Robust and Covariance-assisted Tensor Response Regression

Ning Wang and Xin Zhang*

Beijing Normal University and Florida State University

Abstract

Tensor data analysis is gaining increasing popularity in modern multivariate statistics. In addition to the high-dimensionality and the higher-order structures, tensor data analysis in real-world applications often suffers from the heavy-tail issue, which is also a fundamental challenge for high-dimensional data analysis. Many tensor estimation approaches in the literature can be sensitive to heavy-tailed data and potential outliers. In this article, based on a recently proposed tensor t-distribution, we develop a robust and covariance-assisted tensor response regression model. This new model assumes that the tensor regression coefficient has a low-rank structure that can be learned more effectively by using the additional covariance information. This leads to a fast and robust decomposition-based estimation method. Theoretical and numerical studies confirm the superior performance of the proposed method.

Key Words: CP Decomposition; Dimension Reduction; Envelope Method; Robust Estimation; t-distributions; Tensor Regression.

1 Introduction

Many modern data sets are collected as a multidimensional array, also known as a tensor. Examples include neuroimaging data (Zhou et al. 2013, Karahan et al. 2015), multiple-tissue genetic data (Hore et al. 2016), and data sets studied in economics and finance Chen et al. (2022). In contrast to the traditionally used vector/multivariate data, tensor data has more complex structures and is usually very high-dimensional. Different modes of the tensor represents different natures and aspects of the data collection processes. For example, in the electroencephalography (EEG) data, the first model of the data represents the time/freqency and the second mode represents different

*Ning Wang (ningwangbnu@bnu.edu.cn) is Assistant Professor, Center of Statistics and Data Science, Beijing Normal University, Zhuhai, 519807, China; Xin Zhang (xzhang8@fsu.edu) is Associate Professor, Department of Statistics, Florida State University, Tallahassee, 32312, Florida, USA.

electrodes positions. Traditional vector-based methods vectorize the tensor data directly, which breaks the special tensor structure and usually results in the loss of information. Moreover, because of the multiple modes, the tensor dimension is usually high. The high-dimensionality results in an excessive number of parameters in statistical modeling and thus makes the tensor data analysis even more challenging. In recent years, there has been a rapidly growing literature on the analysis of tensor data, for example, in tensor decomposition (Kolda & Bader 2009, Chi & Kolda 2012, Sidiropoulos et al. 2017, Sun et al. 2017, Zhang & Han 2019), tensor regression (Zhou et al. 2013, Hoff 2015, Li & Zhang 2017, Zhang & Li 2017, Sun & Li 2017, Lock 2018, Raskutti et al. 2019) and tensor classification and clustering (Lyu et al. 2017, Pan et al. 2019, Mai et al. 2021, Wang, Zhang & Li 2022, Wang, Wang & Zhang 2022). These methods, among many others, can avoid vectorization and take advantage of the special tensor structure to achieve more accurate estimations.

Besides the complex structure and high dimensionality, we note that many tensor data analysis tools may suffer from the heavy-tail behaviors of the data and potential outliers. Some observations can be far away from the center of population, and bring more challenges to the estimation of a tensor model. We study the problem of tensor response regression with heavy-tail errors. Tensor response regression is a generalization of the multivariate linear regression model: Li & Zhang (2017) proposed a parsimonious tensor response regression using the tensor envelope; Sun & Li (2017) developed a sparse and low-rank model based on the CP decomposition Kolda & Bader (2009). However, the above-mentioned methods are not designed for tensor response with heavy-tail errors and may suffer from potential outliers. More recently, Wang, Zhang & Mai (2022) proposed a tensor t-distribution and applied it to the tensor response regression model to achieve robust estimations. The tensor t-distribution generalizes the multivariate t-distribution from vector to tensor data. In addition, it includes the tensor normal distribution (e.g. Li & Zhang 2017), as a special case. Both the tensor normal and tensor t-distribution assume that the covariance has a separable structure, which can reduce the free parameters in the covariance matrix substantially. The tensor t-distribution was shown to be closed for various tensor operators such as vectorization, linear transformation, rotation, and sub-tensor extraction. The simple definition and nice properties of the tensor t-distribution bring convenience for both algorithm implementation and theoretical studies and provide insights for its application in tensor analysis.

In this article, we focus on the study of the tensor response regression model with tensor t-distributed errors. The tensor regression coefficient is usually high-dimensional. Additional assumptions are usually required to avoid over-fitting and guarantee estimation accuracy. Wang, Zhang & Mai (2022) considered the sparsity assumption for it and proposed several regularized estimation methods. Although the popular sparsity assumption has been shown to work well for many high-dimensional data sets, it may not be suitable when we have a dense signal (i.e., the true regression coefficient tensor is not sparse). Therefore, we propose a covariance-assisted low-rank structure for the tensor coefficient, which jointly parametrizes the mean and covariance parameters in the tensor data similar to the tensor envelope approach in Li & Zhang (2017). We assume that the tensor response $\mathbf{Y}$ can be decomposed as $\mathbf{Y} = P(\mathbf{Y}) + Q(\mathbf{Y})$. Only $P(\mathbf{Y})$ is linearly associated with the predictor, while $Q(\mathbf{Y})$ has no linear association with the predictor and thus can be viewed as redundant information. More specifically, we assume that $P(\mathbf{Y})$ is a low-rank projection of $\mathbf{Y}$, which takes advantage of the tensor structure. As such, the tensor regression coefficients are in the form of the CP decomposition (e.g. Kolda & Bader 2009). In addition, we assume that the immaterial part $Q(\mathbf{Y})$ is uncorrelated with $P(\mathbf{Y})$ to eliminate its effects on $\mathbf{P}(\mathbf{Y})$. As a consequence, the mean and covariance of the tensor response can be jointly parameterized by the condition that the basics of the low-rank projection are the eigenvectors of the covariance matrices. We will provide more detailed motivations and explanations of the model assumption in Section 3. Thanks to the separable covariance structure, we can construct estimations for the covariance matrices with nice convergence results, which enhance the estimation accuracy of the tensor regression coefficient. The proposed structure shares similarities with the envelope (Cook et al. 2010) and tensor envelope (Li & Zhang 2017, Zhang & Li 2017) models, in which they assume that the regression coefficient has a low-rank structure, with the basis matrices belonging to a reducing subspace of the covariance matrices. Their strategies and ours are common: by projecting the large response onto a low-dimensional subspaces, we identify the part of the response that is relevant to regression and move the immaterial parts, which reduces the number of free parameters and facilitates the estimation efficiency. However, we will discuss later that the tensor envelope finds more directions than ours and estimates a surrogate subspace of ours. Besides, the envelope and tensor envelope are developed based on the normal distribution. As a comparison, the proposed method is developed based on the t-distribution and is robust against outliers. We then developed

a robust and decomposition-based algorithm. The core idea of the algorithm is similar to that of Zhang et al. (2022), which developed a straightforward way of envelope modeling from a principal components regression perspective and decomposition-based algorithms for the envelope method. The common procedure of our algorithm and theirs is that we first eigen-decompose the covariance matrices and then select the eigenvectors that belong to the target subspaces. The difference is that our algorithm is designed for the tensor data and is robust against potential outliers. Our algorithm has several advantages over the likelihood-based estimation method used in most literature about the tensor envelope. Firstly, our estimation considers the heavy-tail issue in the tensor response and is robust against outliers. Secondly, to obtain the estimate, we do not need any iterations. The algorithm only involves matrix multiplication and eigendecomposition. Thirdly, there are no local solution problems in our estimation. As a comparison, the likelihood-based objective function for the tensor envelope is complex and non-convex. The optimization for it is much more challenging and cannot guarantee to obtain the global solution.

The contributions of this article are multi-fold. Firstly, we propose a covariance-assisted low-rank structure. Compared with the first moment-based tensor low-rank method (e.g. Sun & Li 2017, Lock 2018), it jointly parametrizes the mean and covariance parameters to enhance estimation efficiency. Secondly, based on the tensor t-distribution, we propose a robust decomposition-based estimation method, which circumvents the iterations and non-convex problems in likelihood-based methods for the tensor envelope and is more computationally efficient. Thirdly, we obtain a non-asymptotic convergence rate for the proposed decomposition-based estimation method, which is strong enough for most tensor data. Note that we are handling the case where the tensor response is heavy-tail, which makes the theoretical analysis non-trivial. To our best knowledge, the proposed method is the first robust low-rank one for tensor response regression, which jointly parameterizes the tensor mean and covariance.

The rest of the paper is organized as follows. In Section 2, we introduce some tensor notations and review the tensor t-distribution and tensor response regression briefly. In Second 3, we propose the covariance-assisted tensor low-rank regression model. Then, in Second 4, we propose the robust decomposition-base estimation method. Section 5 shows the non-asymptotic convergence result of the proposed estimation method. Section 6 contains the numerical studies. Finally, Section 7 includes a short discussion. Proofs for the theorems are provided in the Supplementary Materials.

2 Preparations

2.1 Tensor notation

The following notation and (multi-)linear algebra will be used in this article. We call a multidimensional array $\mathbf{A} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ an M-way tensor or M-th order tensor, while $M = 1$ corresponds to vectors and $M = 2$ corresponds to matrices. Some key operators on a general M-th order tensor $\mathbf{A}$ are defined as follows.

- Vectorization. The vectorization of $\mathbf{A}$ is denoted by $\text{vec}(\mathbf{A}) \in \mathbb{R}^{\prod_m p_m}$, where the $(i_1, \cdots i_M)$-th scalar in $\mathbf{A}$ is mapped to the j-th entry of $\text{vec}(\mathbf{A})$, $j = 1 + \sum_{m=1}^{M} \{(i_m - 1) \prod_{k=1}^{m-1} p_k\}$.

- Matricization. The *mode-n matricization*, reshapes the tensor $\mathbf{A}$ into a matrix denoted by $\mathbf{A}_{(n)} \in \mathbb{R}^{p_n \times \prod_{m \neq n} p_m}$, so that the $(i_1, \cdots, i_M)$-th element in $\mathbf{A}$ becomes the (i_n, j)-th element of the matrix $\mathbf{A}_{(n)}$, where $j = 1 + \sum_{k \neq n} \{(i_k - 1) \prod_{l<k, l \neq n} p_l\}$.

- Vector product. The *mode-n vector product* of $\mathbf{A}$ and a vector $\mathbf{c} \in \mathbb{R}^{p_n}$ is represented by $\mathbf{A} \bar{\times}_n \mathbf{c} \in \mathbb{R}^{p_1 \times \cdots \times p_{n-1} \times p_{n+1} \times \cdots \times p_M}$ results in an $(M-1)$-th order tensor. This product is the result of the inner products between every *mode-n fiber* in $\mathbf{A}$ with vector $\mathbf{c}$. The mode-n fibers of $\mathbf{A}$ are the vectors obtained by fixing all indices except the n-th index.

- Matrix product. The *mode-n product* of tensor $\mathbf{A}$ and a matrix $\mathbf{G} \in \mathbb{R}^{s \times p_n}$, denoted as $\mathbf{A} \times_n \mathbf{G}$, is an M-th order tensor with dimension $p_1 \times \cdots \times p_{n-1} \times s \times p_{n+1} \times \cdots \times p_M$. Similar to the vector product, the product is a result of multiplication between every mode-n fibers of $\mathbf{A}$ and the matrix $\mathbf{G}$.

- Tucker product. The *Tucker product* of the core tensor $\mathbf{A}$ and a series of factor matrices $\mathbf{G}_1, \ldots, \mathbf{G}_M$, is defined as $\mathbf{A} \times_1 \mathbf{G}_1 \times_2 \cdots \times_M \mathbf{G}_M \equiv [\![\mathbf{A}; \mathbf{G}_1, \ldots, \mathbf{G}_M]\!]$.

- Tensor Mahalanobis distance. The *tensor Mahalanobis* of $\mathbf{A}$ with respect $\mathbf{\Xi} = \{\mathbf{\Sigma}_1, \cdots, \mathbf{\Sigma}_M\}$, where $\mathbf{\Sigma}_m \in \mathbb{P}^{p_m \times p_m}$, $m = 1, \cdots, M$, are positive and symmetric definite matrices, is defined as $\|\mathbf{A}\|_{\mathbf{\Xi}} = \text{vec}(\mathbf{A})^T (\otimes_{m=M}^{1} \mathbf{\Sigma}_m^{-1}) \text{vec}(\mathbf{A})$.

- Inner product of two tensors with the matching dimensions is $\langle \mathbf{A}, \mathbf{B} \rangle = \text{vec}^T(\mathbf{A}) \text{vec}(\mathbf{B})$ and Frobenius norm of $\mathbf{A}$ is $\|\mathbf{A}\|_F = \|\text{vec}(\mathbf{A})\|_2$.

For more background on tensor algebra, see Kolda & Bader (2009).

2.2 Tensor t distribution

In this section, we briefly review the tensor t-distribution (Wang, Zhang & Mai 2022), which aims to handle the heavy-tail issues in the tensor data. We start with the formal definition of it.

Definition 1. *A tensor-variate random variable* $\mathbf{Y} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ *follows the tensor* t*-distribution* $\mathrm{TT}(\boldsymbol{\mu}, \boldsymbol{\Xi}, \nu)$ *if and only if it has probability density function,*

$$f(\mathbf{Y} \mid \boldsymbol{\mu}, \boldsymbol{\Phi}, \nu) = \frac{\Gamma(\frac{\nu+p}{2}) \prod_{m=1}^{M} |\boldsymbol{\Sigma}_m|^{-p_{-m}/2}}{(\pi\nu)^{p/2}\Gamma(\nu/2)} \times (1 + \|\mathbf{Y} - \boldsymbol{\mu}\|_{\boldsymbol{\Xi}}^2/\nu)^{-\frac{\nu+p}{2}}, \tag{1}$$

where $\boldsymbol{\Xi} = \{\boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_M\}$, $p = \prod_{m=1}^{M} p_m$, $p_{-m} = \prod_{j \neq m} p_j$ *and* $\Gamma(\cdot)$ *is the Gamma function.*

The tensor t-distribution can be viewed as a generalization of the tensor normal distribution (e.g. Li & Zhang 2017, Manceur & Dutilleul 2013). A generative definition of tensor normal distribution is that $\mathbf{Y} \sim \mathrm{TN}(\boldsymbol{\mu}, \boldsymbol{\Xi})$ if $\mathbf{Y} = \boldsymbol{\mu} + [\![\mathbf{Z}; \boldsymbol{\Sigma}_1^{1/2}, \cdots, \boldsymbol{\Sigma}_M^{1/2}]\!]$ for some random tensor $\mathbf{Z}$ that consists of independent standard normal entries. The following proposition shows an equivalent representation for the tensor t distribution, which is more intuitive and builds the connection with the tensor normal distribution.

Proposition 1. *Suppose* $\mathbf{X} \sim \mathrm{TN}(0, \boldsymbol{\Xi})$ *and* $G \sim \chi_\nu^2/\nu$ *are independent, where* χ_ν^2 *is the Chi-square distribution with degree freedom* $\nu > 0$*, then* $\mathbf{Y} \sim \mathbf{X}/\sqrt{G} + \boldsymbol{\mu} \sim \mathrm{TT}(\boldsymbol{\mu}, \boldsymbol{\Xi}, \nu)$.

The tensor t distribution makes the tail of the tensor normal distribution heavier by introducing a single random variable $G \sim \chi_\nu^2/\nu$. When ν is small, the tensor t distribution has a much heavier tail compared with the tensor normal distribution and thus can account for the potential outliers in tensor data sets, and when $\nu \to \infty$, the tensor t-distribution reduces to the tensor normal distribution. Another important property of the tensor t-distribution is that $\mathrm{vec}(\mathbf{Y}) \sim t_p\big(\mathrm{vec}(\boldsymbol{\mu}), \bigotimes_{m=M}^{1} \boldsymbol{\Sigma}_m, \nu\big)$. Compared with multivariate t-distribution, the tensor t distribution has a separable covariance structure as the tensor normal distribution, which reduces the number of the free parameters in the scale parameter from $(\prod_{m=1}^{M} p_m)(\prod_{m=1}^{M} p_m + 1)/2$ to $\sum_{m=1}^{M} p_m(p_m + 1)/2 - M + 1$. Intuitively, less free parameters can enhance the estimation accuracy of statistical models. For more properties and interpretations of the tensor t-distribution, please refer to Wang, Zhang & Mai (2022).

2.3 Robust tensor response regression

To model the association between a response tensor $\mathbf{Y} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ and a covariate vector $\mathbf{X} = (X_1, \cdots, X_q)^T$, Wang, Zhang & Mai (2022) considered the following robust response regression model

$$\mathbf{Y} = \mathbf{B}_1 X_1 + \cdots + \mathbf{B}_q X_q + \mathbf{E}, \tag{2}$$

where $\mathbf{B}_k$ are the tensor coefficients for $k = 1, \cdots, q$, and $\mathbf{E} \sim \mathrm{TT}(0, \mathbf{\Xi}, \nu)$ are independent of $\mathbf{X}$. Without loss of generality, we assume that $\mathrm{E}(\mathbf{Y}) = 0$, $\mathrm{E}(\mathbf{X}) = 0$, and that the data are centered. Let $\mathbf{B} \in \mathbb{R}^{p_1 \times \cdots \times p_M \times q}$ be the stacked tensor coefficient $\{\mathbf{B}_1, \cdots, \mathbf{B}_q\}$. Model 2 is equivalent to $\mathbf{Y} = \mathbf{B} \bar{\times}_{M+1} \mathbf{X} + \mathbf{E}$. Compared with most existing approaches that assume the error $\mathbf{E}$ to be tensor normal (e.g., Li & Zhang 2017) or inexplicitly use the least squares loss that corresponds to the isotropic normal distribution (e.g. Rabusseau & Kadri 2016, Sun & Li 2017), the robust tensor response regression model is based on a tensor t-distribution whose tail can be much heavier. To gain more intuition, (Wang, Zhang & Mai 2022) considered the maximum likelihood estimation (MLE) of $\mathbf{B}$. For independent and identically distributed data $\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^n$ from (2), the MLE satisfies

$$\widehat{\mathbf{B}} = \mathbb{Y} \times_{M+1} (\mathbb{X}\mathbb{W}\mathbb{X}^T)^{-1}\mathbb{X}\mathbb{W},$$

where $\mathbb{W} \in \mathbb{R}^{n \times n}$ is a diagonal matrix with $w_i = (\nu + p)/\big(\nu + \|\mathbf{Y}_i - \widehat{\mathbf{B}} \bar{\times}_{M+1} \mathbf{X}_i\|_{\mathbf{\Xi}}^2\big)$, $\mathbb{Y} \in \mathbb{R}^{p_1 \times \cdots \times p_M \times n}$ is the sample tensor for the response, and $\mathbb{X} \in \mathbb{P}^{q \times n}$ is the sample matrix for the predictor. The MLE $\widehat{\mathbf{B}}$ can be viewed as a weighted least squares estimator. For the potential outliers far away from the center of the data, the tensor Mahalanobis distance $\|\mathbf{Y}_i - \widehat{\mathbf{B}} \bar{\times}_{M+1} \mathbf{X}_i\|_{\mathbf{\Xi}}^2$ is large, which makes the weights small. Hence, the outliers have a minor influence on the estimation.

For tensor data sets, the dimension of the response tensor $p = \prod_{m=1}^M p_m$ is usually high, which makes the tensor coefficient less interpretable and the estimation more challenging. Wang, Zhang & Mai (2022) assumed the tensor coefficient $\mathbf{B}$ to be sparse and proposed regularized estimations by using the adaptive lasso or adaptive group lasso penalties. Although the popular sparsity assumption works well for many high-dimensional data sets, the computation can be slow, especially when the model is not sparse enough, p is large, and the covariance information is considered.

3 Model

Instead of the sparsity assumption, we consider a tensor low-rank structure for the regression coefficients. Specifically, we consider the following parameterizations of $\mathbf{B}$ and $\mathbf{\Sigma}_m$, $m = 1, \cdots, M$,

$$\begin{aligned} \mathbf{B}_k &= \sum_{r=1}^{R} \alpha_{rk} \boldsymbol{\gamma}_{1r} \circ \cdots \circ \boldsymbol{\gamma}_{Mr},\ k = 1, \cdots, q, \\ \mathbf{\Sigma}_m \boldsymbol{\gamma}_{mr} &= \lambda_{mr} \boldsymbol{\gamma}_{mr},\ m = 1, \cdots, M,\ r = 1, \cdots, R, \end{aligned} \tag{3}$$

where $\circ$ is the outer product. We refer (3) as the covariance-assisted tensor low-rank model (CATL) and $\mathcal{P}_{\mathbf{\Sigma}}(\mathbf{B}) = \mathrm{span}\{\otimes_{m=M}^{1} \boldsymbol{\gamma}_{m1}, \cdots, \otimes_{m=M}^{1} \boldsymbol{\gamma}_{mR}\}$ as the covariance-assisted tensor low-rank subspace. The structure $\mathbf{B}_k = \sum_{r=1}^{R} \alpha_{rk} \boldsymbol{\gamma}_{1r} \circ \cdots \circ \boldsymbol{\gamma}_{Mr}$ is usually referred to as the CP decomposition (e.g. Kolda & Bader 2009), and R is called the rank of the decomposition. We use Figures 1 and 2 to help explain the motivation of (3). Firstly, as is shown in Figure 1, the basics $\boldsymbol{\gamma}_{mr}$, $m = 1, \cdots, M$, $r = 1, \cdots, R$, are common for all the coefficients $\mathbf{B}_k$, $k = 1, \cdots, q$. As such, when we project the response $\mathbf{Y}$ along each of its mode using basics $\boldsymbol{\gamma}_{1r}, \cdots, \boldsymbol{\gamma}_{Mr}$, (2) reduces to $[\![\mathbf{Y}; \boldsymbol{\gamma}_{1r}, \cdots, \boldsymbol{\gamma}_{Mr}]\!] = \alpha_{r1} X_1 + \cdots + \alpha_{rq} X_q + Z_R$, where $Z_R \sim \mathrm{t}(0, \prod_{m=1}^{M} \lambda_{mr}, \nu)$. Meanwhile, if we project $\mathbf{Y}$ onto subspaces orthogonal to $\mathcal{P}_{\mathbf{\Sigma}}(\mathbf{B})$, the projected response has no linear association with the predictor $\mathbf{X}$. Thus, we can write $\mathbf{Y}$ in the form of $\mathbf{Y} = P(\mathbf{Y}) + Q(\mathbf{Y})$, where $P(\mathbf{Y}) = [\![\mathbf{Y}; \boldsymbol{\gamma}_{11}\boldsymbol{\gamma}_{11}^T, \cdots, \boldsymbol{\gamma}_{M1}\boldsymbol{\gamma}_{M1}^T]\!] + \cdots + [\![\mathbf{Y}; \boldsymbol{\gamma}_{1R}\boldsymbol{\gamma}_{1R}^T, \cdots, \boldsymbol{\gamma}_{MR}\boldsymbol{\gamma}_{MR}^T]\!]$. Under this decomposition of $\mathbf{Y}$, only $P(\mathbf{Y})$, a low-rank projection of $\mathbf{Y}$, has linear association with the predictor, and $Q(\mathbf{Y}) = \mathbf{Y} - P(\mathbf{Y})$ is the immaterial part for regression. Since $\boldsymbol{\gamma}_{mr}$ is also a eigenvector of $\mathbf{\Sigma}_m$, we have $\mathrm{cov}\big(P(\mathbf{Y}), Q(\mathbf{Y})\big) = 0$, which means that the immaterial part will not influence the material part by correlation. Hence, if we can identify the subspace $\mathcal{P}_{\mathbf{\Sigma}}(\mathbf{B})$ successfully, the regression problem will reduce to a low-dimensional one.

We make two remarks for CATL. Firstly, although the CP decomposition is easy to be interpreted and widely used, obtaining the CP decomposition is computationally intractable (Kolda & Bader 2009). One of the most popular methods of obtaining CP decomposition is the alternating least squares (ALS) method, which may take many iterations to converge and is not guaranteed to converge to a global minimum or even a stationary point. As a comparison, by linking the mean and covariance, we will propose a non-iterative approach, which does not involve local-solution issues and is much more computationally efficient. Secondly, CATL finds a smaller subspace than

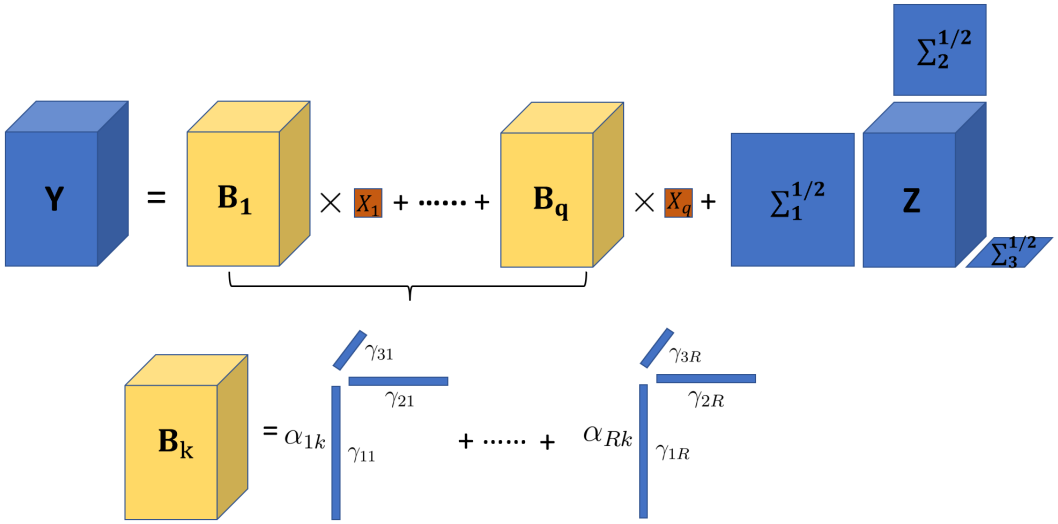

Figure 1: Visualization of Model 3

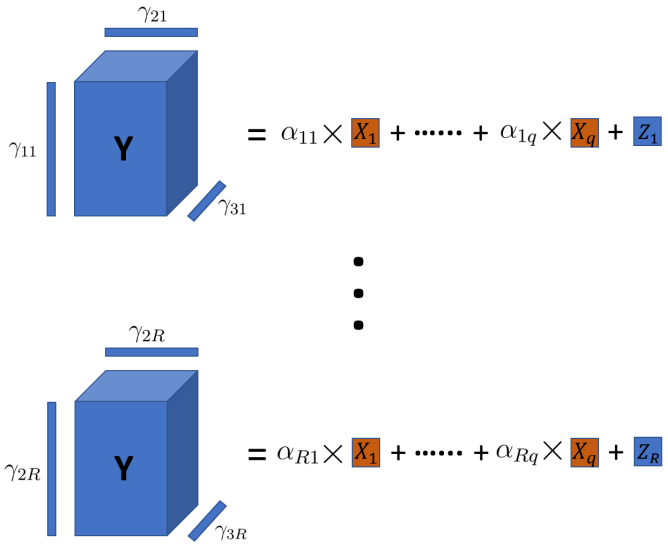

Figure 2: Explanation for Model 3

the tensor response envelope method (Li & Zhang 2017), which assumes that

$$\mathbf{B}_k = [\![\boldsymbol{\Theta}_k; \boldsymbol{\Gamma}_1, \cdots, \boldsymbol{\Gamma}_M]\!] \text{ for some } \boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times \cdots \times d_M},\ k = 1, \cdots, q,$$

$$\boldsymbol{\Sigma}_m = \boldsymbol{\Gamma}_m \boldsymbol{\Omega}_m \boldsymbol{\Gamma}_m^T + \boldsymbol{\Gamma}_{m0} \boldsymbol{\Omega}_{m0} \boldsymbol{\Gamma}_{m0}^T,\ m = 1, \cdots, M,$$

where $(\boldsymbol{\Gamma}_m, \boldsymbol{\Gamma}_{m0}) \in \mathbb{R}^{p_m \times p_m}$ is an orthogonal matrix, $d_m = \dim(\text{span}(\boldsymbol{\Gamma}_m))$, and $\boldsymbol{\Omega}_m$ and $\boldsymbol{\Omega}_{m0}$ are positive and symmetric matrices. Note that $\mathbf{B}_k$ can also be written as

$$\mathbf{B}_k = \sum_{j_1, \cdots, j_m} \theta^{(k)}_{j_1, \cdots, j_M} \boldsymbol{\Gamma}_{1 j_1} \circ \cdots \circ \boldsymbol{\Gamma}_{M j_M}.$$

where $\theta^{(k)}_{j_1,\cdots,j_M}$ is the $(j_1,\cdots,j_m)$-th element of $\mathbf{\Theta}_k$ and $\mathbf{\Gamma}_{mj_m}$ is the j_m-th column of $\mathbf{\Gamma}_m$. Let $\mathcal{T}_{\mathbf{\Sigma}}(\mathbf{B}) = \text{span}(\otimes^1_{m=M}\mathbf{\Gamma}_m)$. It is obvious that $\mathcal{P}_{\mathbf{\Sigma}}(\mathbf{B}) \subseteq \mathcal{T}_{\mathbf{\Sigma}}(\mathbf{B})$ because $\theta_{j_1,\cdots,j_M,k}$ can be zero. Hence, the tensor envelope subspace $\mathcal{T}_{\mathbf{\Sigma}}(\mathbf{B})$ can be viewed as a surrogate subspace of $\mathcal{P}_{\mathbf{\Sigma}}(\mathbf{B})$ and $R \leq \prod_{m=1}^M d_m$.

To gain more intuition, we use a toy example to show the connection and difference between $\mathcal{P}_{\mathbf{\Sigma}}(\mathbf{B})$ and $\mathcal{T}_{\mathbf{\Sigma}}(\mathbf{B})$. Suppose $p_1 = p_2 = 5$, $q = 1$, $\mathbf{B} \in \mathbb{R}^{5\times 5}$ with its first and second diagonal elements b_{11} and b_{22} to be 1 and the other elements to be 0, and $\mathbf{\Sigma}_m \in \mathbb{R}^{5\times 5}$, $m = 1$ and 1, are diagonal matrices, whose diagonal elements are all different. For this example, $\mathcal{P}_{\mathbf{\Sigma}}(\mathbf{B}) = \text{span}(\mathbf{e}_1 \otimes \mathbf{e}_1, \mathbf{e}_2 \otimes \mathbf{e}_2)$, which is a 2-dimensional linear subspace. As a comparison $\mathcal{T}_{\mathbf{\Sigma}}(\mathbf{B}) = \text{span}(\mathbf{e}_1\otimes\mathbf{e}_1, \mathbf{e}_1\otimes\mathbf{e}_2, \mathbf{e}_2\otimes\mathbf{e}_1, \mathbf{e}_2\otimes\mathbf{e}_2)$ is a 4-dimensional subspace. For tensor response regression models, CATL can always identify a smaller or at most the same subspace as the tensor envelop.

4 A robust decomposition-based estimation method

In this section, we propose a non-iterative decomposition-based estimation method for CATL model. By definition, γ_{rm}, $r = 1,\cdots,R$, are the eigenvectors of $\mathbf{\Sigma}_m$. Hence, we first obtain the eigen-decomposition of $\mathbf{\Sigma}_m$, $m = 1,\cdots,M$, and then identify the eigenvectors that belong to $\mathcal{P}_{\mathbf{\Sigma}}(\mathbf{B})$. The detailed algorithm in population is as follows.

1. Obtain the eigenvectors of $\mathbf{\Sigma}_m$: $\mathbf{v}^{(m)}_1,\cdots,\mathbf{v}^{(m)}_{p_m}$, with ordered eigenvalues $\lambda^{(m)}_1 \geq \cdots \geq \lambda^{(m)}_{p_m}$.

2. Calculate the envelope scores: $\phi_{l_1,\cdots,l_M} = \|[\![\mathbf{B}; \mathbf{v}^{(1)}_{l_1},\cdots,\mathbf{v}^{(M)}_{l_M}]\!]\|_2$, for $l_m = 1,\cdots,p_m$, $m = 1,\cdots,M$. Organize the envelope scores in the descending order $\phi_{(1)} \geq \phi_{(2)} \geq \cdots \geq \phi_{(\prod_{m=1}^M p_m)}$, and let $(\mathbf{v}^{(1)}_{l^{(j)}_1},\cdots,\mathbf{v}^{(M)}_{l^{(j)}_M})$ be the eigenvectors corresponding to $\phi_{(j)}$.

3. Output: $\mathcal{F}_{\mathbf{\Sigma}}(\mathbf{B}) = \text{span}(\otimes^1_{m=M}\mathbf{v}^{(m)}_{l^{(1)}_m}, \otimes^1_{m=M}\mathbf{v}^{(m)}_{l^{(2)}_m},\cdots,\otimes^1_{m=M}\mathbf{v}^{(m)}_{l^{(\widetilde{R})}_m})$, where $\widetilde{R}$ is the number such that $\phi_{(\widetilde{R}+1)} = 0$.

Let $\boldsymbol{\gamma}_m = (\boldsymbol{\gamma}_{m1},\cdots,\boldsymbol{\gamma}_{mR})$, $\mathbf{P}_{\boldsymbol{\gamma}_m}$ be the projection matrix onto the subspace spanned by the columns of $\boldsymbol{\gamma}_m$, and $\mathbf{Q}_{\boldsymbol{\gamma}_m} = \mathbf{I}_{p_m} - \mathbf{P}_{\mathbf{\Gamma}_m}$. The following Lemma shows the connection between the estimated subspace $\mathcal{F}_{\mathbf{\Sigma}}(\mathbf{B})$ and $\mathcal{P}_{\mathbf{\Sigma}}(\mathbf{B})$.

Lemma 1. *If the eigenvalues of* $\mathbf{P}_{\boldsymbol{\gamma}_m}\boldsymbol{\Sigma}_m\mathbf{P}_{\boldsymbol{\gamma}_m}$ *are all different and are distinct from those of* $\mathbf{Q}_{\boldsymbol{\gamma}_m}\boldsymbol{\Sigma}_m\mathbf{Q}_{\boldsymbol{\gamma}_m}$*, for* $m = 1, \cdots, M$*, then* $\mathcal{F}_{\boldsymbol{\Sigma}}(\mathbf{B}) = \mathcal{P}_{\boldsymbol{\Sigma}}(\mathbf{B})$ *and* $\widetilde{R} = R$*; If the eigenvalues of* $\mathbf{P}_{\boldsymbol{\gamma}_m}\boldsymbol{\Sigma}_m\mathbf{P}_{\boldsymbol{\gamma}_m}$ *are distinct from those of* $\mathbf{Q}_{\boldsymbol{\gamma}_m}\boldsymbol{\Sigma}_m\mathbf{Q}_{\boldsymbol{\gamma}_m}$*, then* $\mathcal{P}_{\boldsymbol{\Sigma}}(\mathbf{B}) \subseteq \mathcal{F}_{\boldsymbol{\Sigma}}(\mathbf{B}) \subseteq \mathcal{T}_{\boldsymbol{\Sigma}}(\mathbf{B})$ *and* $R \leq \widetilde{R} \leq \prod_{m=1}^{M} d_m$*. More specifically, let* $\mathcal{V}_{l_m}^{(m)}$*,* $l_m = 1, \cdots, u_m$*, where* $u_m \leq d_m$ *be the eigenspaces with non-zero eigenvalue of* $\mathbf{P}_{\boldsymbol{\gamma}_m}\boldsymbol{\Sigma}_m\mathbf{P}_{\boldsymbol{\gamma}_m}$*, for* $m = 1, \cdots, M$*, and* $\mathbf{V}_{l_m}^{(m)}$ *be a basis matrix of* $\mathcal{V}_{l_m}^{(m)}$*. We have* $\mathcal{T}_{\boldsymbol{\Sigma}}(\mathbf{B}) = \mathrm{span}(\otimes_{m=M}^{1}\widetilde{\boldsymbol{\gamma}}_{m1}, \cdots, \otimes_{m=M}^{1}\widetilde{\boldsymbol{\gamma}}_{mR})$*, where* $\widetilde{\boldsymbol{\gamma}}_{mr} = \mathbf{V}_{l_m}^{(m)}$ *if* $\boldsymbol{\gamma}_{mr} \in \mathcal{V}_{l_m}^{(m)}$ *for some* l_m *and* $\dim(\mathcal{V}_{l_m}^{(m)}) > 1$*, and* $\widetilde{\boldsymbol{\gamma}}_{mr} = \boldsymbol{\gamma}_{mr}$ *otherwise.*

Lemma 1 indicates that when the eigenvalues of $\mathbf{P}_{\boldsymbol{\gamma}_m}\boldsymbol{\Sigma}_m\mathbf{P}_{\boldsymbol{\gamma}_m}$ are all different and are distinct from those of $\mathbf{Q}_{\boldsymbol{\gamma}_m}\boldsymbol{\Sigma}_m\mathbf{Q}_{\boldsymbol{\gamma}_m}$, for $m = 1, \cdots, M$, the subspace $\mathcal{F}_{\boldsymbol{\Sigma}}(\mathbf{B})$ obtained by the algorithm is exactly the same as $\mathcal{P}_{\boldsymbol{\Sigma}}(\mathbf{B})$. For the worst case, the algorithm can estimate the tensor envelope subspace $\mathcal{T}_{\boldsymbol{\Sigma}}(\mathbf{B})$.

Next, we consider the finite sample case. Suppose we have n independent and identical distributed samples $\{(\mathbf{Y}_i, \mathbf{X}_i)\}_{i=1}^{n}$ from CATL model. Recall that the tail of the tensor error is heavier than the tensor normal distribution and our goal is to provide a robust estimation for the tensor coefficient $\mathbf{B}$. We construct the following robust estimates for the coefficient and the covariance matrices. Let $\widehat{\mathbf{B}}^{\mathrm{OLS}} = \mathbb{Y} \times_{M+1} \{(\mathbb{X}\mathbb{X}^T)^{-1}\mathbb{X}\}$ be the OLS estimator, $\widehat{\omega}_i = p/\|\mathbf{Y}_i - \widehat{\mathbf{B}}^{\mathrm{OLS}} \bar{\times}_{M+1} \mathbf{X}_i\|_F$ and $\mathbb{W} \in \mathbb{R}^{n\times n}$ with its ith diagonal elements to be ω_i. Define

$$\begin{aligned} \widehat{\mathbf{B}} &= \mathbb{Y} \times_{M+1} (\mathbb{X}\mathbb{W}\mathbb{X}^T)^{-1}\mathbb{X}\mathbb{W}, \\ \widehat{\boldsymbol{\Sigma}}_m &= \frac{1}{np_{-m}} \sum_{i=1}^{n} \widehat{\omega}_i (\mathbf{Y}_i - \widehat{\mathbf{B}}^{\mathrm{OLS}} \bar{\times}_{M+1} \mathbf{X}_i)_{(m)} (\mathbf{Y}_i - \widehat{\mathbf{B}}^{\mathrm{OLS}} \bar{\times}_{M+1} \mathbf{X}_i)_{(m)}^T. \end{aligned} \tag{4}$$

The sample algorithm is readily available by replacing $\boldsymbol{\Sigma}_m$ and $\mathbf{B}$ with $\widehat{\boldsymbol{\Sigma}}_m$ and $\widehat{\mathbf{B}}$. Then, the robust low-rank estimation $\widehat{\mathbf{B}}^{\mathrm{CATL}}$ is given by

$$\widehat{\mathbf{B}}^{\mathrm{CATL}} = \sum_{j=1}^{R} [\![\widehat{\mathbf{B}}; \mathbf{v}_{l_1^{(j)}}^{(1)} (\mathbf{v}_{l_1^{(j)}}^{(1)})^T, \cdots, \mathbf{v}_{l_M^{(j)}}^{(M)} (\mathbf{v}_{l_M^{(j)}}^{(M)})^T]\!].$$

In practice, we use 5-fold cross validation to select the rank $\widetilde{R}$.

We make several remarks for estimations $\widehat{\mathbf{B}}$ and $\widehat{\boldsymbol{\Sigma}}_m$. Firstly, in $\widehat{\mathbf{B}}$ and $\widehat{\boldsymbol{\Sigma}}_m$, the weights $\widehat{\omega}_i$ are different from those used in the MLE. We replace the tensor Mahalanobius distance $\|\mathbf{Y}_i - \widehat{\mathbf{B}} \bar{\times}_{M+1} \mathbf{X}_i\|_{\boldsymbol{\Xi}}$ by the Euclidean distance $\|\mathbf{Y}_i - \widehat{\mathbf{B}} \bar{\times}_{M+1} \mathbf{X}_i\|_F$. By making this adjustment, we avoid the iterations between the weights and the covariance matrices and thus accelerate the computation. Although those two weights are usually different from each other, they both measures

how far the sample is away from the center of the data. Thus, the potential outliers are assigned with small weights. Theoretically, we proved that $\widehat{\omega}_i$ estimates the latent variable G_i in the tensor t-error consistently up to a constant. Thus, it guarantees the consistency of $\widehat{\mathbf{B}}$ an $\widehat{\mathbf{\Sigma}}_m$ and enhances the robustness in the estimation. Secondly, we estimate $\widehat{\mathbf{\Sigma}}_m$ with an explicit formula. As a comparison, to obtain the MLE for $\mathbf{\Sigma}_m$, $m = 1, \cdots, M$, we need cyclically updates between all the covariance matrices, which can be time-consuming for high-dimensional tensor data sets. Although the iterations are omitted, $\widehat{\mathbf{\Sigma}}_m$ is still a consistent estimation for $\mathbf{\Sigma}_m$ up to a universal constant and thus the consistency of $\widehat{\mathbf{B}}^{\mathrm{CATL}}$ is not affected. Thirdly, the formula of $\widehat{\omega}_i$ does not involve the degree of freedom. Note that the expectation of $\|\mathbf{Y}_i - \widehat{\mathbf{B}} \bar{\times}_{M+1} \mathbf{X}_i\|_F$ is in the same order as p. In most tensor analyses, the dimension $p = \prod_{m=1}^{M}$ is large. Thus, the weights are very insensitive to the choice of ν. The $\widehat{\omega}_i$ we use circumvents the problem of selecting the degree of freedom ν.

5 Theory

In this section, we will establish the non-asymptotic convergence results for $\widehat{\mathbf{B}}^{\mathrm{CATL}}$. Due to technical reasons, we split the data into two batches and use the first batch to estimate $\widehat{\mathbf{B}}$ and the second batch to estimate $\widehat{\mathbf{\Sigma}}_m$, $m = 1, \cdots, M$. In our theoretical analysis, we allow the dimension of the tensor response to diverge and treat the dimension of the vector predictor q as a fixed number. Besides, we assume that $d_m = \dim(\mathrm{span}(\boldsymbol{\gamma}_{m1}, \cdots, \boldsymbol{\gamma}_{mR}))$ does not diverge with p and n. We first introduce some technical assumptions. Throughout this section, C and c represent generic constants that can vary line by line.

- (A1) The eigenvalues of $\mathbf{\Sigma}_m$, $m = 1, \cdots, M$, are all bounded between positive constants c_1 and c_2.
- (A2) The eigenvalues of $\mathbf{\Sigma}_{\mathbf{X},G} = \sum_{i=1}^{n} G_i \mathbf{X}_i \mathbf{X}_i^T$, where G_i is the latent random variable for the ith observation, are all bounded between positive constants c_3 and c_4.
- (A4) The absolute value of X_{ij}, which is the j-th element of $\mathbf{X}_i$, are upper bounded by M_X for all i and j.
- (A4) $c_5 \leq \alpha_{rk} \leq c_6$ for all r and k.

- (A5) $\sqrt{\frac{p_m}{np_{-m}}} = o(1)$ for all $m = 1, \cdots, M$.
- (A6) The degree of freedom $\nu > 4$.

Assumption (A1) implies that the population parameter $\mathbf{\Sigma}_m$ is well-conditioned regardless of how p_m grows. Since q is fixed, when n is large enough, Assumption (A2) is true for many distributions of $\mathbf{X}_i$, such as sub-Gaussian and sub-exponential distributions with positive definite covariance. Assumption (A3) assumes all the elements of $\mathbf{X}_i$ are bounded, which is a mild assumption in practice. Assumption (A4) states that the signal α_{rk} in (3) is well-conditioned, which is greater than a generic constant. Assumption (A5) is about the growth rate of n and p_m, $m = 1, \cdots, M$. When $M \geq 3$, it is true for most cases since p_{-m} is usually greater than p_m. It is even a mild assumption when $M = 2$ as long as the growth rates of p_m are consistent for $m = 1$ and 2. It guarantees the convergence of the proposed estimator. Assumption (A6) requires the existence of the fourth moment of the response. Note that the requirement of the fourth moment is only for facilitating theoretical studies. It is not required in numerical studies.

Theorem 1. *Under Assumptions (A1)–(A6),*

$$\|\widehat{\mathbf{B}}^{\mathrm{CATL}} - \mathbf{B}\|_2 = O(C_M\sqrt{1/n} + \max_m \sqrt{\frac{p_m}{np_{-m}}} + \sqrt{\log(n)/p})$$

with probability at least $1 - C_1 n^{-C_2} - C_3 \exp(-C_4 C_M) - C_5 \sum_{m=1}^{M} \exp(-p_m)$.

Corollary 1. *Under Assumptions (A1)–(A6), when* $p_m \to \infty$ *and* $n \to \infty$, $\widehat{\mathbf{B}}^{\mathrm{CATL}} \to \mathbf{B}$ *in probability.*

Note that the distribution of the tensor response in our model does not satisfy the popular sub-Gaussian or sub-exponential assumption. The moment generating function for each element of the response does not exist. Both this heavy-tail issue and the complex structure of the tensor make the theoretical analysis more challenging. The result in Theorem 1 is sufficiently strong for most tensor data applications since p_{-m} is usually greater than p_m, especially when the $M \geq 3$. If the dimensions p_m, $m = 1, \cdots, M$, grow at the same rate, the ratio p_m/p_{-m} either converges to zero ($M \geq 3$) or is bounded from above by a constant ($M = 2$). Then we have $\sqrt{n}$-consistency for arbitrarily high-dimensional p_m when $M \geq 2$. However, for vector data, the rate becomes $(p/n)^{1/2}$, which means p can not grow too fast. By aggregating the information from different modes, we

obtained a consistent estimation of $\mathbf{\Sigma}_m$, for which the convergence rate is much faster than the conventional sample covariance matrix. As such, by the joint parametrization (3), we obtained a consistent estimation for $\mathbf{B}$ with faster convergence rate than the vector-based approaches.

6 Numerical studies

In this section, we use several simulation models to investigate the finite sample performance of the proposed robust decomposition-based estimator. We include several estimators for tensor response regression as competitors: 1) The OLS estimator; 2) The weighted OLS estimator (WOLS) $\widehat{\mathbf{B}}$ defined in Section 4; 3) Non-robust version of proposed estimation method (CATL(N)), where the weights $\widehat{\omega}_i$ in (4) are replaced by 1; 4) Tensor envelope estimator (TE) proposed by Li & Zhang (2017), which is a likelihood-based low-rank estimator for tensor response regression. The proposed estimation method is represented by CATL. The evaluation criterion we use is the tensor Frobenius norm of the difference between the estimated $\mathbf{B}$ and the true parameter $\mathbf{B}$.

6.1 Simulation settings

We generate data from (2) with $\mathbf{\Xi} = \{\sigma^2\mathbf{\Sigma}_1, \cdots, \mathbf{\Sigma}_M\}$. We let $\mathbf{B}_k = \sum_{r=1}^{R} \alpha_{rk}\boldsymbol{\gamma}_{1r} \circ \cdots \circ \boldsymbol{\gamma}_{Mr}$ with R, α_{rk}, for $r = 1, \cdots, R$, to be specified later. The basics $\boldsymbol{\gamma}_{mr}$, $m = 1, \cdots, M$, $r = 1, \cdots, R$, unlike other specified, are randomly generated, which are othorgonal to each other for $r = 1, \cdots, R$. To make $\boldsymbol{\gamma}_{mr}$, $r = 1, \cdots, R$, be the eigenvectors for $\mathbf{\Sigma}_m$, we assume that $\mathbf{\Sigma}_m = \sum_{r=1}^{R} \lambda_{mr}\boldsymbol{\gamma}_{mr}\boldsymbol{\gamma}_{mr}^T + \boldsymbol{\gamma}_{m0}\mathbf{\Omega}_{m0}\boldsymbol{\gamma}_{m0}^T$, where $\boldsymbol{\gamma}_{m0}$ is a basis matrix of the complement subspace for $\mathrm{span}(\boldsymbol{\gamma}_{m1}, \cdots, \boldsymbol{\gamma}_{mr})$ and $\mathbf{\Omega}_0$ is a diagonal matrix, for $m = 1, \cdots, M$. The sample size $n = 100$ and the dimension of the predictor $q = 5$ unless other specified. Each element of the predictor $\mathbf{X}$ is generated from standard normal distribution independently. We set the degree of freedom ν to be 2. The covariance matrices we consider include three types:

- (C1) $\lambda_{mr} = r$, for $r = 1, \cdots, R$, and $\mathbf{\Omega}_{0m} = 0.5\mathbf{I}_{p_m - u_m}$.
- (C2) Each element of $(\lambda_{m1}, \cdots, \lambda_{mR})$ and $\mathrm{diag}(\mathbf{\Omega}_{m0})$ are randomly generated from $\mathrm{Uniform}(0.5, 2)$ independently.
- (C3) $\lambda_{mr} = 2^r$, for $r = 1, \cdots, R$, and $\mathrm{diag}(\mathbf{\Omega}_{m0}) = \exp(k_{m,1}, \ldots, k_{m,p_m - u_m})$, where $(k_{m,1}, \ldots, k_{m,p_m - u_m})$ are $p_m - u_m$ evenly spaced numbers between -2 and 2.

We consider the following 4 simulation models.

- (M1) We consider a 2-way matrix response $\mathbf{Y} \in \mathbb{R}^{p_1 \times p_2}$. We set $p_1 = p_2 = 30$ and $R = 2$ and α_{1k} and α_{2k} be randomly generated from $\mathrm{Uniform}(0, 1)$ independently, for $k = 1, \cdots, q$. The parameter σ^2 is 6, 12, and 15 for covariances (C1)-(C3), respectively. For this model, envelope rank (d_1, d_2) for $\mathcal{T}_{\mathbf{\Sigma}}(\mathbf{B})$ is $(2, 2)$.

- (M2) Similar to (M1) but $R = 4$. For basics $\boldsymbol{\gamma}_{m1}, \cdots, \boldsymbol{\gamma}_{m4}$, we first generate $\boldsymbol{\gamma}_{m1}$ and $\boldsymbol{\gamma}_{m2}$ randomly, then let $\boldsymbol{\gamma}_{m3} = \boldsymbol{\gamma}_{m1}$ and $\boldsymbol{\gamma}_{m4} = \boldsymbol{\gamma}_{m2}$. For this model, the envelope rank is $(2, 2)$.

- (M3) Similar to (M2) but with higher dimensions $p_1 = p_2 = 64$.

- (M4) Similar to (M1) but a 3-way example. We consider $\mathbf{Y} \in \mathbb{R}^{20 \times 30 \times 40}$ and $R = 3$. The parameter σ^2 is 6, 8, and 8 for covariances (C1)-(C3), respectively. For this model, the envelope rank (d_1, d_2, d_3) is $(3, 3, 3)$.

In simulation studies, we use true rank R for CATL and CATL(N), true envelope rank $(d_1, \cdots, d_M)$ for TE.

Model	M1			M2		
	(C1)	(C2)	(C3)	(C1)	(C2)	(C3)
OLS	22.45 (3.62)	47.42 (7.45)	43.63 (6.75)	22.44 (3.62)	47.42 (7.45)	43.63 (6.75)
WOLS	7.56 (0.12)	16.01 (0.24)	14.89 (0.23)	7.56 (0.12)	16.01 (0.24)	14.89 (0.23)
CATL	**1.52** (0.02)	**1.96** (0.02)	**1.99** (0.04)	**2.05** (0.03)	**2.98** (0.03)	**2.38** (0.05)
CATL(N)	3.12 (0.16)	6.27 (0.46)	2.23 (0.08)	4.48 (0.25)	9.02 (0.65)	3.44 (0.14)
TE	2.94 (0.15)	4.65 (0.35)	3.71 (0.39)	3.44 (0.14)	5.16 (0.34)	4.30 (0.38)
Model	M3			M4		
	(C1)	(C2)	(C3)	(C1)	(C2)	(C3)
OLS	31.30 (1.89)	62.99 (0.23)	28.79 (0.49)	58.77 (5.52)	184.63 (17.37)	58.14 (3.48)
WOLS	11.36 (0.12)	22.87 (0.23)	17.96 (0.11)	21.08 (0.26)	66.25 (0.83)	21.16 (0.21)
CATL	**1.79** (0.04)	**3.01** (0.02)	**3.67** (0.03)	**2.38** (0.02)	**2.68** (0.01)	**2.64** (0.02)
CATL(N)	4.86 (0.27)	8.98 (0.67)	5.00 (0.10)	3.31 (0.31)	5.20 (0.94)	3.03 (0.07)
TE	3.26 (0.10)	5.35 (0.07)	4.06 (0.06)	3.49 (0.29)	7.16 (0.72)	3.66 (0.23)

Table 1: The averaged estimation error for $\mathbf{B}$ (in Frobenius norm) and the associated standard errors (in parentheses) over 100 replicates.

Under Covariance (C1), the material variation in the response is larger than the immaterial variation. The estimation results for OLS and WOLS are not too bad. Covariances (C2) and (C3)

are more complex, and the immaterial variation can also be large, which makes the models more challenging. For those two covariances, OLS and WOLS fail to give meaningful estimation results. From Table 1, we can see that CATL is the best method for all simulation examples by considering both the heavy-tail issue of the data and using the tensor low-rank structure. Compared with CATL(N), by assigning small weights for the potential outliers, CATL is more robust and accurate. We have the same observation for OLS and WOLS that WOLS can improve the performance of OLS by considering the heavy-tail issue. Due to the high dimensionality of the tensor response, the estimation errors for OLS and WOLS are quite large, especially for the 3-way tenor model. By introducing the tensor low-rank structure, we obtain a substantial improvement in CATL, CATL(N) and TE. Note that CATL(N) has comparable performance as TE, a likelihood-based method, which demonstrates the estimation efficiency of the proposed decomposition-based method.

In Table 2, we also report the rank selection results based on a 5-fold cross-validation for M1-M4 with covariance (C1). With the increase in the sample size, cross-validation gains more accuracy in rank selection. When $n = 500$, the rank selection accuracy is over 80% for M1-M3. For M4, due to the high dimensionality, larger sample size is required for higher rank selection accuracy. Note that choosing a slightly larger rank is not problematic for the proposed method as long as the eigenvectors belonging to $\mathcal{T}_{\mathbf{\Sigma}}(\mathbf{B})$ are all selected. Thus, we can use a slightly larger rank than that selected by BIC in practice. A small number of over-selected eigenvectors will not influence the estimation results much.

	M1	M2	M3	M4
n=100	18	17	23	20
n=500	90	83	82	32

Table 2: Dimension selection accuracy for M1-M4 with covariance (C1). For each model setup, we repeated 100 simulations and report the number of cases (out of 100) where the rank R is correctly selected.

6.2 Signal recovery

To further show the outperformance of CATL, we use another simulation example to visualize the tensor coefficients estimated by different methods. We generate data from CATL model (2) with $p_1 = p_2 = 64$ and $q = 1$. The model settings are parallel to those of (M1), except that the

regression coefficient $\mathbf{B} \in \mathbb{R}^{64\times 64}$. We assume that some elements of $\mathbf{B}$ are 1, while the others are all 0. The shape of $\mathbf{B}$ we consider includes a square, a cross and a bat. For covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, we use the (C2) structure with $\sigma^2 = 15$. To guarantee $\mathcal{T}_{\mathbf{\Sigma}}(\mathbf{B})$ is made up of the eigenvalues of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, we eigendecompose the coefficient matrix as $\mathbf{B} = \mathbf{G}_1\mathbf{D}\mathbf{G}_2^T$. Then we let $\boldsymbol{\gamma}_{mr} = \mathbf{G}_{mr}$, where r takes value from 1 to the rank of $\mathbf{G}_m$, for $m = 1$ and 2. The CP rank for the three shapes is 1, 2 and 169, respectively. Figure 1 visualizes the estimated coefficient matrix $\mathbf{B}$ by different methods. It is clearly seen that CATL performs much better than the OLS and WOLS estimator for square and cross shapes. For the bat shape, the coefficient is of relatively large rank, CATL performs similarly as WOLS and better than the other methods. Compared with CATL(N) and TE, by considering the heavy-tail issue and assigning small weights to the outliers, the estimated shapes of CATL are more clear.

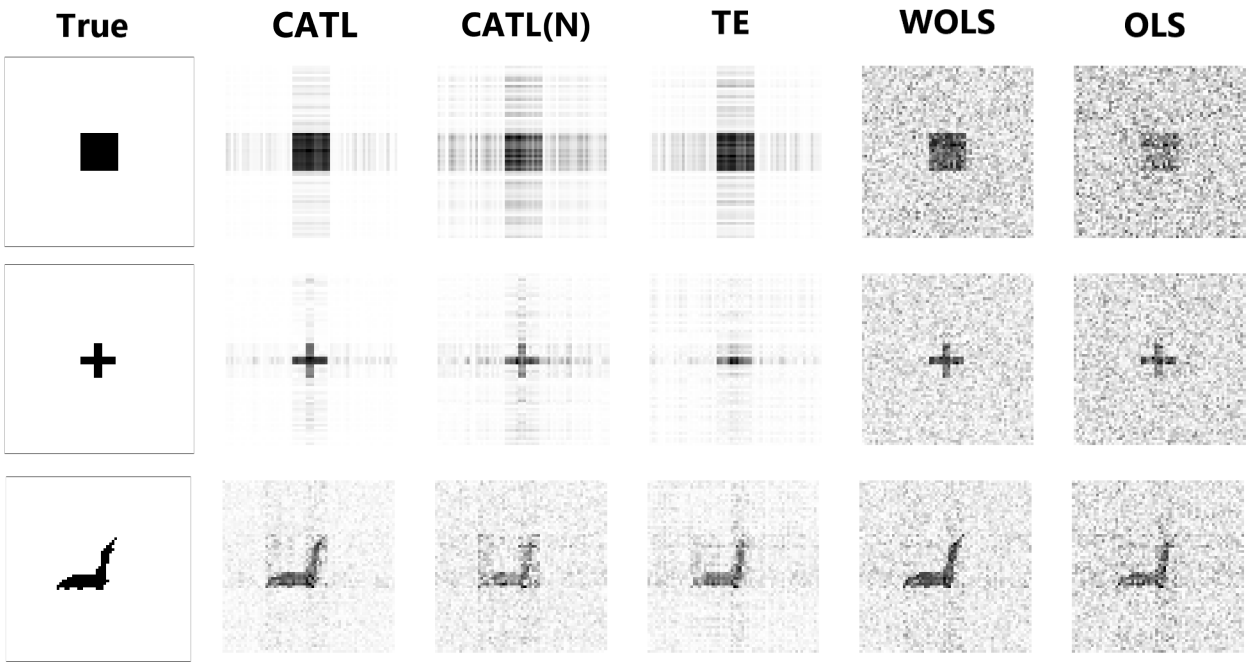

Figure 3: Pattern recovery results. Reported values are the absolute values of the estimated coefficients. The larger the absolute value is, the darker the pixel is.

6.3 Real data

We analyze an electroencephalography (EEG) data for an alcoholism study. The data was obtained from *https://archive.ics.uci.edu/ml/datasets/EEG+Database*. It contains 77 alcoholic individuals

and 44 controls. Each individual was measured with 64 electrodes placed on the scalp sampled at 256 Hz for one second, resulting an EEG image of 64 channels by 256 time points. More information about data collection and some analysis can be found in Li et al. (2010) and Li & Zhang (2017). To facilitate the analysis and visualization, we downsized the data along the time domain by averaging four consecutive time points, yielding a 64×64 matrix response. We draw the QQ plot for the EEG data set to check its normality. Specifically, we first regress the response tensor in predictors using least squares estimation and then standardize the residual tensor along each mode of it by the estimated covariance matrices in (4). We compare the quantiles of the standardized residuals with those of a χ^2 distribution with degree of freedom 64×64. From Figure 4, the heavy-tailed behavior is clear, and the potential outliers are possibly due to poor scan quality or problematic scan registration.

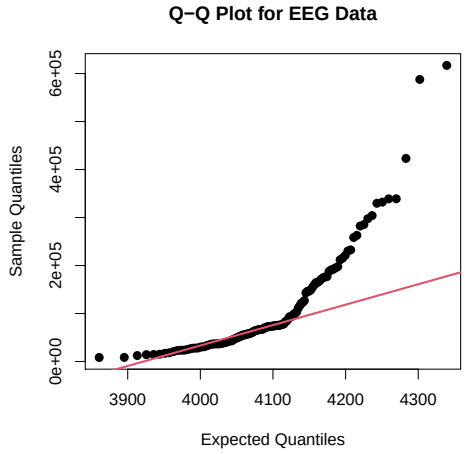

Figure 4: Quantile-Quantile (Q-Q) plot for EEG data.

We report the results of several estimators in Figure 5. For CATL and CATL(N), 5-fold cross validation selects rank $R = 24$. For TE, following Li & Zhang (2017), we use envelope dimension $(1, 1)$. In Figures 5 and 6, we report the estimated tensor coefficient and its truncated version. To get truncated coefficients, we first calculate the maximum of the absolute value for the coefficients and then set the elements whose absolute value is smaller than 0.2 times the maximum absolute value to be 0, while the others to be 1. We observe that CATL identifies the channels between about 0 to 10, 20 to 40 and 45 to 60 at time range from 80 to 160 and 200 to 240, that are most relevant to distinguish the alcoholic group from the control. The results of TE is similar to CATL. As a

comparison, the other estimators, especially CATL(N), are much more variable, with the revealed signal regions being less clear.

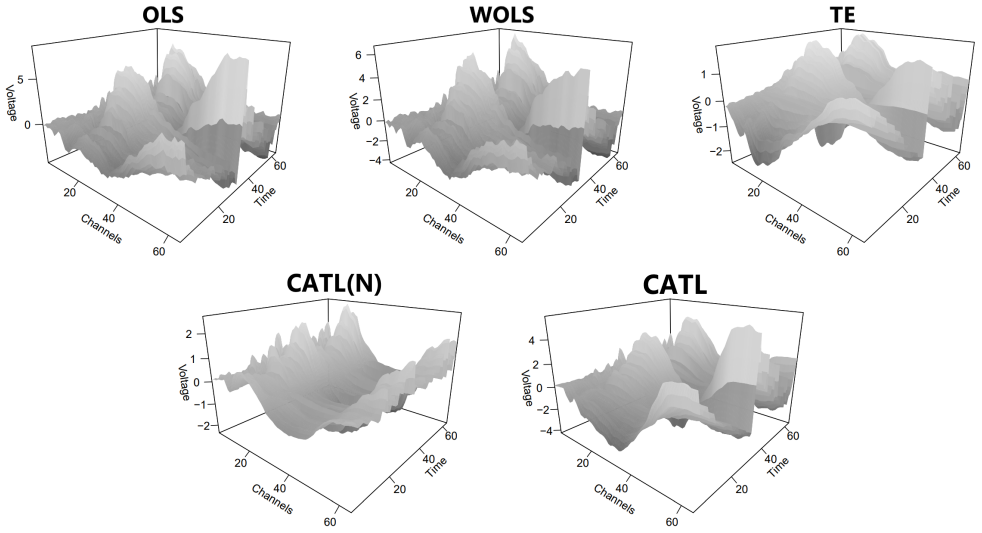

Figure 5: EEG data analysis: The five panels show the estimated coefficient tensor using different methods.

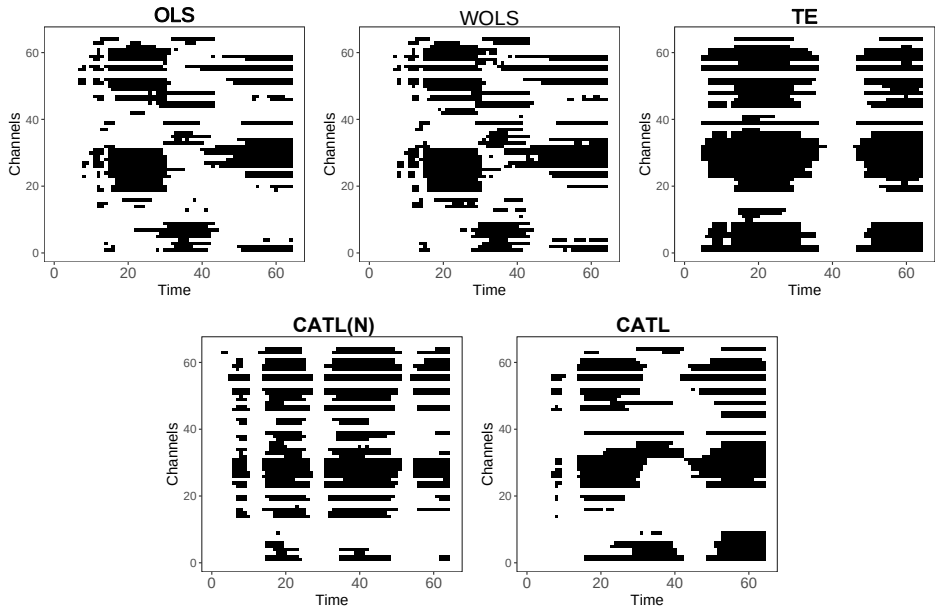

Figure 6: EEG data analysis: the five panels are the truncated tensor coefficient estimated by different methods at level 0.2.

7 Discussion

This paper considered a covariance-assisted low-rank estimation for the robust tensor response regression model and proposed a fast decomposition-based estimation method. Both the theoretical studies and numerical results demonstrated the superior performance of the proposed method. Although this article focus on the tensor regression model, the proposed covariance-assisted tensor low-rank structure and robust decomposition-based method can be extended to tensor classification, tensor clustering, and tensor graphical models. Another interesting future work is developing the tensor t-distribution to more general tensor distribution families and building robust statistical models.

References

Bi, X., Qu, A. & Shen, X. (2018), 'Multilayer tensor factorization with applications to recommender systems', *The Annals of Statistics* **46**(6B), 3308–3333.

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

Chi, E. C. & Kolda, T. G. (2012), 'On tensors, sparsity, and nonnegative factorizations', *SIAM Journal on Matrix Analysis and Applications* **33**(4), 1272–1299.

Cook, R., Li, B. & Chiaromonte, F. (2010), 'Envelope models for parsimonious and efficient multivariate linear regression', *Statistica Sinica* **20**(3), 927–960.

Hoff, P. D. (2015), 'Multilinear tensor regression for longitudinal relational data', *The Annals of Applied Statistics* **9**(3), 1169.

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

Karahan, E., Rojas-Lopez, P. A., Bringas-Vega, M. L., Valdés-Hernández, P. A. & Valdes-Sosa, P. A. (2015), 'Tensor analysis and fusion of multimodal brain images', *Proceedings of the IEEE* **103**(9), 1531–1559.

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

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

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

Lock, E. F. (2018), 'Tensor-on-tensor regression', *Journal of Computational and Graphical Statistics* **27**(3), 638–647.

Lyu, T., Lock, E. F. & Eberly, L. E. (2017), 'Discriminating sample groups with multi-way data', *Biostatistics* **18**(3), 434–450.

Mai, Q., Zhang, X., Pan, Y. & Deng, K. (2021), 'A doubly enhanced em algorithm for model-based tensor clustering', *Journal of the American Statistical Association* pp. 1–15.

Manceur, A. M. & Dutilleul, P. (2013), 'Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion', *Journal of Computational and Applied Mathematics* **239**, 37–49.

Pan, Y., Mai, Q. & Zhang, X. (2019), 'Covariate-adjusted tensor classification in high dimensions', *Journal of the American Statistical Association* **114**(527), 1305–1319.

Rabusseau, G. & Kadri, H. (2016), Low-rank regression with tensor responses, *in* 'NIPS'.

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

Sidiropoulos, N. D., De Lathauwer, L., Fu, X., Huang, K., Papalexakis, E. E. & Faloutsos, C. (2017), 'Tensor decomposition for signal processing and machine learning', *IEEE Transactions on Signal Processing* **65**(13), 3551–3582.

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

Sun, W. W., Lu, J., Liu, H. & Cheng, G. (2017), 'Provable sparse tensor decomposition', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**(3), 899–916.

Wang, N., Wang, W. & Zhang, X. (2022), 'Parsimonious tensor discriminant analysis', *Statistica Sinica* (accepted).

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

Wang, N., Zhang, X. & Mai, Q. (2022), 'High-dimensional tensor response regression using the t-distribution', *Manuscript* .

Zhang, A. & Han, R. (2019), 'Optimal sparse singular value decomposition for high-dimensional high-order data', *Journal of the American Statistical Association* **114**(528), 1708–1725.

Zhang, X., Deng, K. & Mai, Q. (2022), 'Envelopes and principal component regression', *arXiv preprint arXiv:2207.05574* .

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

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