathbb{R}^p, k = 1, \dots, K$. Figure 2 illustrates the rank-K CPD. Denote the factor matrices composed of the vectors from the rank-one components by $\boldsymbol{U}^* = (\boldsymbol{u}_1^*, \boldsymbol{u}_2^*, \dots, \boldsymbol{u}_K^*) \in \mathbb{R}^{p_1 \times K}$, $\boldsymbol{V}^* \in \mathbb{R}^{p_2 \times K}$ and $\boldsymbol{W}^* \in \mathbb{R}^{p \times K}$ respectively. The CPD can
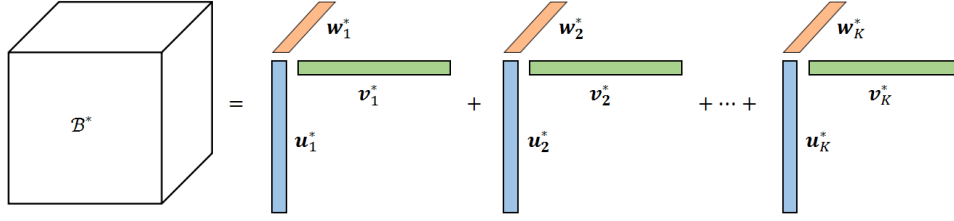
Figure 2: Illustration of the rank-K CPD of the tensor $\mathcal{B}^*$.

be rewritten as $\mathcal{B}^* = [\![\boldsymbol{U}^*, \boldsymbol{V}^*, \boldsymbol{W}^*]\!]$. The LROD assumes that the factor matrices $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$ produced by CPD are column-orthogonal. We define the rank of LROD tensor as the CP rank K. Moreover, we consider the row sparsity settings on the mode-3 factor matrix $\boldsymbol{W}^*$, which can translate to the sparsity of the frontal slices for the tensor $\mathcal{B}^*$.

To identify significant features and estimate the coefficient tensor $\mathcal{B}^*$, we develop a penalized sparse and low-rank tensor regression method which optimizes the following gSLOPE penalized LROD tensor optimization problem (TgSLOPE)

$$
\begin{aligned}
\min_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}} \quad & \frac{1}{2}\big\|\mathcal{Y} - [\![\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}]\!] \times_3 \boldsymbol{X}\big\|_F^2 + P_{\boldsymbol{\lambda}}\big(\|\boldsymbol{W}\|_r\big) \\
\text{s.t.} \quad & \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_K,
\end{aligned}
\tag{3}
$$

where $P_{\boldsymbol{\lambda}}(\boldsymbol{x}) = \sum_{j=1}^p \lambda_j |x|_{(j)}$ is the SLOPE penalty with the regularization parameter vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^\top$ satisfying

$$
\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0 \text{ and } \lambda_1 > 0,
$$

and $|x|_{(j)}$ the jth largest component of $\boldsymbol{x}$ in magnitude.

According to the property of the matrix Khatri-Rao product that for any $\boldsymbol{A} \in \mathbb{R}^{m \times k}$ and $\boldsymbol{B} \in \mathbb{R}^{n \times k}$, $(\boldsymbol{A} \odot \boldsymbol{B})^\top (\boldsymbol{A} \odot \boldsymbol{B}) = (\boldsymbol{A}^\top \boldsymbol{A}) * (\boldsymbol{B}^\top \boldsymbol{B})$, we have the following lemma.

Lemma 1 *For any two column-orthogonal matrices $\boldsymbol{A} \in \mathbb{R}^{m \times k}$ and $\boldsymbol{B} \in \mathbb{R}^{n \times k}$, we have $\boldsymbol{C} = \boldsymbol{A} \odot \boldsymbol{B} \in \mathbb{R}^{mn \times k}$ is column-orthogonal, that is, $\boldsymbol{C}^\top \boldsymbol{C} = \boldsymbol{I}_k$.*

Following the matricized form of a tensor, we have $\mathcal{M}_3(\mathcal{B}^*) = \boldsymbol{W}^* (\boldsymbol{V}^* \odot \boldsymbol{U}^*)^\top$. Denote $\boldsymbol{H}^* = \boldsymbol{V}^* \odot \boldsymbol{U}^* \in \mathbb{R}^{p_1 p_2 \times K}$. Without loss of generality, we assume that $K \leq \min\{p, p_1 p_2\}$. It is known from Lemma 1 that $\boldsymbol{H}^{*\top} \boldsymbol{H}^* = \boldsymbol{I}_K$. Thus, the TgSLOPE problem (3) can be simplified as

$$
\begin{aligned}
\min_{\boldsymbol{W},\boldsymbol{H}} \quad & L(\boldsymbol{W},\boldsymbol{H}) + P_{\boldsymbol{\lambda}}\big(\|\boldsymbol{W}\|_r\big) \\
\text{s.t.} \quad & \boldsymbol{H}^\top \boldsymbol{H} = \boldsymbol{I}_K,
\end{aligned}
\tag{4}
$$

where the loss $L(\boldsymbol{W},\boldsymbol{H}) = \frac{1}{2}\big\|\mathcal{M}_3(\mathcal{Y}) - \boldsymbol{X}\boldsymbol{W}\boldsymbol{H}^\top\big\|_F^2$. The estimator $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ produced by (4) and the tensor estimator $\hat{\mathcal{B}}$ are linked via $\mathcal{M}_3(\hat{\mathcal{B}}) = \hat{\boldsymbol{W}}\hat{\boldsymbol{H}}^\top$.

Notably, the LROD tensor decomposition with the additional column-orthogonality on U and V turns to be a special CPD. While the best CP low-rank approximation of a

tensor may not exist and the exact CP rank of a tensor is hard to compute, the imposed column-orthogonality on U and V can make sure that there is always an optimal rank-k approximation for a given tensor (Sørensen et al., 2012). The resulting approximation tensor that admits such a special CPD has been coined as an orthogonally decomposable tensor, which has been studied in the communities of tensor analysis and tensor learning, see, e.g., the perturbation bounds and the applications to the unsupervised learning scenario of tensor SVD and the supervised task of tensor regression by Auddy and Yuan (2023), and the linear convergence of an alternating polar decomposition method for low rank orthogonal tensor approximations by Hu and Ye (2023). Additionally, Poythressa et al. (2022) have illustrated that the orthogonally decomposable tensor model has the potential to perform better than CPD model in terms of predictive performance and model interpretation in numerical experiments. Technically, the column-orthogonal constraints on factor matrices can help to fix the indeterminacy and the non-uniqueness of CPD, thereby improving identifiability results of CPD methods (see, e.g., Zhou et al. (2013)).

2.2 Identifiability

The identifiability of parameters is analyzed in this subsection. Considering CPD of a tensor $\mathcal{B} = [\![U, V, W]\!]$, Zhou et al. (2013) have stated that the parameters in the tensor model is not identifiable due to two complications. One is the indeterminacy coming from scaling and permutation, and the second is possibly non-uniqueness of decomposition. For LROD tensor, the scaling indeterminacy can be avoided automatically due to the column-orthogonal U and V. To fix the permutation indeterminacy, we adopt the convention that the first row of the factor matrix U is assumed to be arranged in descending order. Moreover, it follows from the Kruskal's condition of uniqueness of CPD (Theorem 4b, Kruskal (1977)) that the LROD is unique up to permutation if $k(W) \geq 2$, where $k(W)$ is the k-rank of W defined by the largest integer k such that every subset of k columns of W is linearly independent. Under the assumptions of determinacy and uniqueness of LROD, the imposed frontal slice sparsity on coefficient tensor $\mathcal{B}$ can be equivalently translated to the row sparsity settings on the mode-3 factor matrix W, which yields the feasible set of problem (4) being

$$\mathbb{B}_T = \left\{ \mathcal{B} = [\![U, V, W]\!] : U^\top U = V^\top V = I_K, \left\| \|W\|_r \right\|_0 \leq s \right\}, \qquad (5)$$

where $s > 0$ is a prescribed parameter that controls the frontal slice sparsity.

We give conditions of global identifiability result in the following proposition. The proof is presented in Appendix A.

Proposition 2 *Consider the tensor regression model (2) with the coefficient tensor $\mathcal{B}^* = [\![U^*, V^*, W^*]\!] \in \mathbb{B}_T$. Then $\mathcal{B}^*$ is globally identifiable up to permutation if $k(W^*) \geq 2$ and $\left\| \|\mathcal{B}^*\|_f \right\|_0 \leq k(X)/2$.*

Remark 3 *For sparse parameter models, Donoho and Elad (2003) have discussed the conditions for the uniqueness of sparse coefficient vector in classical linear regression frameworks. The identifiability result in Proposition 2 can be regarded as an extension of Corollary 1 (Donoho and Elad, 2003) to the sparse and decomposable tensor regression models. Provided that the LROD tensor decomposition is unique up to permutation, the full coefficient*

tensor $\mathcal{B}^*$ with the frontal slice sparsity is globally identifiable and thus the LROD of $\mathcal{B}^*$ is identifiable.

3. Statistical Results

This section is devoted to the TgFDR control and the estimate accuracy for our proposed TgSLOPE procedure.

3.1 TgFDR Control

FDR is a commonly used error rate that counts the expected proportion of errors among the rejected hypotheses in multiple testing. In this subsection, we develop the classic FDR notion to the setting of tensor regression and show that it can be well controlled by our proposed TgSLOPE.

Definition 4 *Consider the tensor regression model (2) and let $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ be an estimator given by the optimization problem (4). We define the tensor group false discovery rate (*TgFDR*) for TgSLOPE as*

$$\text{TgFDR} = \mathbb{E}\left[\frac{V}{\max\{R, 1\}}\right], \tag{6}$$

where V, R are defined as follows

$$V = \sharp\{j \in [p] : \boldsymbol{W}_j^* = \boldsymbol{0}, \hat{\boldsymbol{W}}_j \neq \boldsymbol{0}\}, \quad R = \sharp\{j \in [p] : \hat{\boldsymbol{W}}_j \neq \boldsymbol{0}\}$$

with $\boldsymbol{W}_j^*$ and $\hat{\boldsymbol{W}}_j$ the jth rows of $\boldsymbol{W}^*$ and $\hat{\boldsymbol{W}}$, respectively.

Define the regularization parameters of the TgSLOPE procedure as

$$\lambda_j = \sigma F_{\chi_K}^{-1}\big(1 - q \cdot j/p\big), \; j \in [p], \tag{7}$$

where $0 < q < 1$, $F_{\chi_K}^{-1}(\alpha)$ is the αth quantile of the χ distribution with K degrees of freedom. We give an upper bound of TgFDR in the following theorem. The technical proof is deferred to Appendix B.1.

Theorem 5 *Consider the tensor regression model (2) with the predictor matrix $\boldsymbol{X}$ satisfying $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}_p$. Then, for any solution $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ given by the TgSLOPE problem (4) with the regularization parameters in (7), TgFDR obeys*

$$\text{TgFDR} = \mathbb{E}\left[\frac{V}{\max\{R, 1\}}\right] \leq q \cdot \frac{p-s}{p}$$

with s the number of nonzero rows of $\boldsymbol{W}^*$.

Remark 6 *Under the guarantee of the uniqueness for LROD, our row sparsity settings on the matrix $\boldsymbol{W}^*$ can be equivalently interpreted as the sparsity in the frontal slices of the coefficient tensor $\mathcal{B}^*$. Figure 3 illustrates this equivalence relation. Therefore, the definition of TgFDR in (6) can be redefined based on the following random variables*

$$TV = \sharp\{j \in [p] : \boldsymbol{B}_j^* = \boldsymbol{O}, \hat{\boldsymbol{B}}_j \neq \boldsymbol{O}\}, \quad TR = \sharp\{j \in [p] : \hat{\boldsymbol{B}}_j \neq \boldsymbol{O}\}$$
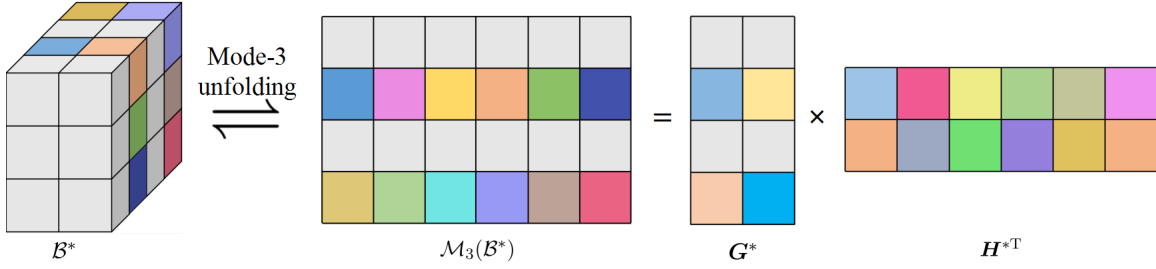
Figure 3: Illustration of the sparsity equivalence relation between factor matrix $\boldsymbol{W}^*$ and tensor $\mathcal{B}^*$. Here the elements of matrices and tensors are zero (gray) and nonzero (color).

with $\boldsymbol{B}_j^*$ and $\hat{\boldsymbol{B}}_j$ the jth frontal slices of $\mathcal{B}^*$ and $\hat{\mathcal{B}}$, respectively. This indicates that the estimator $\hat{\mathcal{B}} = \mathcal{M}_3^{-1}(\hat{\boldsymbol{W}}\hat{\boldsymbol{H}}^\top)$ given by (4) provably controls TgFDR at any prespecified level $q \in (0, 1)$.

3.2 Estimate Accuracy

This subsection aims to show that TgSLOPE enjoys minimax optimal estimation property under the assumption that the ground truth $\mathcal{B}^*$ is bounded, that is, $\max\{\|\boldsymbol{B}_j^*\|_F, j \in [p]\} < \infty$. We measure the deviation of an estimator from $\mathcal{B}^*$ in the following theorem. The technical proof is given in Appendix B.2.

Theorem 7 *Consider the TgSLOPE optimization problem (4) with the predictor matrix $\boldsymbol{X}$ satisfying $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}_p$ and the regularization parameters in (7). Then, under $p \to \infty$ and $s/p \to 0$, the TgSLOPE estimator enjoys the asymptotically minimax rate over $\mathbb{B}_s = \{\mathcal{B} : \big\|\|\mathcal{B}\|_f\big\|_0 \leq s\}$, that is,*

$$\sup_{\mathcal{B}^* \in \mathbb{B}_s} \mathbb{E}\|\hat{\mathcal{B}} - \mathcal{B}^*\|_F^2 \sim \inf_{\tilde{\mathcal{B}}} \sup_{\mathcal{B}^* \in \mathbb{B}_s} \mathbb{E}\|\tilde{\mathcal{B}} - \mathcal{B}^*\|_F^2,$$

where the infimum is taken over all estimators $\tilde{\mathcal{B}}$ based on date set $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{Y}_i)\}_{i=1}^n$.

Remark 8 *The asymptotically minimax result in Theorem 7 is an extension of Theorem 1.1 in Su and Candès (2016) (for the SLOPE estimate) and Theorem 2.2 in Brzyski et al. (2019) (for the group SLOPE estimate) under the classical linear regression frameworks. In addition, we note that the choice of the regularization parameters in Theorem 7 does not depend on the sparsity level. In such a sense, TgSLOPE is adaptive to a range of sparsity in achieving the minimax optimality. It would also be interesting to work on estimation bounds of minimax rates in a non-asymptotic manner, such as that in Raskutti et al. (2019) and Hao et al. (2020). We leave this as one of our research topics in the future.*

4. Proximal DCA with Extrapolation

In this section, we propose a proximal difference-of-convex algorithm with extrapolation (pDCAe) to solve the TgSLOPE optimization problem and establish its global convergence.

4.1 DC Reformulation

Given an optimal solution $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ of the problem (4), it is known from the orthogonal Procrustes problem (Gower and Dijksterhuis, 2004) that $\hat{\boldsymbol{H}}$ can be determined by $\hat{\boldsymbol{W}}$ in the way that $\hat{\boldsymbol{H}} = \hat{\boldsymbol{U}}_{[:,1:K]}\hat{\boldsymbol{V}}^\top$, where $\hat{\boldsymbol{U}}, \hat{\boldsymbol{V}}$ are obtained from the SVD $\mathcal{M}_3(\mathcal{Y})^\top \boldsymbol{X}\hat{\boldsymbol{W}} = \hat{\boldsymbol{U}}\boldsymbol{D}\hat{\boldsymbol{V}}^\top$ with $\hat{\boldsymbol{U}} \in \mathbb{R}^{p_1 p_2 \times p_1 p_2}$, $\boldsymbol{D} \in \mathbb{R}^{p_1 p_2 \times K}$ and $\hat{\boldsymbol{V}} \in \mathbb{R}^{K \times K}$. Plugging $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ into the objective function of problem (4), we obtain that

$$
\begin{aligned}
L(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}) &= \frac{1}{2}\big\|\mathcal{M}_3(\mathcal{Y})\big\|_F^2 + \frac{1}{2}\big\|\boldsymbol{X}\hat{\boldsymbol{W}}\hat{\boldsymbol{H}}^\top\big\|_F^2 - \operatorname{tr}(\mathcal{M}_3(\mathcal{Y})^\top \boldsymbol{X}\hat{\boldsymbol{W}}\hat{\boldsymbol{H}}^\top) \\
&= \frac{1}{2}\big\|\mathcal{M}_3(\mathcal{Y})\big\|_F^2 + \frac{1}{2}\big\|\boldsymbol{X}\hat{\boldsymbol{W}}\big\|_F^2 - \operatorname{tr}(\boldsymbol{D}\hat{\boldsymbol{U}}_{[:,1:K]}^\top \hat{\boldsymbol{U}}) \\
&= \frac{1}{2}\big\|\mathcal{M}_3(\mathcal{Y})\big\|_F^2 + \frac{1}{2}\big\|\boldsymbol{X}\hat{\boldsymbol{W}}\big\|_F^2 - \operatorname{tr}(\boldsymbol{D}) \\
&= \frac{1}{2}\big\|\mathcal{M}_3(\mathcal{Y})\big\|_F^2 + \frac{1}{2}\big\|\boldsymbol{X}\hat{\boldsymbol{W}}\big\|_F^2 - \|\mathcal{M}_3(\mathcal{Y})^\top \boldsymbol{X}\hat{\boldsymbol{W}}\|_*.
\end{aligned}
$$

Therefore, the TgSLOPE problem (4) can be equivalently transformed into

$$
\begin{cases}
\hat{\boldsymbol{W}} = \arg\min\left\{F(\boldsymbol{W}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{W}\|_F^2 + P_{\boldsymbol{\lambda}}\big(\|\boldsymbol{W}\|_r\big) - \|\mathcal{M}_3(\mathcal{Y})^\top \boldsymbol{X}\boldsymbol{W}\|_*\right\}, \\[2mm]
\hat{\boldsymbol{H}} = \hat{\boldsymbol{U}}_{[:,1:K]}\hat{\boldsymbol{V}}^\top, \text{ where } \hat{\boldsymbol{U}} \text{ and } \hat{\boldsymbol{V}} \text{ are from the SVD } \mathcal{M}_3(\mathcal{Y})^\top \boldsymbol{X}\hat{\boldsymbol{W}} = \hat{\boldsymbol{U}}\boldsymbol{D}\hat{\boldsymbol{V}}^\top.
\end{cases}
\tag{8}
$$

Denote $F1(\boldsymbol{W}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{W}\|_F^2$ and $F2(\boldsymbol{W}) = \frac{L}{2}\|\boldsymbol{W}\|_F^2 - F1(\boldsymbol{W})$. Since $F1$ is gradient Lipschitz continuous with modulus $L = \|\boldsymbol{X}^\top \boldsymbol{X}\|_2$, the convexity of $F2$ follows. Inspired by the DC-representable function from Gotoh et al. (2018), we rewrite the optimization problem in (8) as

$$
\min_{\boldsymbol{W}\in\mathbb{R}^{p\times K}} F(\boldsymbol{W}) = \underbrace{\frac{L}{2}\|\boldsymbol{W}\|_F^2 + P_{\boldsymbol{\lambda}}\big(\|\boldsymbol{W}\|_r\big)}_{C1(\boldsymbol{W})} - \underbrace{\left(\frac{L}{2}\|\boldsymbol{W}\|_F^2 - \frac{1}{2}\|\boldsymbol{X}\boldsymbol{W}\|_F^2 + N(\boldsymbol{W})\right)}_{C2(\boldsymbol{W})}, \tag{9}
$$

where $N(\boldsymbol{W}) = \|\mathcal{M}_3(\mathcal{Y})^\top \boldsymbol{X}\boldsymbol{W}\|_*$. It is easy to verify that both $C1$ and $C2$ are convex, leading to a DC program as in (9).

By now, we constructively reformulate the manifold optimization problem (4) as a DC program (9). This allows us to solve the reformulation problem of TgSLOPE by a DC-type algorithm.

4.2 Proximal DCA

Considering the DC program (9), we first give the subdifferential of $N(\boldsymbol{W})$ in the following lemma, which can be clearly derived from the subdifferential of the matrix nuclear norm (Watson, 1992) and the chain rule of Theorem 10.6 in Rockafellar and Wets (2009).

Lemma 9 *The subdifferential of $N(\boldsymbol{W})$ is given by*

$$\partial N(\boldsymbol{W}) = \{\boldsymbol{X}^\top \mathcal{M}_3(\mathcal{Y})(\tilde{\boldsymbol{U}}_{[:,1:K]}\tilde{\boldsymbol{V}}^\top + \tilde{\boldsymbol{W}}) : \tilde{\boldsymbol{U}}_{[:,1:K]}^\top \tilde{\boldsymbol{W}} = \boldsymbol{O}, \tilde{\boldsymbol{W}}\bar{\boldsymbol{V}} = \boldsymbol{O}, \|\tilde{\boldsymbol{W}}\|_2 \le 1\}, \quad (10)$$

where $\tilde{\boldsymbol{U}}, \tilde{\boldsymbol{V}}$ are obtained by the SVD of $\mathcal{M}_3(\mathcal{Y})^\top \boldsymbol{X}\boldsymbol{W} = \tilde{\boldsymbol{U}}\boldsymbol{D}\tilde{\boldsymbol{V}}^\top$.

The DCA iterative scheme for (9) takes the following form

$$
\begin{aligned}
\boldsymbol{W}^{(k+1)} &= \arg\min_{\boldsymbol{W}} \left\{ \frac{L}{2}\|\boldsymbol{W}\|_F^2 - \left\langle \boldsymbol{W}, (L\boldsymbol{I} - \boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{W}^{(k)} + \boldsymbol{Q}_1^{(k)} \right\rangle + P_{\boldsymbol{\lambda}}(\|\boldsymbol{W}\|_r) \right\} \\
&= \arg\min_{\boldsymbol{W}} \left\{ \frac{1}{2}\left\| \boldsymbol{W} - \left(\boldsymbol{W}^{(k)} - \frac{1}{L}\left(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{W}^{(k)} - \boldsymbol{Q}_1^{(k)} \right) \right) \right\|_F^2 + P_{\boldsymbol{\lambda}/L}(\|\boldsymbol{W}\|_r) \right\},
\end{aligned}
$$
$$(11)$$

where $\boldsymbol{Q}_1^{(k)} \in \partial N(\boldsymbol{W}^{(k)})$. Then the optimal solution of problem (11) can be computed by the proximal operator of the penalty function $P_{\boldsymbol{\lambda}/L}$. Specifically,

$$
\begin{cases}
\boldsymbol{\eta}^{(k+1)} = \arg\min_{\boldsymbol{\eta}} \left\{ \frac{1}{2}\big\|\|\boldsymbol{Q}^{(k)}\|_r - \boldsymbol{\eta}\big\|^2 + P_{\boldsymbol{\lambda}/L}(\boldsymbol{\eta}) \right\}, \\
\boldsymbol{W}_j^{(k+1)} = \big(\mathrm{Prox}_{P_{\boldsymbol{\lambda}/L}}(\boldsymbol{Q}^{(k)})\big)_j = \eta_j^{(k+1)} \frac{\boldsymbol{q}_j^{(k)}}{\|\boldsymbol{q}_j^{(k)}\|}, j \in [p],
\end{cases}
\quad (12)
$$

where $\boldsymbol{Q}^{(k)} = \boldsymbol{W}^{(k)} - \big(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{W}^{(k)} - \boldsymbol{Q}_1^{(k)}\big)/L$. The resulting DCA is called as the proximal DCA (Gotoh et al., 2018; Wen et al., 2018). Consequently, the $\boldsymbol{W}$-subproblem is reduced to identifying the general SLOPE proximal operator, which can be efficiently solved by FastProxSL1 (Bogdan et al., 2015). Furthermore, in order to possibly accelerate the algorithm, we entertain extrapolation techniques (Nesterov, 2013) in the proximal DCA. The algorithmic framework is summarized in Algorithm 1.

Algorithm 1 pDCAe for solving TgSLOPE (4)

Initialize $\boldsymbol{W}^{(0)} \in \mathbb{R}^{p \times K}$, $\{\beta^{(k)}\} \subseteq [0,1)$ with $\sup_k \beta^{(k)} < 1$, set $\boldsymbol{W}^{(-1)} = \boldsymbol{W}^{(0)}$, $k = 0$;

1: Choose $\boldsymbol{Q}_1^{(k)} \in \partial N(\boldsymbol{W}^{(k)})$ by (10);
2: Compute $\boldsymbol{A}^{(k)} = \boldsymbol{W}^{(k)} + \beta^{(k)}(\boldsymbol{W}^{(k)} - \boldsymbol{W}^{(k-1)})$ and update

$$\boldsymbol{Q}^{(k)} = \boldsymbol{A}^{(k)} - \big(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{A}^{(k)} - \boldsymbol{Q}_1^{(k)}\big)/L;$$

3: Compute $\boldsymbol{W}^{(k+1)}$ by the two-step form in (12);
4: Set $k = k + 1$. If the stopping criterion is met, perform the SVD $\mathcal{M}_3(\mathcal{Y})^\top \boldsymbol{X}\boldsymbol{W}^{(k)} = \boldsymbol{U}^{(k)}\boldsymbol{D}^{(k)}\boldsymbol{V}^{(k)\top}$ to get

$$\boldsymbol{H}^{(k)} = \big(\boldsymbol{U}^{(k)}\big)_{[:,1:K]}\boldsymbol{V}^{(k)\top},$$

then stop and return $\mathcal{B}^{(k)} = \mathcal{M}_3^{-1}(\boldsymbol{W}^{(k)}\boldsymbol{H}^{(k)\top})$; otherwise, go to the step 1.

Note that the pDCAe algorithm is reduced to the general proximal DCA if $\beta^{(k)} = 0$ for all k. Moreover, Algorithm 1 can be coupled with many popular choices of extrapolation parameters $\{\beta^{(k)}\}$ including that used in accelerated proximal gradient (APG) for solving

SLOPE (Bogdan et al., 2015). In our numerical experiments section, we follow Bogdan et al. (2015) and set $\theta^{(-1)} = 1$,

$$\beta^{(k)} = \frac{\theta^{(k-1)} - 1}{\theta^{(k)}} \quad \text{with} \quad \theta^{(k)} = \frac{1 + \sqrt{1 + 4(\theta^{(k-1)})^2}}{2}, \quad \forall k \geq 0.$$

4.3 Global Convergence

This subsection is dedicated to the global convergence of pDCAe. We start with the level-coercivity of the DC function $F(\boldsymbol{W})$.

Lemma 10 *Let the predictor matrix $\boldsymbol{X}$ in problem (4) be full column rank. Then the DC function $F(\boldsymbol{W})$ in (9) is level-coercive in the sense that $\liminf_{\|\boldsymbol{W}\|_F \to \infty} \frac{F(\boldsymbol{W})}{\|\boldsymbol{W}\|_F} > 0$.*

Based on the level-coercivity as obtained in Lemma 10, the desired global convergence of pDCAe for solving problem (4) is stated as follows.

Theorem 11 *Suppose that the prediction matrix $\boldsymbol{X}$ is full column rank and let $\{\boldsymbol{W}^{(k)}\}$ be the sequence generated by Algorithm 1 for solving problem (4). Then the following properties hold:*

(a) The sequence $\{\boldsymbol{W}^{(k)}\}$ is bounded;

(b) $\lim_{k \to \infty} \|\boldsymbol{W}^{(k+1)} - \boldsymbol{W}^{(k)}\|_F = 0$;

(c) Every limit point $\bar{\boldsymbol{W}}$ of the sequence $\{\boldsymbol{W}^{(k)}\}$ is a stationary point of F in the sense that

$$\boldsymbol{O} \in \boldsymbol{X}^\top \boldsymbol{X} \bar{\boldsymbol{W}} + \partial P_{\boldsymbol{\lambda}}\big(\|\bar{\boldsymbol{W}}\|_r\big) - \partial \|\mathcal{M}_3(\mathcal{Y})^\top \boldsymbol{X} \bar{\boldsymbol{W}}\|_*.$$

The proofs of results in this subsection are provided in Appendix C.

5. Numerical Experiments

This section gives some experiments on synthetic data and a real human brain connection data. All numerical experiments are implemented in MATLAB (R2021a), running on a laptop with Intel Core i5-8265U CPU (1.60GHz) and 16 GB RAM.

5.1 Comparative Methods

We verify the performance of proposed TgSLOPE by comparing it with the following three approaches:

(a) TBMM: the block majorization minimization (BMM) algorithm proposed by Wei et al. (2021) to solve the TgSLOPE problem (4), whose corresponding iterative scheme is

$$\begin{cases} \boldsymbol{W}^{(k+1)} = \arg\min_{\boldsymbol{W}} \big\{ \frac{1}{2} \|\boldsymbol{W} - \boldsymbol{R}^{(k)}\|_F^2 + P_{\boldsymbol{\lambda}/\boldsymbol{L}}\big(\|\boldsymbol{W}\|_r\big) \big\}, \\ \boldsymbol{H}^{(k+1)} = \big(\boldsymbol{U}^{(k+1)}\big)_{[:,1:K]} \boldsymbol{V}^{(k+1)\top}, \end{cases}$$

where $\boldsymbol{R}^{(k)} = \boldsymbol{W}^{(k)} - \big(\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{W}^{(k)} - \boldsymbol{X}^\top \mathcal{M}_3(\mathcal{Y}) \boldsymbol{H}^{(k)}\big)/L$, $\boldsymbol{U}^{(k+1)}$ and $\boldsymbol{V}^{(k+1)}$ are from the SVD $\mathcal{M}_3(\mathcal{Y})^\top \boldsymbol{X} \boldsymbol{W}^{(k+1)} = \boldsymbol{U}^{(k+1)} \boldsymbol{D}^{(k+1)} \boldsymbol{V}^{(k+1)\top}$;

(b) TgLASSO: the group LASSO penalized LROD tensor regression approach, in which the pDCAe algorithm in Section 4 is used to solve

$$\min_{\boldsymbol{W},\boldsymbol{H}} \left\{ \frac{1}{2} \left\| \mathcal{M}_3(\mathcal{Y}) - \boldsymbol{X}\boldsymbol{W}\boldsymbol{H}^\top \right\|_F^2 + \lambda \sum_{j=1}^p \|\boldsymbol{W}_j\| : \boldsymbol{H}^\top \boldsymbol{H} = \boldsymbol{I}_K \right\},$$

where $\lambda > 0$ is the tuning parameter;

(c) TLRR: the LROD tensor regression (without sparse penalty), in which the pDCAe is applied to solve

$$\min_{\boldsymbol{W},\boldsymbol{H}} \left\{ \frac{1}{2} \left\| \mathcal{M}_3(\mathcal{Y}) - \boldsymbol{X}\boldsymbol{W}\boldsymbol{H}^\top \right\|_F^2 : \boldsymbol{H}^\top \boldsymbol{H} = \boldsymbol{I}_K \right\}.$$

In our numerical experiments, we simply adopt the stopping criterion proposed by Wen et al. (2018) for all approaches, i.e., $\frac{\|\boldsymbol{W}^{(k)} - \boldsymbol{W}^{(k+1)}\|_F}{\max\{\|\boldsymbol{W}^{(k)}\|_F, 1\}} \le \epsilon$ with some given tolerance $\epsilon > 0$.

5.2 Synthetic Data

In our numerical experiments on synthetic data, we adopt the measures including TgFDR defined in (6) and the tensor power (TP) to evaluate selection performance of the estimator $\hat{\mathcal{B}}$ generated by a given method. Here TP is defined as

$$\text{TP} = \mathbb{E}(T)/s, \text{ where } T = \sharp\{j \in [p] : \boldsymbol{B}_j^* \ne \boldsymbol{O}, \hat{\boldsymbol{B}}_j \ne \boldsymbol{O}\} \text{ and } s = \sharp\{j \in [p] : \boldsymbol{B}_j^* \ne \boldsymbol{O}\}.$$

For estimation accuracy, we evaluate the performance of $\hat{\mathcal{B}}$ in terms of the relative group estimate error (RgEE) and the mean squared error (MSE) defined respectively as

$$\text{RgEE} = \frac{\left\| \|\hat{\mathcal{B}}\|_f - \|\mathcal{B}^*\|_f \right\|^2}{\left\| \|\mathcal{B}^*\|_f \right\|^2}, \quad \text{MSE} = \left\| (\hat{\mathcal{B}} - \mathcal{B}^*) \times_3 \boldsymbol{X} \right\|_F^2 / n p_1 p_2.$$

Meanwhile, to evaluate the time efficiency of our proposed pDCAe algorithm, the CPU time (Time) is reported for each testing instance.

The ground truth is simulated via $\mathcal{B}^* = \mathcal{M}_3^{-1}(\boldsymbol{W}^* \boldsymbol{H}^{*\top}) \in \mathbb{R}^{p_1 \times p_2 \times p}$, in which $\boldsymbol{W}^* \in \mathbb{R}^{p \times K}$ is generated in a similar manner as in Brzyski et al. (2019) with s nonzero rows. Specifically, each nonzero row of $\boldsymbol{W}^*$ is generated from the uniform distribution $U[0.1, 1.1]$ and then we scale it such that $\|\boldsymbol{w}_j^*\| = a\sqrt{K}$ with $a = \sqrt{4\ln(p)/(1 - p^{-2/K}) - K}$. The explanation of such a special simulation procedure is presented in Appendix D. The column-orthogonal matrix $\boldsymbol{H}^* \in \mathbb{R}^{p_1 p_2 \times K}$ is simulated as the first K left singular vectors of a $p_1 p_2 \times p_1 p_2$ matrix with i.i.d. standard normal distributed entries. The response tensor $\mathcal{Y} \in \mathbb{R}^{p_1 \times p_2 \times n}$ is simulated by $\mathcal{Y} = \mathcal{B}^* \times_3 \boldsymbol{X} + \mathcal{E}$, where the entries of the noise tensor $\mathcal{E}$ are i.i.d. drawn from $N(0, 1)$.

We first verify the TgFDR controlling performance of TgSLOPE under the following two situations of the predictor matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$: (a) the orthogonal design, where $\boldsymbol{X}$ is generated from the orthogonalization for a $n \times p$ matrix with i.i.d. standard normal distributed entries; (b) the commonly used Gaussian random design, such as in Kong et al.
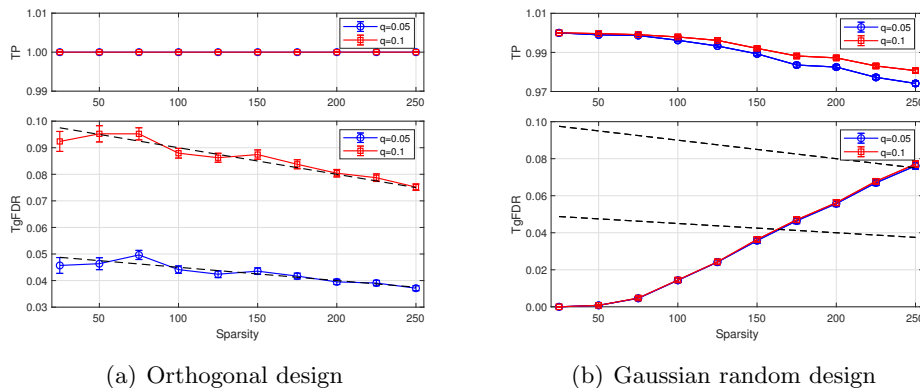
(a) Orthogonal design　　　　(b) Gaussian random design

Figure 4: Estimated TgFDR and TP with $n = 2000, p = 1000,\ p_1 = p_2 = 10, K = 20$ under the orthogonal and Gaussian situations of the matrix $\boldsymbol{X}$. Bars correspond to $\pm$SE (standard error), black dotted lines represent the 'nominal' TgFDR level $q \cdot (p - s)/p$.

(2020), where the predictor vectors $\boldsymbol{x}_i, i = 1, \cdots, n$ are i.i.d. drawn from Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{C})$, where $\boldsymbol{C}$ is a $p \times p$ covariance matrix with the element $\boldsymbol{C}_{j_1 j_2} = 0.5^{|j_1 - j_2|}$ for any $j_1, j_2 \in [p]$. The orthogonal situation works with the regularization parameter sequence defined in (7). For Gaussian random design, we select regularization parameters of TgSLOPE according to Procedure 2 of Brzyski et al. (2019).

Set $n = 2000, p = 1000, p_1 = p_2 = 10, K = 20$ and sparsity $s = 25 : 25 : 250$. We perform 100 independent replications for each sparsity and target TgFDR level ($q = 0.05, 0.1$). As shown in Figure 4 (a), TgSLOPE maintains a comparable TgFDR with the 'nominal' level and reports the estimated tensor power TP $=1$ for all testing instances. For the Gaussian situation, TgSLOPE reports the relatively low TgFDR with the strong power, especially for the sparse cases as shown in Figure 4 (b). The reported TgFDR of TgSLOPE does not exceed the 'nominal' level of the case $q = 0.1$ in all sparsity settings except for the case $s = 250$.

In the following numerical comparisons, we consider the Gaussian random situation in which the entries of $\boldsymbol{X}$ are i.i.d. drawn from $N(0, 1/n)$, and test the effect of sparsity s, model size p and LROD rank K respectively under the target TgFDR level $q = 0.05$. For TgLASSO, the tuning parameter is selected by 5-fold cross-validation.

Sparsity effect: set $n = 3000, p = 1000, p_1 = p_2 = 10$ and $K = 20$. We test the performance of four approaches under various sparsity with $s = 25 : 25 : 250$. Simulation results report TP $=1$ for all competitors in each testing instance. In addition, average results based on 100 independent replications have been shown in Figure 5. (i) Figure 5 (a) illustrates that TgSLOPE and TBMM have significant superiority in terms of TgFDR, with TgFDR below 'nominal' for all testing sparsity. (ii) TgLASSO fails to control TgFDR, although it reports the relatively small RgEE and MSE as shown in Figure 5 (b) and (c). (iii) Figure 5 (d) shows that TgSLOPE reports a relatively short CUP time, especially in the
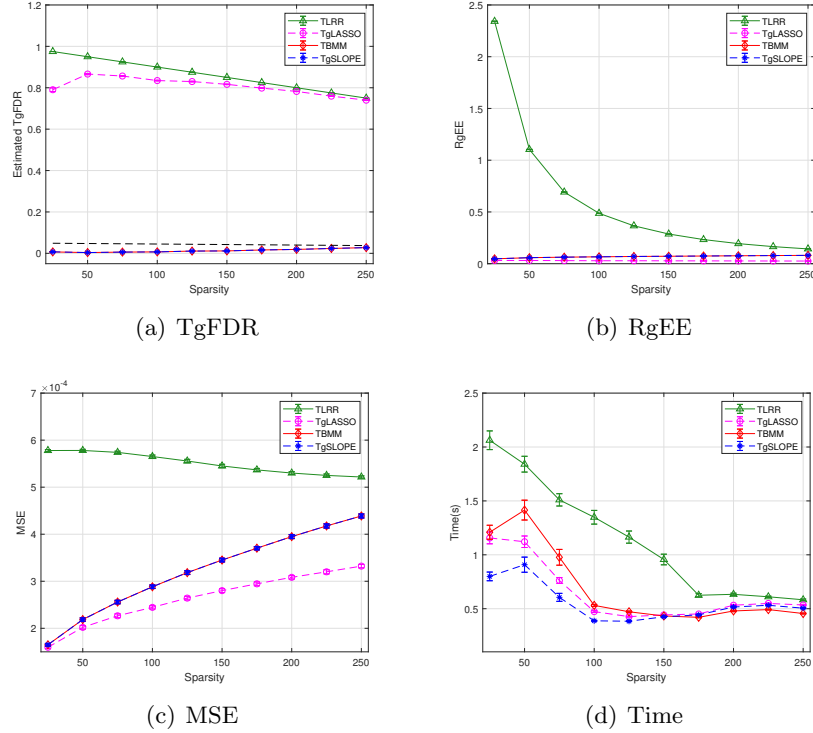
14

Figure 5: Average results with various sparsity under $n = 3000, p = 1000, p_1 = p_2 = 10, K = 20$ and for the Gaussian random design situation of $\boldsymbol{X}$. Bars correspond to $\pm$SD (standard deviation). In (a), black dotted lines represent the 'nominal' TgFDR level $q \cdot (p - s)/p$.

cases of small sparsity. (iv) It is intuitive that those four methods tend to have comparable performance as the sparsity increases.

Model size effect: in this example, we test the effect of model size p comparing among four approaches. Set $n = 3000, p_1 = p_2 = 10, K = 20$ and the sparsity $s = 0.02p$ with model size $p = 2000, 4000, 6000$. In each testing instance, all these competitors report TP =1. Moreover, Table 1 collects the average results based on 100 independent replications for each model size. (i) As suggested in Table 1, TLRR gives the worst performance. For the other three approaches, all of the four evaluation metrics tend to become larger as the model size grows. (ii) Similar to the results shown in Figure 5, TgLASSO fails to control TgFDR, while TgSLOPE and TBMM can maintain the TgFDR below the 'nominal' TgFDR level for all testing model sizes, with predictable sacrifice on the estimation accuracy. (iii) We can also see from Table 1 that TgSLOPE gives the smaller TgFDR than TBMM for the model size $p = 6000$, and it reports the least CPU time for all cases.

LROD rank effect: we now examine the impact of LROD rank on the four approaches under $p_1 = p_2 = 10$ and $p_1 = p_2 = 20$, respectively. Set $n = 1000, p = 2000$ and the sparsity $s = 0.02p$. It is known from Kolda and Bader (2009) that LROD rank $K \leq \{p_1p_2, pp_1, pp_2\}$, then we set LROD rank $K = 5 : 5 : 50$. All testing instances are simulated based on 100

Size	Method	TgFDR	RgEE	MSE	Time(s)
$p = 2000$	TLRR	0.980	5.10e-0 (6.24e-2)	7.77e-4 (1.89e-6)	3.325
	TgLASSO	0.891	3.44e-2 (3.97e-3)	1.92e-4 (4.45e-6)	1.566
	TBMM	0.006	5.59e-2 (3.91e-3)	2.02e-4 (2.73e-6)	2.242
	TgSLOPE	0.007	5.59e-2 (3.91e-3)	2.02e-4 (2.76e-6)	1.081
$p = 4000$	TLRR	0.980	3.92e-0 (4.26e-2)	9.28e-4 (2.45e-6)	10.163
	TgLASSO	0.921	3.50e-2 (2.16e-3)	2.40e-4 (4.70e-6)	3.727
	TBMM	0.008	6.46e-2 (3.28e-3)	2.70e-4 (3.48e-6)	6.104
	TgSLOPE	0.008	6.45e-2 (3.27e-3)	2.70e-4 (3.49e-6)	2.962
$p = 6000$	TLRR	0.980	1.69e-0 (1.71e-2)	9.19e-4 (2.94e-6)	15.146
	TgLASSO	0.924	3.66e-2 (2.17e-3)	2.77e-4 (5.16e-6)	6.179
	TBMM	0.025	6.91e-2 (3.78e-3)	3.24e-4 (4.02e-6)	9.108
	TgSLOPE	0.014	6.88e-2 (3.75e-3)	3.24e-4 (4.01e-6)	5.666

Table 1: Average results with different model size under $n = 3000, p_1 = p_2 = 10, K = 20, s = 0.02p$ for the Gaussian random design situation of $\boldsymbol{X}$. Standard deviations are presented in brackets.

independent replications. Simulation results report that TP =1 for all the competitors. Figure 6 depicts changes of the four evaluation metrics with the increase of LROD rank. (i) Our proposed TgSLOPE procedure works well in terms of TgFDR, especially for small LROD ranks. See, e.g., as shown in Figure 6 (a) and (b), TgSLOPE and TBMM report the lower TgFDR than the 'nominal' level as $K \leq 25$. (ii) Figure 6 (g) and (h) illustrate that TgSLOPE reports a relatively short CPU time, especially compared with TBMM. For example, as $p_1 = p_2 = 20$ and $K \leq 35$, the CPU time reported by TgSLOPE is no more than 1/5 of that of TBMM. (iii) For the fixed tensor size, we can see from Figure 6 (a) and (b) that the TgFDR tends to become larger as the LROD rank increases. In addition, with the increase of the LROD rank, Figure 6 (c) and (d) show that RgEE reduces, while Figure 6 (e) and (f) show that MSE grows. This may give us some inspiration to choose a moderate LROD rank.

5.3 Human Brain Connection Data

In this subsection, we test our proposed TgSLOPE comparing with the other three approaches on a real human brain connection (HBC) data from the Human Connectome Project (HCP), which aims to build a network map between the anatomical and functional connectivity within healthy human brains (Van Essen et al., 2013). The preprocessed HBC data set is provided by Hu et al. (2022), in which the response is a 68×68 binary matrix with entries encoding the presence or absence of fiber connections between 68 brain regions-of-interest, the predictor matrix is collected from different personal features for each observed individual. After removing those missing values, the HBC data set consists of 111 individuals and 549 personal features, including gender, age, etc. (The HBC data can be found at https://wiki.humanconnectome.org/display/PublicData/). For HBC analysis, the predictor matrix is normalized to have unit column vectors. We choose the tuning parameters
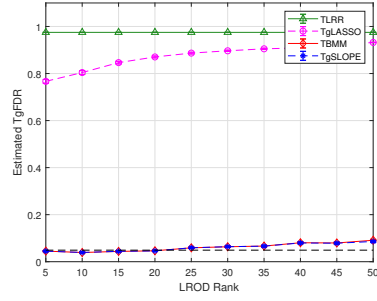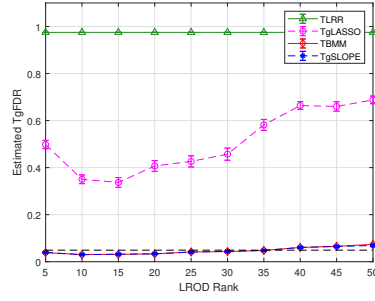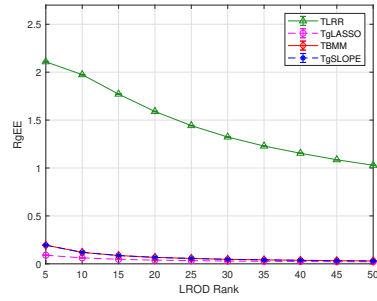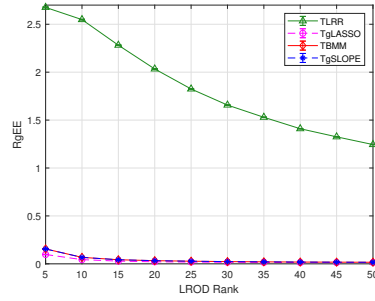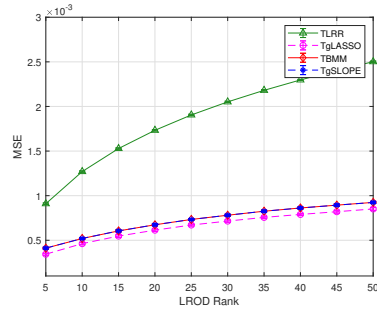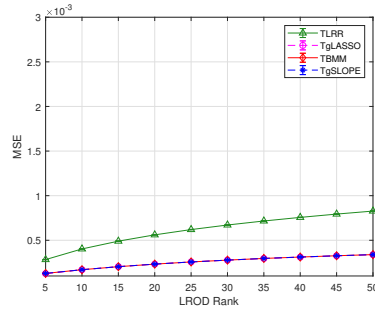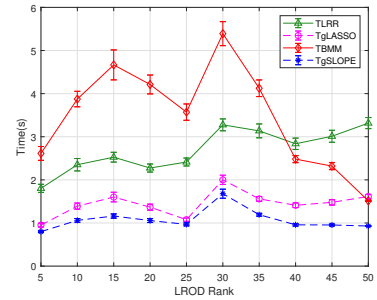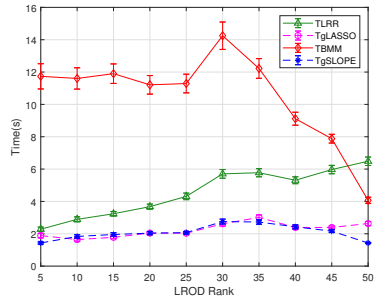
(a) TgFDR

(b) TgFDR

(c) RgEE

(d) RgEE

(e) MSE

(f) MSE

(g) Time

(h) Time

Figure 6: Simulation results with different LROD rank under $n = 1000, p = 2000, s = 0.02p$ for the Gaussian random design situation of $\boldsymbol{X}$. Left column: $p_1 = p_2 = 10$, right column: $p_1 = p_2 = 20$. In (a) and (b), black dotted lines represent the 'nominal' TgFDR level $q \cdot (p - s)/p$.
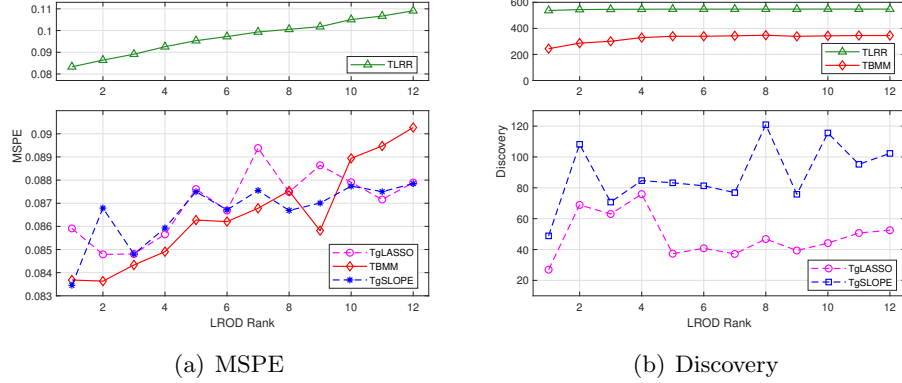
(a) MSPE

(b) Discovery

Figure 7: Average results with the different LROD rank based on 20 randomly 8:2 splits for HBC data set.

via a BIC-type criterion on the whole data set, which minimizes

$$\text{BIC} = \|\mathcal{Y} - \hat{\mathcal{B}} \times_3 \boldsymbol{X}\|_F^2 + \left(\left\|\|\hat{\mathcal{B}}\|_f\right\|_0 + p_1 + p_2\right) K \log(np_1p_2).$$

In addition, the number of discovered features (termed as Discovery), the mean squared prediction error (MSPE) and the CPU time are adopted to evaluate the performance of four competitors. Here, Discovery and MSPE are defined respectively as

$$\text{Discovery} = \sharp\{j \in [p] : (\hat{\boldsymbol{B}}_{training})_j \neq \boldsymbol{O}\},$$

$$\text{MSPE} = \|\mathcal{Y}_{test} - \hat{\mathcal{B}}_{training} \times_3 \boldsymbol{X}_{test}\|_F^2 / p_1 p_2 n_{test},$$

where $\hat{\mathcal{B}}_{training}$ is the estimator for the training set, $\boldsymbol{X}_{test}$, $\mathcal{Y}_{test}$ and n_{test} are the predictor matrix, response tensor and the sample size of the testing set, respectively.

For HBC data, the experimental results in Zhang et al. (2016) have revealed that the whole-brain functional magnetic resonance imaging (fMRI) signals can be well characterized by sparse representations, which is supportive on the interpretability with fewer significant features in the HBC data analysis. In this regard, it may be reasonable to recognize the performance superiority of a method with the smaller Discovery and the comparable or even less MSPE. We will test the performance of our proposed approach in such a sense with comparison to other three methods.

Notably, the rank of true coefficient tensor is unknown in real data applications. We first test the performance of four approaches under different LROD ranks. Set the tolerance for termination in algorithms to be $\epsilon = 10^{-4}$. For each LROD rank selected from 1 to 12, HBC data set is randomly divided into 80 percent of training set and 20 percent of testing set 20 times. The average numerical results are depicted in Figure 7. We can see from Figure 7 (a) that small LROD ranks would be sufficient for an acceptable prediction errors of all methods, which suggests us to fit the true rank with a relatively small value. In addition, as shown in Figure 7 (b), TLRR fails to feature selection for all rank testing instances, while

18

ϵ	Method	MSPE	Discovery	Time(s)
10^{-4}	TLRR	0.0837	536.30	2.766
	TgLASSO	0.0848	28.45	9.396
	TBMM	0.0840	242.50	7.137
	TgSLOPE	0.0838	47.95	8.171
10^{-5}	TLRR	0.0831	536.20	9.363
	TgLASSO	0.0844	27.70	30.561
	TBMM	0.0833	121.90	92.278
	TgSLOPE	0.0833	47.15	29.724
10^{-6}	TLRR	0.0836	537.20	25.629
	TgLASSO	0.0861	26.70	111.271
	TBMM	0.0838	73.65	668.040
	TgSLOPE	0.0838	49.60	99.964

Table 2: Average results with BIC selected LROD rank and different tolerance ϵ based on 20 randomly 8:2 splits for HBC data set.

the other three methods all report the sparse numerical approximate estimates reflected by relatively small values of Discovery, with the smaller MSPE compared to TLRR. This further supports the reasonability of taking the group sparsity of coefficient tensor into consideration for HBC data set.

To test the comparative performance of four methods, we next choose the combination of LROD rank and regularization parameters via the BIC criterion. Table 2 collects the average results with different tolerances $\epsilon = \{10^{-4}, 10^{-5}, 10^{-6}\}$ based on 20 independent replications, which demonstrate the superior performance of TgSLOPE among all comparative methods. Specifically, as shown in Table 2, TLRR reports the competitive MSPE with TgSLOPE, but unsurprisingly fails to feature selection for all tolerance settings. In addition, TgLASSO gives the larger MSPE than TgSLOPE in all testing instances, although it has the fewest discovered features.

For the method of TBMM, it solves the same optimization model (4) by the alternating minimization scheme. While our proposed TgSLOPE method solves problem (4) by pDCAe based on the DC reformulation (8). We can see from Table 2 that these two algorithms report the similar MSPE in different tolerance settings. While the Discovery generated by TBMM is larger than that of TgSLOPE, and the difference of the values for Discovery decreases as the tolerance ϵ of the numerical algorithms reduces, as shown in Table 2 Columns "ϵ" and "Discovery". In addition, TgSLOPE reports the less CPU time than TBMM, especially for the small tolerance $\epsilon = 10^{-6}$. One possible explanation for the superior efficiency of TgSLOPE is that pDCAe has more power of achieving numerical approximate coefficient tensor solution with higher estimate accuracy, leading to better performance on time efficiency and group sparsity reflected by Discovery. The comparison results on Discovery, MSPE and CPU Time of both algorithms with varying ϵ have been collected in the boxplots in Figure 8.
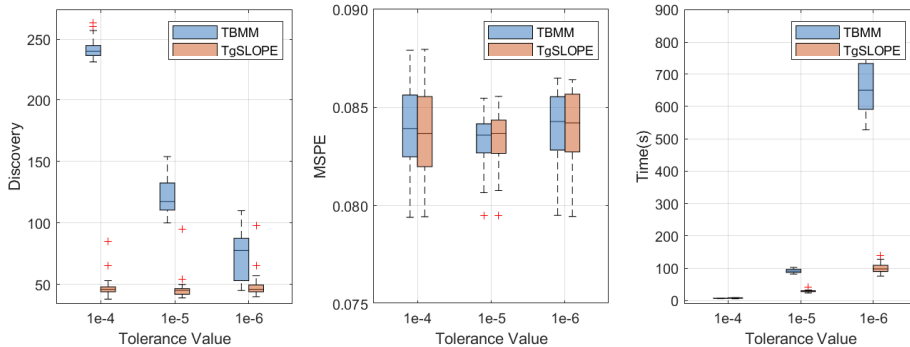
Figure 8: Boxplot for results of TBMM and TgSLOPE with BIC selected LROD rank and different tolerance ϵ based on 20 randomly 8:2 splits for HBC data set.

6. Conclusions

In this article we propose a sparse and low-rank tensor regression method, which optimizes a gSLOPE penalized low-rank, orthogonally decomposable tensor minimization problem. Under the assumption of column-orthogonality for the predictor matrix, we show that our proposed TgSLOPE procedure controls TgFDR at a pre-set level, and achieves the asymptotically minimax convergence with respect to the produced estimation risk. This provides theoretical guarantees for feature selection and coefficient estimation in finite samples. Moreover, a globally convergent pDCAe algorithm is applied to solve the TgSLOPE estimator by constructively reformulating our TgSLOPE problem into a DC program. Numerical experiments verify the superiority of our method in terms of TgFDR control, estimation accuracy and CPU time against three state-of-the-art approaches.

For the TgSLOPE method, it would be interesting to investigate statistical properties including TgFDR control and estimate accuracy in more general cases besides the orthogonal design. Moreover, Luo et al. (2019) propose a sparse semismooth Newton-based augmented Lagrangian method (Newt-ALM) to solve the SLOPE model of the classical linear regression (Bogdan et al., 2015) and show that Newt-ALM offers a notable computational advantage in the high-dimensional settings comparing with the first-order algorithms, such as APG and alternating direction method of multipliers (ADMM). How to design the robust and highly efficient Newton-type algorithm for TgSLOPE based models is also one of our research topic in the future.

Acknowledgments

Appendix A. Proof of Proposition 2

To check the identifiability for the frontal slice sparse and LROD tensor coefficient $\mathcal{B} \in \mathbb{B}_T$ defined in (5), we need to state that the equation system produced by tensor regression model (2), i.e., $\bar{\mathcal{Y}} = \mathcal{B} \times_3 \boldsymbol{X}$, has a unique solution $\mathcal{B}^*$ satisfying $\mathcal{B}^* \in \mathbb{B}_T$. Under the uniqueness guarantee $k(\boldsymbol{W}^*) \geq 2$ of the LROD tensor decomposition $\mathcal{B}^* = [\![\boldsymbol{U}^*, \boldsymbol{V}^*, \boldsymbol{W}^*]\!]$ (Kruskal, 1977, Theorem 4b), it suffices to show that

$$\mathcal{B}^* = \underset{\mathcal{B} \in \mathbb{R}^{p_1 \times p_2 \times p}}{\arg\min} \left\| \|\mathcal{B}\|_f \right\|_0, \quad \text{s.t. } \bar{\mathcal{Y}} = \mathcal{B} \times_3 \boldsymbol{X}$$

is a unique solution. Following the matricized form of a tensor, the above optimization problem can be equivalent to

$$\underset{\boldsymbol{B} \in \mathbb{B}_M}{\min} \left\| \|\boldsymbol{B}\|_r \right\|_0, \tag{13}$$

where $\mathbb{B}_M = \{\mathcal{B} \in \mathbb{R}^{p \times p_1 p_2} : \mathcal{M}_3(\bar{\mathcal{Y}}) = \boldsymbol{X}\boldsymbol{B}\}$. Thus, it suffices to prove that problem (13) has a unique solution $\boldsymbol{B}^* = \mathcal{M}_3^{-1}(\mathcal{B}^*)$ under the condition that $\left\| \|\boldsymbol{B}^*\|_r \right\|_0 \leq k/2$, where k is the k-rank of the matrix $\boldsymbol{X}$.

Assume on the contrary that $\tilde{\boldsymbol{B}} \in \mathbb{B}_M$ is another optimal solution of (13), which gives that $\left\| \|\boldsymbol{B}^*\|_r \right\|_0 = \left\| \|\tilde{\boldsymbol{B}}\|_r \right\|_0$, $\boldsymbol{\Delta} = \boldsymbol{B}^* - \tilde{\boldsymbol{B}} \neq \boldsymbol{O}$ and $\boldsymbol{X}\boldsymbol{\Delta} = \boldsymbol{O}$. This further indicates that there exist $\left\| \|\boldsymbol{\Delta}\|_r \right\|_0$ columns of $\boldsymbol{X}$ which are linearly dependent. By virtue of the definition of the k-rank of $\boldsymbol{X}$, it yields that

$$k < \left\| \|\boldsymbol{\Delta}\|_r \right\|_0 \leq \left\| \|\boldsymbol{B}^*\|_r \right\|_0 + \left\| \|\tilde{\boldsymbol{B}}\|_r \right\|_0 \leq k,$$

where the second inequality is from the fact that ℓ_0-norm obeys the triangle inequality, and the last inequality from $\left\| \|\tilde{\boldsymbol{B}}\|_r \right\|_0 = \left\| \|\boldsymbol{B}^*\|_r \right\|_0 \leq k/2$. Contradiction arrives and the proof is completed. ∎

Appendix B. Proofs for Section 3

B.1 Proof of Theorem 5

Assume that $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ is a local minimizer of the TgSLOPE problem (4). Let $\hat{\boldsymbol{H}}^\perp \in \mathbb{R}^{p_1 p_2 \times (p_1 p_2 - K)}$ such that $[\hat{\boldsymbol{H}}, \hat{\boldsymbol{H}}^\perp]$ is orthogonal. Then the loss function of problem (4)

$$L(\boldsymbol{W}, \hat{\boldsymbol{H}}) = \left\| [\mathcal{M}_3(\mathcal{Y}) - \boldsymbol{X}\boldsymbol{W}\hat{\boldsymbol{H}}^\top][\hat{\boldsymbol{H}}, \hat{\boldsymbol{H}}^\perp] \right\|_F^2$$
$$= \left\| \mathcal{M}_3(\mathcal{Y})\hat{\boldsymbol{H}} - \boldsymbol{X}\boldsymbol{W} \right\|_F^2 + \left\| \mathcal{M}_3(\mathcal{Y})\hat{\boldsymbol{H}}^\perp \right\|_F^2.$$

Together with the column-orthogonality of the predictor matrix $\boldsymbol{X}$, we have

$$\hat{\boldsymbol{W}} = \underset{\boldsymbol{W} \in \mathbb{R}^{p \times K}}{\arg\min} \left\{ \frac{1}{2} \left\| \hat{\boldsymbol{Y}} - \boldsymbol{W} \right\|_F^2 + P_{\boldsymbol{\lambda}}(\|\boldsymbol{W}\|_r) \right\}, \tag{14}$$

where $\hat{\boldsymbol{Y}} = \boldsymbol{X}^\top \mathcal{M}_3(\mathcal{Y})\hat{\boldsymbol{H}} \in \mathbb{R}^{p \times K}$. It follows from (2) that

$$\hat{\boldsymbol{Y}} = \boldsymbol{W}^* \boldsymbol{H}^{*\top} \hat{\boldsymbol{H}} + \boldsymbol{X}^\top \mathcal{M}_3(\mathcal{E})\hat{\boldsymbol{H}},$$

which implies that $\hat{\boldsymbol{Y}}$ follows the matrix normal distribution $\mathcal{N}(\boldsymbol{W}^* \boldsymbol{H}^{*\top} \hat{\boldsymbol{H}}, \sigma^2 \boldsymbol{I}_p \otimes \boldsymbol{I}_K)$. Note that (14) is a convex problem with strongly convex, piecewise linear-quadratic objective function and hence admits a unique optimal solution. Similar to the proximal operator for gSLOPE (Brzyski et al., 2019), the optimization problem (14) can be solved in two steps

$$\begin{cases} \hat{\boldsymbol{\eta}} = \arg\min_{\boldsymbol{\eta}} \left\{ \frac{1}{2} \sum_{j=1}^p \left(\|\hat{\boldsymbol{y}}_j\| - \eta_j \right)^2 + P_{\boldsymbol{\lambda}}(\boldsymbol{\eta}) \right\}, \\ \hat{\boldsymbol{W}}_j = \hat{\eta}_j \frac{\hat{\boldsymbol{y}}_j}{\|\hat{\boldsymbol{y}}_j\|}, j \in [p]. \end{cases} \tag{15}$$

Define $V_{\boldsymbol{\eta}} = \sharp\{j \in [p] : \boldsymbol{W}_j^* = \boldsymbol{0}, \hat{\eta}_j \neq 0\}, R_{\boldsymbol{\eta}} = \sharp\{j \in [p] : \hat{\eta}_j \neq 0\}$. It suffices to show that

$$\mathbb{E}\left[\frac{V_{\boldsymbol{\eta}}}{\max\{R_{\boldsymbol{\eta}}, 1\}} \right] \leq q \cdot \frac{p-s}{p}.$$

Without loss of generality, assume that $\boldsymbol{W}_j^* = \boldsymbol{0}$ for $0 \leq j \leq p-s$ and $\boldsymbol{W}_j^* \neq \boldsymbol{0}$ otherwise. By the definition of TgFDR, we derive that

$$\begin{aligned} \mathbb{E}\left[\frac{V_{\boldsymbol{\eta}}}{\max\{R_{\boldsymbol{\eta}}, 1\}} \right] &= \sum_{r=1}^p \frac{1}{r} \mathbb{E}\left[V_{\boldsymbol{\eta}} \mathbb{1}_{\{R_{\boldsymbol{\eta}} = r\}} \right] \\ &= \sum_{r=1}^p \frac{1}{r} \sum_{j=1}^{p-s} \mathbb{E}\left[\mathbb{1}_{\{\hat{\eta}_j \neq 0\}} \mathbb{1}_{\{R_{\boldsymbol{\eta}} = r\}} \right] \\ &= \sum_{r=1}^p \frac{1}{r} \sum_{j=1}^{p-s} \Pr\left(\hat{\eta}_j \neq 0, R_{\boldsymbol{\eta}} = r \right), \end{aligned} \tag{16}$$

where $\mathbb{1}_{\{\cdot\}} = 1$ if the event occurs, and $\mathbb{1}_{\{\cdot\}} = 0$ otherwise. Next we focus on the events $\{\hat{\eta}_j \neq 0, R_{\boldsymbol{\eta}} = r\}, j = 1, \ldots, p-s, r = 1, \ldots, p$. Denote $\overline{\boldsymbol{Y}} = (\hat{\boldsymbol{y}}_1, \ldots, \hat{\boldsymbol{y}}_{j-1}, \hat{\boldsymbol{y}}_{j+1}, \ldots, \hat{\boldsymbol{y}}_p)^\top \in \mathbb{R}^{(p-1) \times K}$, $\overline{\boldsymbol{\lambda}} = (\lambda_2, \ldots, \lambda_p)^\top \in \mathbb{R}^{p-1}$. Applying the simplified TgSLOPE problem (14) to $\overline{\boldsymbol{Y}}$ with tuning parameter $\overline{\boldsymbol{\lambda}}$, the optimization problem is given by

$$\overline{\boldsymbol{W}} = \arg\min_{\boldsymbol{W} \in \mathbb{R}^{(p-1) \times K}} \left\{ \frac{1}{2} \sum_{j=1}^{p-1} \|\overline{\boldsymbol{y}}_j - \boldsymbol{W}_j\|^2 + P_{\overline{\boldsymbol{\lambda}}}(\|\boldsymbol{W}\|_r) \right\}.$$

Define $\overline{R}^j = \sharp\{j \in [p-1] : \overline{\boldsymbol{W}}_j \neq \boldsymbol{0}\}$. It follows from Lemmas E.6 and E.7 in Brzyski et al. (2019) that

$$\left\{ \|\hat{\boldsymbol{Y}}\|_r : \hat{\eta}_j \neq 0, R_{\boldsymbol{\eta}} = r \right\} \subset \left\{ \|\hat{\boldsymbol{Y}}\|_r : \|\hat{\boldsymbol{y}}_j\| > \lambda_r, \overline{R}^j = r - 1 \right\},$$

where $\|\hat{\boldsymbol{Y}}\|_r = (\|\hat{\boldsymbol{y}}_1\|, \ldots, \|\hat{\boldsymbol{y}}_p\|)^\top$ with $\|\hat{\boldsymbol{y}}_j\|^2 \sim \chi_K^2 \left(\|\boldsymbol{W}_j^* \boldsymbol{H}^{*\top} \hat{\boldsymbol{H}}\|^2 \right)$, $j = 1, \ldots, p$. Then we have

$$\begin{aligned} \Pr(\hat{\eta}_j \neq 0, R_{\boldsymbol{\eta}} = r) &\leq \Pr(\|\hat{\boldsymbol{y}}_j\| > \lambda_r, \overline{R}^j = r - 1) \\ &= \Pr(\|\hat{\boldsymbol{y}}_j\|/\sigma > \lambda_r^*) \Pr(\overline{R}^j = r - 1) \\ &\leq \frac{q \cdot r}{p} \Pr(\overline{R}^j = r - 1), \end{aligned}$$

where $\lambda_r^* = F_{\chi_K}^{-1}(1 - q \cdot r/p)$, the equality is due to the independence between $\|\hat{\boldsymbol{y}}_j\|$ and $\overline{R}^j$, the last inequality follows from the definition of the tuning parameters in (7). Therefore, TgFDR in (16) can be bounded by

$$\mathbb{E}\left[\frac{V_{\boldsymbol{\eta}}}{\max\{R_{\boldsymbol{\eta}}, 1\}}\right] \leq \sum_{r=1}^{p} \frac{1}{r} \sum_{j=1}^{p-s} \frac{q \cdot r}{p} \Pr(\overline{R}^j = r - 1)$$

$$= \sum_{j=1}^{p-s} \frac{q}{p} \sum_{r=1}^{p} \Pr(\overline{R}^j = r - 1) = q \cdot \frac{p - s}{p}.$$

The proof is completed. ∎

B.2 Proof of Theorem 7

Let $(\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}})$ be an optimal solution of the problem (4). Then $\hat{\mathcal{B}} = \mathcal{M}_3^{-1}(\hat{\boldsymbol{W}}\hat{\boldsymbol{H}}^\top)$. Under the column-orthogonality assumption on the predictor matrix $\boldsymbol{X}$, we know that the statistically equivalent model of (2) is

$$\widetilde{\boldsymbol{Y}} = \mathcal{M}_3(\mathcal{B}^*) + \boldsymbol{X}^\top \mathcal{M}_3(\mathcal{E}), \tag{17}$$

which has the distribution $\mathcal{N}(\mathcal{M}(\mathcal{B}^*), \sigma^2 \boldsymbol{I}_p \otimes \boldsymbol{I}_K)$. Denote $\widetilde{\mathbb{B}}$ as the set of all coefficient tensors for which only the elements in the first column of the mode-3 unfolded matrix are possibly nonzero and the rest are fixed to be zero. Let $\widetilde{\mathbb{B}}_s = \mathbb{B}_s \cap \widetilde{\mathbb{B}}$. Then, for any $\mathcal{B}^* \in \widetilde{\mathbb{B}}_s$, (17) is reduced to a general Gaussian sequence model with length p and sparsity at most s. As $s/p \to 0$, this sequence model has minimax risk $(1 + o(1))2\sigma^2 s \log(p/s)$ (Donoho and Johnstone, 1994). Thus, we have

$$\sup_{\mathcal{B}^* \in \widetilde{\mathbb{B}}_s} \mathbb{E}\left\|\mathcal{M}(\hat{\mathcal{B}}) - \mathcal{M}(\mathcal{B}^*)\right\|_F^2 \sim (1 + o(1))2\sigma^2 s \log(p/s),$$

which yields that

$$\sup_{\mathcal{B}^* \in \mathbb{B}_s} \mathbb{E}\|\hat{\mathcal{B}} - \mathcal{B}^*\|_F^2 \geq (1 + o(1))2\sigma^2 s \log(p/s).$$

We next show that the ℓ_2-loss which measures the deviation of the TgSLOPE estimator from the ground truth $\mathcal{B}^*$ is bounded above by $(1 + o(1))2\sigma^2 s \log(p/s)$. For simplicity, we assume that $\|\boldsymbol{w}_j^*\| \neq 0$ for $j \leq s$ and $\|\boldsymbol{w}_j^*\| = 0$ otherwise. Denote $\mu_j = \|\boldsymbol{w}_j^*\|$ and $\nu_j = \|\hat{\boldsymbol{H}}^\top \boldsymbol{H}^* \boldsymbol{w}_j^*\|$. Then, it follows from the proof of Theorem 5 that $\hat{\boldsymbol{W}}$ can be obtained

by the two step format (15), in which $\|\hat{\boldsymbol{y}}_j\|^2 \sim \chi_K^2(\nu_j^2)$, $j = 1, \ldots, p$. The ℓ_2-loss is

$$
\mathbb{E}\|\hat{\mathcal{B}} - \mathcal{B}^*\|_F^2 \overset{(a)}{=} \mathbb{E}\|\hat{\boldsymbol{W}}\hat{\boldsymbol{H}}^\top - \boldsymbol{W}^*\boldsymbol{H}^{*\top}\|_F^2 = \mathbb{E}\left[\sum_{j=1}^p \|\hat{\boldsymbol{H}}\hat{\boldsymbol{w}}_j - \boldsymbol{H}^*\boldsymbol{w}_j^*\|^2\right]
$$

$$
\overset{(b)}{\leq} \mathbb{E}\left[\sum_{j=1}^p (\|\hat{\boldsymbol{H}}\hat{\boldsymbol{w}}_j\| + \|\boldsymbol{H}^*\boldsymbol{w}_j^*\|)^2\right] \overset{(c)}{=} \mathbb{E}\left[\sum_{j=1}^p (\|\hat{\boldsymbol{w}}_j\| + \|\boldsymbol{w}_j^*\|)^2\right],
$$

$$
= \mathbb{E}\left[\sum_{j=1}^s (\|\hat{\boldsymbol{w}}_j\| + \|\boldsymbol{w}_j^*\|)^2\right] + \mathbb{E}\left[\sum_{j=s+1}^p \|\hat{\boldsymbol{w}}_j\|^2\right]
$$

$$
= \underbrace{\mathbb{E}\left[\sum_{j=1}^s (\hat{\eta}_j + \mu_j)^2\right]}_{E1} + \underbrace{\mathbb{E}\left[\sum_{j=s+1}^p \hat{\eta}_j^2\right]}_{E2},
$$

where (a) comes from $\hat{\mathcal{B}} = \mathcal{M}_3^{-1}(\hat{\boldsymbol{W}}\hat{\boldsymbol{H}}^\top)$, (b) is due to the triangle inequality and (c) follows from the column-orthogonality of $\hat{\boldsymbol{H}}$ and $\boldsymbol{H}$. Thus, it suffices to show

$$
E1 \leq (1 + o(1))2\sigma^2 s \log(p/s) \quad and \quad E2 = o(1)2\sigma^2 s \log(p/s).
$$

To proceed, define random variables $\psi_j^2 = \psi_{j,1}^2 + \psi_{j,2}^2 + \cdots + \psi_{j,K}^2$ and $\phi_j^2 = (\psi_{j,1} + \nu_j)^2 + \psi_{j,2}^2 + \cdots + \psi_{j,K}^2$ with i.i.d. $\psi_{j,k} \sim N(0,1)$, $j \in [p], k \in [K]$. Then $\psi_j^2 \sim \chi_K^2$ and $\phi_j^2 \sim \chi_K^2(\nu_j^2)$ for all $j \in [p]$. Denoting $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_p)^\top$, we have

$$
\mathbb{E}\|\hat{\boldsymbol{\eta}}_{[1:s]} + \boldsymbol{\mu}_{[1:s]}\|^2 = \mathbb{E}\|\hat{\boldsymbol{\eta}}_{[1:s]} - \boldsymbol{\phi}_{[1:s]} + \boldsymbol{\phi}_{[1:s]} + \boldsymbol{\mu}_{[1:s]}\|^2
$$
$$
\leq \mathbb{E}\left(\|\hat{\boldsymbol{\eta}}_{[1:s]} - \boldsymbol{\phi}_{[1:s]}\| + \|\boldsymbol{\phi}_{[1:s]} + \boldsymbol{\mu}_{[1:s]}\|\right)^2 \tag{18}
$$
$$
\leq \|\boldsymbol{\lambda}_{[1:s]}\|^2 + \mathbb{E}\|\boldsymbol{\phi}_{[1:s]} + \boldsymbol{\mu}_{[1:s]}\|^2 + 2\|\boldsymbol{\lambda}_{[1:s]}\|\mathbb{E}\|\boldsymbol{\phi}_{[1:s]} + \boldsymbol{\mu}_{[1:s]}\|,
$$

where the last inequality is obtained by $\|\hat{\boldsymbol{\eta}}_{[1:s]} - \boldsymbol{\phi}_{[1:s]}\| \leq \|\boldsymbol{\lambda}_{[1:s]}\|$, owing to Fact 3.3 of Su and Candès (2016) with conditions ϕ_j^2 and $\|\hat{\boldsymbol{y}}_j\|^2$ are i.i.d. for any $j \leq s$. Moreover, it follows from Inglot (2010) that as $s/p \to 0$, $F_{\chi_K}^{-1}\left(1 - q \cdot j/p\right) \sim \sqrt{2\log(p/(q \cdot j))}$ for all $j \leq s$, which yields

$$
\|\boldsymbol{\lambda}_{[1:s]}\|^2 \sim 2\sigma^2 s \log(p/s). \tag{19}
$$

Then, it is easy to see that

$$
|\phi_j + \mu_j|^2 = \left|\sqrt{(\psi_{j,1} + \nu_j)^2 + \psi_{j,2}^2 + \cdots + \psi_{j,K}^2} + \mu_j\right|^2
$$
$$
\leq \left|\sqrt{\psi_{j,2}^2 + \cdots + \psi_{j,K}^2} + |\psi_{j,1}| + \nu_j + \mu_j\right|^2 \tag{20}
$$
$$
\leq \left(\sqrt{2(\psi_{j,1}^2 + \psi_{j,2}^2 + \cdots + \psi_{j,K}^2)} + \nu_j + \mu_j\right)^2
$$
$$
\leq 4\psi_j^2 + 2(\nu_j + \mu_j)^2.
$$

Plugging $\mu_j = \|\boldsymbol{w}_j^*\|$ and $\nu_j = \|\hat{\boldsymbol{H}}^\top \boldsymbol{H}^* \boldsymbol{w}_j^*\|$ into the part $|\nu_j - \mu_j|$ derives that

$$
\begin{aligned}
|\nu_j + \mu_j| &\overset{(a)}{\leq} \|\hat{\boldsymbol{H}}^\top \boldsymbol{H}^* \boldsymbol{w}_j^* + \boldsymbol{w}_j^*\| = \|(\hat{\boldsymbol{H}} + \boldsymbol{H}^*)^\top (\mathcal{M}_3(\mathcal{B}^*))_j\| \\
&\leq \|(\mathcal{M}_3(\mathcal{B}^*))_j\| \|\hat{\boldsymbol{H}} + \boldsymbol{H}^*\|_2 \leq \|(\mathcal{M}_3(\mathcal{B}^*))_j\| (\|\hat{\boldsymbol{H}}\|_2 + \|\boldsymbol{H}^*\|_2) \\
&\overset{(b)}{\leq} 2\|(\mathcal{M}_3(\mathcal{B}^*))_j\|,
\end{aligned}
\tag{21}
$$

where (a) follows from the Cauchy-Schwarz inequality, (b) comes from the fact that $\|\hat{\boldsymbol{H}}\|_2 = \|\boldsymbol{H}^*\|_2 = 1$ for the column-orthogonal matrices $\hat{\boldsymbol{H}}$ and $\boldsymbol{H}^*$. Setting $\bar{\alpha} = \max\{\|\boldsymbol{B}_j^*\|_F, j \in [p]\}$, we know that $(\nu_j - \mu_j)^2 \leq 4\bar{\alpha}^2$. Combining with (20) and (21), we obtain

$$
\begin{aligned}
\mathbb{E}\|\boldsymbol{\phi}_{[1:s]} + \boldsymbol{\mu}_{[1:s]}\|^2 = \mathbb{E}\left[\sum_{j=1}^s (\phi_j + \mu_j)^2\right] &\leq \mathbb{E}\left[\sum_{j=1}^s (4\psi_j^2 + 8\bar{\alpha}^2)\right] \\
&= 8s\bar{\alpha}^2 + 4\mathbb{E}(\zeta^2) = 4s(2\bar{\alpha}^2 + K),
\end{aligned}
\tag{22}
$$

where $\zeta^2 = \sum_{j=1}^s \psi_j^2 \sim \chi_{sK}^2$. Therefore, combining with (18), (19) and (22) yields that

$$
\begin{aligned}
E1 &\leq (1 + o(1))2\sigma^2 s \log(p/s) + 4s(2\bar{\alpha}^2 + K) \\
&\quad + 2\sqrt{4s(2\bar{\alpha}^2 + K)}\sqrt{(1 + o(1))2\sigma^2 s \log(p/s)} \\
&\sim (1 + o(1))2\sigma^2 s \log(p/s),
\end{aligned}
\tag{23}
$$

where the last step makes use of $s/p \to 0$.

We claim that $E2 = o(1)2\sigma^2 s \log(p/s)$ in the following. Note that $|\phi_j| = |\psi_j| \sim \chi_K$ since $\nu_j = 0$ for $j > s$. Denote $|\psi|_{(1)} \geq \cdots \geq |\psi|_{(p-s)}$ as the order statistics of $|\psi_{s+1}|, \ldots, |\psi_p|$. It follows from the proof of Lemma 3.3 in Su and Candès (2016) that

$$
\mathbb{E}\left[\sum_{j=s+1}^p \hat{\eta}_j^2\right] \leq \sum_{j=1}^{p-s} \mathbb{E}(|\psi|_{(j)} - \lambda_{s+j})_+^2,
$$

where $x_+ = \max\{0, x\}$. Then, we can partition the sum into three parts

$$
\begin{aligned}
\sum_{j=1}^{p-s} \mathbb{E}(|\psi|_{(j)} - \lambda_{s+j})_+^2 &= \sum_{j=1}^{\lfloor As \rfloor} \mathbb{E}(|\psi|_{(j)} - \lambda_{s+j})_+^2 \\
&\quad + \sum_{j=\lceil As \rceil}^{\lfloor ap \rfloor} \mathbb{E}(|\psi|_{(j)} - \lambda_{s+j})_+^2 + \sum_{j=\lceil ap \rceil}^{p-s} \mathbb{E}(|\psi|_{(j)} - \lambda_{s+j})_+^2,
\end{aligned}
$$

for a sufficiently large constant $A > 0$ and a sufficiently small constant $a > 0$. Note that Lemmas F.1, F.2 and F.3 given by Brzyski et al. (2019) show that each part is negligible compared with $2\sigma^2 s \log(p/s)$. This indicates that $E2 = o(1)2\sigma^2 s \log(p/s)$ and consequently completes the proof together with (23). ∎

Appendix C. Proofs for Section 4

C.1 Proof of Lemma 10

Denoting $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_K$ as singular values of $\boldsymbol{W}$, we can obtain that

$$\|\boldsymbol{W}\|_*^2 = \left(\sum_{i=1}^{K} \sigma_i\right)^2 \leq K \sum_{i=1}^{K} \sigma_i^2 = K\|\boldsymbol{W}\|_F^2. \tag{24}$$

In addition, we know from the singular value inequality (Chatelin, 1983) that $\sigma_{i+j-1}(\boldsymbol{AB}) \leq \sigma_i(\boldsymbol{A})\sigma_j(\boldsymbol{B})$ for $1 \leq i, j \leq K, i + j \leq K + 1$, which implies that $\|\boldsymbol{AB}\|_* \leq \|\boldsymbol{A}\|_*\|\boldsymbol{B}\|_*$ for any two matrices $\boldsymbol{A} \in \mathbb{R}^{p_1 p_2 \times p}$ and $\boldsymbol{B} \in \mathbb{R}^{p \times K}$. Together with (24), we derive that as $\|\boldsymbol{W}\|_F \to \infty$,

$$\begin{aligned}
\frac{F(\boldsymbol{W})}{\|\boldsymbol{W}\|_F} &= \frac{\|\boldsymbol{XW}\|_F^2/2 + P_{\boldsymbol{\lambda}}(\|\boldsymbol{W}\|_r) - \|\mathcal{M}_3(\mathcal{Y})^\top \boldsymbol{XW}\|_*}{\|\boldsymbol{W}\|_F} \\
&\geq \frac{\|\boldsymbol{XW}\|_F^2/2 + P_{\boldsymbol{\lambda}}(\|\boldsymbol{W}\|_r)}{\|\boldsymbol{W}\|_F} - \frac{\|\mathcal{M}_3(\mathcal{Y})^\top \boldsymbol{X}\|_*\|\boldsymbol{W}\|_*}{\|\boldsymbol{W}\|_F} \\
&\geq \frac{\|\boldsymbol{XW}\|_F^2/2 + P_{\boldsymbol{\lambda}}(\|\boldsymbol{W}\|_r)}{\|\boldsymbol{W}\|_F} - \sqrt{K}\|\mathcal{M}_3(\mathcal{Y})^\top \boldsymbol{X}\|_* \to \infty,
\end{aligned}$$

where '$\to$' is from the fact that the inequality $\|\boldsymbol{XW}\|_F^2 \geq \sigma_p^2(\boldsymbol{X})\|\boldsymbol{W}\|_F^2$ holds with $\sigma_p(\boldsymbol{X}) > 0$ when $\boldsymbol{X}$ is full column rank. This completes the proof. ∎

C.2 Proof of Theorem 11

It follows from Lemma 10 that the DC function F in (9) is also level-bounded, that is, the level set $\mathsf{L}_\alpha = \{\boldsymbol{W} \in \mathbb{R}^{p \times K} : F(\boldsymbol{W}) \leq \alpha\}$ is bounded for any $\alpha \in \mathbb{R}$. In addition, we know from Wen et al. (2018) that pDCAe enjoys the global convergence if F is level-bounded. Thus, using Lemma 10, the desired properties in (a)-(c) are consequences of Theorem 4.1 in Wen et al. (2018). ∎

Appendix D. Scaling of nonzero ground truth in simulations

In simulations, we generate the row sparsity factor matrix $\boldsymbol{W}^* \in \mathbb{R}^{p \times K}$ in a similar manner as in Brzyski et al. (2019), where each nonzero row of $\boldsymbol{W}^*$ is scaled such that $\|\boldsymbol{w}_j^*\| = a\sqrt{K}$ with $a = \sqrt{4\ln(p)/(1 - p^{-2/K}) - K}$. This specific form allows the generated signals to be comparable to the maximal noise such that nonzero signals (i.e., significant variables) can be identified with moderate power. We give the calculation of the scale value a in the following.

Considering the case of the column-orthogonal predictor matrix, the TgSLOPE estimator $\hat{\boldsymbol{W}}$ can be obtained from (14). We see from (14) that the identification of the R significant variables (where the number, R, is determined by $\boldsymbol{\lambda}$) corresponds to discovering indices of the R largest values among $\|\hat{\boldsymbol{y}}_1\|, \ldots, \|\hat{\boldsymbol{y}}_p\|$. Note that $\hat{\boldsymbol{y}}_1, \ldots, \hat{\boldsymbol{y}}_p$ are generated respectively by the random vectors $\hat{Y}_j = (\hat{Y}_{j_1}, \ldots, \hat{Y}_{j_K})^\top, j = 1, \ldots, p$, where

$$\hat{Y}_j = \boldsymbol{v}_j + E_j \sim \mathcal{N}(\boldsymbol{v}_j, \sigma^2 \boldsymbol{I}_K)$$

with $\boldsymbol{v}_j = \hat{\boldsymbol{H}}^\top \boldsymbol{H}^* \boldsymbol{w}_j^*$ and i.i.d. random noise vectors $E_j = \hat{\boldsymbol{H}}^\top (\mathcal{M}_3(\mathcal{E}))^\top \boldsymbol{X}_{:j} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_K)$ for $j = 1, \ldots, p$. Thus $\|\hat{Y}_j\| = \sqrt{\sum_{k=1}^K \hat{Y}_{jk}^2}$ has a χ_K distribution with the noncentrality parameter $\|\boldsymbol{v}_j\|$ and $\|E_j\|$ has a central χ_K distribution for $j \in [p]$. Then, the nonzero $\|\boldsymbol{v}_j\|$ (or the nonzero $\boldsymbol{w}_j^*$) could be perceived as a strong signal and thus be identified by TgSlOPE if with high probability the value $\|\hat{\boldsymbol{y}}_j\|$ generated from the noncentral χ distribution $\hat{Y}_j$ is large compared to the background noise produced by the random disturbance E_j with central χ_K distributions. Otherwise, the signal could be easily covered by random disturbances. Theorem H.1 of Brzyski et al. (2019) gives that for independent variables $Z_1, \ldots Z_p$ with $Z_j \sim \chi_K^2, j \in [p]$, one has

$$\mathbb{E}\Big(\max_{j \in [p]}\{Z_j\}\Big) \leq \frac{4\ln(p)}{1 - m^{-2/K}}.$$

Thus an idea is to use the quantity $\sqrt{4\ln(p)/(1 - m^{-2/K})}$ as the upper bound on the expected value of maximum over p independent χ_K-distributed variables. To investigate the discovery performance of TgSLOPE in simulation, we aim at $\mathbb{E}(\|\hat{\boldsymbol{y}}_j\|) = \sqrt{4\ln(p)/(1 - m^{-2/K})}$, which yields the setting that $\|\boldsymbol{v}_j\| = \sqrt{4\ln(p)/(1 - m^{-2/K})} - K$ since $\mathbb{E}(\|\hat{\boldsymbol{y}}_j\|) \approx \sqrt{\|\boldsymbol{v}_j\| + K}$. Note that $\|\boldsymbol{v}_j\| = \|\hat{\boldsymbol{H}}^\top \boldsymbol{H}^* \boldsymbol{w}_j^*\| \leq \|\hat{\boldsymbol{H}}\|_2 \|\boldsymbol{H}^*\|_2 \|\boldsymbol{w}_j^*\| = \|\boldsymbol{w}_j^*\|$ due to the column-orthogonality of $\hat{\boldsymbol{H}}$ and $\boldsymbol{H}^*$. In such a sense, we use the value $\sqrt{4K\ln(p)/(1 - p^{-2/K}) - K}$ to scale $\|\boldsymbol{w}_j^*\|$ in our simulations. This yields signals comparable to the maximal noise such that nonzero signals can be detected with moderate power.

References

T. Ahmed, H. Raja, and W. U. Bajwa. Tensor regression using low-rank and sparse tucker decompositions. *SIAM Journal on Mathematics of Data Science*, 2(4):944–966, 2020.

G. I. Allen. Sparse higher-order principal components analysis. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, pages 27–36, 2012.

A. Auddy and M. Yuan. Perturbation bounds for (nearly) orthogonally decomposable tensors and their applications in high dimensional data analysis. *Information and Inference*, 12(2):1044–1072, 2023.

P. C. Bellec, G. Lecué, and A. B. Tsybakov. SLOPE meets LASSO: Improved oracle bounds and optimality. *The Annals of Statistics*, 46(6):3603–3642, 2018.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300, 1995.

M. Bogdan, E. Van Den Berg, C. Sabatti, W. Su, and E. J. Candès. SLOPE-adaptive variable selection via convex optimization. *Annals of Applied Statistics*, 9(3):1103–1140, 2015.

D. Brzyski, A. Gossmann, W. Su, and M. Bogdan. Group SLOPE-adaptive selection of groups of predictors. *Journal of the American Statistical Association*, 114(525):419–433, 2019.

E. Bura, S. Duarte, L. Forzani, E. Smucler, and M. Sued. Asymptotic theory for maximum likelihood estimates in reduced-rank multivariate generalized linear models. *Statistics*, 52 (5):1005–1024, 2018.

F. Chatelin. *Eigenvalues of Matrices*. New York: Weily, 1983.

H. Chen, G. Raskutti, and M. Yuan. Non-convex projected gradient descent for generalized low-rank tensor regression. *Journal of Machine Learning Research*, 20:1–37, 2019.

Y. Chen, Z. Luo, and L. Kong. $\ell_{2,0}$-norm based selection and estimation for multivariate generalized linear models. *Journal of Multivariate Analysis*, 185:104782, 2021.

D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences - PNAS*, 100(5):2197–2202, 2003.

D. L. Donoho and I. M. Johnstone. Minimax risk over ℓ_p-balls for ℓ_q-error. *Probability Theory and Related Fields*, 99(2):227–303, 1994.

J. Fan, W. Gong, and Z. Zhu. Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of Econometrics*, 212(1):177–202, 2019.

J. Gotoh, Akiko Takeda, and Katsuya Tono. DC formulations and algorithms for sparse optimization problems. *Mathematical Programming*, 169:141–176, 2018.

J. C. Gower and G. B. Dijksterhuis. *Procrustes Problems*. Oxford University Press, 2004.

R. Han, R. Willett, and A. Zhang. An optimal statistical and computational framework for generalized tensor estimation. *The Annals of Statistics*, 50(1):1–29, 2022.

B. Hao, A. Zhang, and G. Cheng. Sparse and low-rank tensor estimation via cubic sketchings. *IEEE Transactions on Information Theory*, 66(9):5927–5964, 2020.

J. Hu, C. Lee, and M. Wang. Generalized tensor decomposition with features on multiple modes. *Journal of Computational and Graphical Statistics*, 31(1):204–218, 2022.

S. L. Hu and K. Ye. Linear convergence of an alternating polar decomposition method for low rank orthogonal tensor approximations. *Mathematical Programming*, 199(1-2): 1305–1364, 2023.

W. Hu, W. Shen, H. Zhou, and D. Kong. Matrix linear discriminant analysis. *Technometrics*, 62(2):196–205, 2020.

W. Hu, T. Pan, D. Kong, and W. Shen. Nonparametric matrix response regression with application to brain imaging data analysis. *Biometrics*, 77(4):1227–1240, 2021.

T. Inglot. Inequalities for quantiles of the Chi-square distribution. *Probability and Mathematical Statistics*, 30(2):339–351, 2010.

I. Johnstone and A. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

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

D. Kong, B. An, J. Zhang, and H. Zhu. L2RM: Low-rank linear regression models for high-dimensional matrix responses. *Journal of the American Statistical Association*, 115 (529):403–424, 2020.

J. B. Kruskal. Three-way arrays: Rank and uniqueness of trilinear decompositions, with applications to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18(2):95–138, 1977.

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

Z. Luo, D. Sun, K. C. Toh, and N. Xiu. Solving the OSCAR and SLOPE models using a semismooth newton-based augmented lagrangian method. *Journal of Machine Learning Research*, 20(106):1–25, 2019.

Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140:125–161, 2013.

G. Obozinski, M. J. Wainwright, and M. I. Jordan. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39:1–47, 2011.

J. C. Poythressa, J. Ahnb, and C. Parkc. Low-rank, orthogonally decomposable tensor regression with application to visual stimulus decoding of fMRI data. *Journal of Computational and Graphical Statistics*, 31(1):190–203, 2022.

G. Raskutti, M. Yuan, and H. Chen. Convex regularization for high-dimensional multiresponse tensor regression. *The Annals of Statistics*, 47(3):1554–1584, 2019.

R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer, 2009.

V. De Silva and L.-H. Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1084–1127, 2008.

M. Sørensen, L. D. Lathauwer, P. Comon, S. Icart, and L. Deneire. Canonical polyadic decomposition with a columnwise orthonormal factor matrix. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1190–1213, 2012.

W. Su and E. J. Candès. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3):1038–1068, 2016.

W. W. Sun and L. Li. STORE: Sparse tensor response regression and neuroimaging analysis. *Journal of Machine Learning Research*, 18:1–37, 2017.

D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. J. Behrens, E. Yacoub, and K. Ugurbil. The WU-Minn human connectome project: An overview. *NeuroImage*, 80(15):62–79, 2013.

G. A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33–45, 1992.

Q. Wei, Y. Zhang, and Z. Zhao. Sparse reduced-rank regression with adaptive selection of groups of predictors. In *55th Annual Asilomar Conference on Signals, Systems, and Computers*, pages 145–149, 2021.

B. Wen, X. Chen, and T. Pong. A proximal difference-of-convex algorithm with extrapolation. *Computational Optimization and Applications*, 69:297–324, 2018.

D. Yu, L. Wang, D. Kong, and H. Zhu. Mapping the genetic-imaging-clinical pathway with applications to alzheimer's disease. *Journal of the American Statistical Association*, 117 (540):1656–1668, 2022.

A. Zhang, Y. Luo, G. Raskutti, and M. Yuan. ISLET: Fast and optimal low-rank tensor regression via importance sketching. *SIAM Journal on Mathematics of Data Science*, 2 (2):444–479, 2020.

S. Zhang, X. Li, J. Lv, X. Jiang, L. Guo, and T. Liu. Characterizing and differentiating task-based and resting state fmri signals via two-stage sparse representations. *Brain Imaging and Behavior*, 10(1):21–32, 2016.

Y. Zhang, W. Shen, and D. Kong. Covariance estimation for matrix-valued data. *Journal of the American Statistical Association*, ahead-of-print:1–12, 2022. doi: 10.1080/01621459. 2022.2068419.

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