Statistics and Its Interface Volume 17 (2024) 291–303

Robust and covariance-assisted tensor response regression

Ning Wang and Xin Zhang*

Tensor data analysis is gaining increasing popularity in modern multivariate statistics. When analyzing real-world tensor data, many existing tensor estimation approaches are sensitive to heavy-tailed data and outliers, in addition to the apparent high-dimensionality. In this article, we develop a robust and covariance-assisted tensor response regression model based on a recently proposed tensor t-distribution to address these issues in tensor data. This model assumes that the tensor regression coefficient has a low-rank structure that can be learned more effectively using the additional covariance information. This enables a fast and robust decomposition-based estimation method. Theoretical analysis and numerical experiments demonstrate the superior performance of our approach. By addressing the heavy-tail, high-order, and high-dimensional issues, our work contributes to robust and effective estimation methods for tensor response regression, with broad applicability in various domains.

Keywords and phrases: Dimension reduction, Envelope method, t-distributions, Tensor decomposition.

1. INTRODUCTION

Many modern data sets are collected as a multidimensional array, also known as a tensor. Examples include neuroimaging data [28, 8], multiple-tissue genetic data [7], and data sets studied in economics and finance [2]. In contrast to the traditionally used vector/multivariate data, tensor data has more complex structures and usually higher dimensions. Different modes of the tensor represent 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/frequency, and the second mode represents different electrode 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 [9, 3, 19, 21, 25], tensor regression [28, 6, 11, 27, 20, 12, 18] and tensor classification and clustering [13, 16, 14, 23, 22]. 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 heavy-tail issues and potential outliers. Some observations can be far from the population's center, bringing more challenges to estimating 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: [11] proposed a parsimonious tensor response regression using the tensor envelope; [20] developed a sparse and low-rank model based on the CP decomposition. However, the above-mentioned methods are not designed for tensor response with heavy-tail errors and may suffer from potential outliers. More recently, [24] 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. 11] 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 studying the tensor response regression model with tensor t-distributed errors. The tensor regression coefficient is high-dimensional for many tensor applications. Additional assumptions are usually required to avoid over-fitting and guarantee estimation accuracy. [24] considered the sparsity assumption and proposed several penalized 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

*Corresponding author.

low-rank structure for the tensor coefficient, which jointly parametrizes the mean and covariance parameters in the tensor data. We assume that the tensor response $\mathbf{Y}$ can be decomposed as $\mathbf{Y} = \mathcal{P}(\mathbf{Y}) + \mathcal{Q}(\mathbf{Y})$. Only $\mathcal{P}(\mathbf{Y})$ is linearly associated with the predictor, while $\mathcal{Q}(\mathbf{Y})$ has no linear association with the predictor and thus can be viewed as redundant information. More specifically, we assume that $\mathcal{P}(\mathbf{Y})$ 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. 9]. In addition, we assume that the immaterial part $\mathcal{Q}(\mathbf{Y})$ is uncorrelated with $\mathcal{P}(\mathbf{Y})$ to eliminate its effects on $\mathcal{P}(\mathbf{Y})$. 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 [4] and tensor envelope [11, 27] 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 subspace, we identify the part of the response 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 [26], 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, we do not need any iterations to obtain the estimate. 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 methods [e.g. 20, 12], 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. Section 2 introduces tensor notations and reviews the tensor t-distribution and tensor response regression. We propose the covariance-assisted tensor low-rank regression model in Second 3 and develop the robust decomposition-base estimation in Second 4. Section 5 shows the non-asymptotic convergence result of the proposed estimation method. Section 6 compares the proposed method with several related articles. Section 7 contains the numerical studies. Finally, Section 8 includes a short discussion. Proofs are provided in the Supplementary Materials (http://intlpress.com/site/pub/files/_supp/sii/2024/0017/0002/sii-2024-0017-0002-s002.zip).

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 $(i_1, \ldots i_M)$-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 $(i_1, \ldots, i_M)$-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 $\boldsymbol{\Xi}=\{\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_M\}$, where $\boldsymbol{\Sigma}_m \in \mathbb{P}^{p_m\times p_m}$, $m=1,\ldots,M$, are positive and symmetric definite matrices, is defined as $\|\mathbf{A}\|_{\boldsymbol{\Xi}} = \mathrm{vec}(\mathbf{A})^T(\otimes_{m=M}^1\boldsymbol{\Sigma}_m^{-1})\mathrm{vec}(\mathbf{A})$.
- Inner product of two tensors with the matching dimensions is $\langle\mathbf{A},\mathbf{B}\rangle = \mathrm{vec}^T(\mathbf{A})\mathrm{vec}(\mathbf{B})$ and Frobenius norm of $\mathbf{A}$ is $\|\mathbf{A}\|_F = \|\mathrm{vec}(\mathbf{A})\|_2$.

For more background on tensor algebra, see [9]. For a tensor $\mathbf{A}\in\mathbb{R}^{p_1\times\cdots\times p_M}$, we define $p=\prod_{m=1}^M p_m$ and $p_{-m}=\prod_{j\neq m}p_j$.

2.2 Tensor t-distribution

In this section, we briefly review the tensor t-distribution [24], 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 $\boldsymbol{\Xi}=\{\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_M\}$, $p=\prod_{m=1}^M p_m$, $p_{-m}=\prod_{j\neq m}p_j$ and $\Gamma(\cdot)$ is the Gamma function.

The tensor t-distribution can be viewed as a generalization of the tensor normal distribution [e.g. 11, 15]. A generative definition of tensor normal distribution is that $\mathbf{Y}\sim\mathrm{TN}(\boldsymbol{\mu},\boldsymbol{\Xi})$ if $\mathbf{Y}=\boldsymbol{\mu}+[\![\mathbf{Z};\boldsymbol{\Sigma}_1^{1/2},\ldots,\boldsymbol{\Sigma}_M^{1/2}]\!]$ for a 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* $G\sim\chi^2_\nu/\nu$ *are independent, where* χ^2_ν *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 $G\sim\chi^2_\nu/\nu$. 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 $\mathrm{vec}(\mathbf{Y})\sim t_p\big(\mathrm{vec}(\boldsymbol{\mu}),\bigotimes_{m=M}^1\boldsymbol{\Sigma}_m,\nu\big)$, where $t_p\big(\mathrm{vec}(\boldsymbol{\mu}),\bigotimes_{m=M}^1\boldsymbol{\Sigma}_m,\nu\big)$ is a p dimensional multivariate t-distribution with mean $\mathrm{vec}(\boldsymbol{\mu})$, scale parameter $\bigotimes_{m=M}^1\boldsymbol{\Sigma}_m$, and degree of freedom ν. 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 [24].

2.3 Robust tensor response regression

To model the linear association between a response tensor $\mathbf{Y}\in\mathbb{R}^{p_1\times\cdots\times p_M}$ and a covariate vector $\mathbf{X}=(X_1,\ldots,X_q)^T$, [24] considered the following robust response regression model

$$\mathbf{Y}=\mathbf{B}_1X_1+\cdots+\mathbf{B}_qX_q+\mathbf{E}, \tag{2}$$

where $\mathbf{B}_k$ are the tensor coefficients, for $k=1,\ldots,q$, and $\mathbf{E}\sim\mathrm{TT}(0,\boldsymbol{\Xi},\nu)$ is independent of $\mathbf{X}$. Without loss of generality, we assume that $\mathrm{E}(\mathbf{Y})=0$, $\mathrm{E}(\mathbf{X})=0$, and the data is centered. Let $\mathbf{B}\in\mathbb{R}^{p_1\times\cdots\times p_M\times q}$ be the stacked tensor coefficient $\{\mathbf{B}_1,\ldots,\mathbf{B}_q\}$. 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., 11] or inexplicitly use the least squares loss that corresponds to the isotropic normal distribution [e.g. 17, 20], the robust tensor response regression model is based on a tensor t-distribution whose tail can be much heavier. To gain more intuition, [24] 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

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

where $\mathbb{W}\in\mathbb{R}^{n\times n}$ is a diagonal matrix with the i-th diagonal element $w_i=(\nu+p)/\big(\nu+\|\mathbf{Y}_i-\widehat{\mathbf{B}}^{\mathrm{MLE}}\bar{\times}_{M+1}\mathbf{X}_i\|_{\boldsymbol{\Xi}}^2\big)$, $\mathbb{Y}\in\mathbb{R}^{p_1\times\cdots\times p_M\times n}$ is the sample tensor for the response, and $\mathbb{X}\in\mathbb{R}^{q\times n}$ is the sample matrix for the predictor. The MLE $\widehat{\mathbf{B}}^{\mathrm{MLE}}$ 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 $\|\mathbf{Y}_i-\widehat{\mathbf{B}}^{\mathrm{MLE}}\bar{\times}_{M+1}\mathbf{X}_i\|_{\boldsymbol{\Xi}}^2$ 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. [24] assumed the tensor coefficient $\mathbf{B}$ to be sparse and

proposed regularized estimations 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 $\mathbf{\Sigma}_m$, $m = 1, \ldots, M$,

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

where $\circ$ is the outer product. We refer (3) as the covariance-assisted tensor low-rank model (CATL) and $\mathcal{P}_{\mathbf{\Sigma}}(\mathbf{B}) = \text{span}\{\otimes_{m=M}^{1} \boldsymbol{\gamma}_{m1}, \ldots, \otimes_{m=M}^{1} \boldsymbol{\gamma}_{mR}\}$ 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. 9], 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 $\boldsymbol{\gamma}_{mr}$, $m = 1, \ldots, M$, $r = 1, \ldots, R$, are common for all the coefficients $\mathbf{B}_k$, $k = 1, \ldots, q$. As such, when we project the response $\mathbf{Y}$ along each of its modes using basics $\boldsymbol{\gamma}_{1r}, \ldots, \boldsymbol{\gamma}_{Mr}$, (2) reduces to $[\![\mathbf{Y}; \boldsymbol{\gamma}_{1r}, \ldots, \boldsymbol{\gamma}_{Mr}]\!] = \alpha_{r1} X_1 + \cdots + \alpha_{rq} X_q + Z_r$, where $Z_r \sim \text{t}(0, \prod_{m=1}^{M} \lambda_{mr}, \nu)$. 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 $\mathbf{Y} = \mathcal{P}(\mathbf{Y}) + \mathcal{Q}(\mathbf{Y})$, where $\mathcal{P}(\mathbf{Y}) = [\![\mathbf{Y}; \boldsymbol{\gamma}_{11}\boldsymbol{\gamma}_{11}^T, \ldots, \boldsymbol{\gamma}_{M1}\boldsymbol{\gamma}_{M1}^T]\!] + \cdots + [\![\mathbf{Y}; \boldsymbol{\gamma}_{1R}\boldsymbol{\gamma}_{1R}^T, \ldots, \boldsymbol{\gamma}_{MR}\boldsymbol{\gamma}_{MR}^T]\!]$. Under this decomposition of $\mathbf{Y}$, only $\mathcal{P}(\mathbf{Y})$, a low-rank projection of $\mathbf{Y}$, has a linear association with the predictor, and $\mathcal{Q}(\mathbf{Y}) = \mathbf{Y} - \mathcal{P}(\mathbf{Y})$ is the immaterial part for regression. Since $\boldsymbol{\gamma}_{mr}$ is also an eigenvector of $\mathbf{\Sigma}_m$, we have $\text{cov}\big(\mathcal{P}(\mathbf{Y}), \mathcal{Q}(\mathbf{Y})\big) = 0$, 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 interpret and widely used, obtaining the CP decomposition is computationally intractable [9]. 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

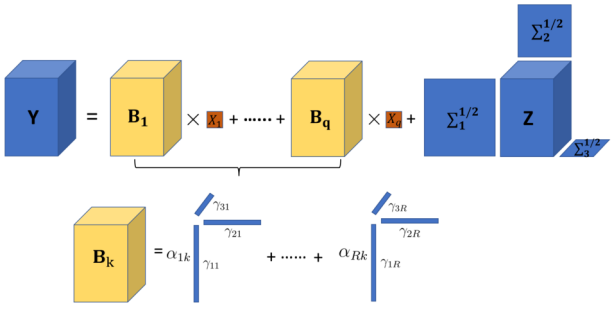

Figure 1. Parameterization for Model (3).

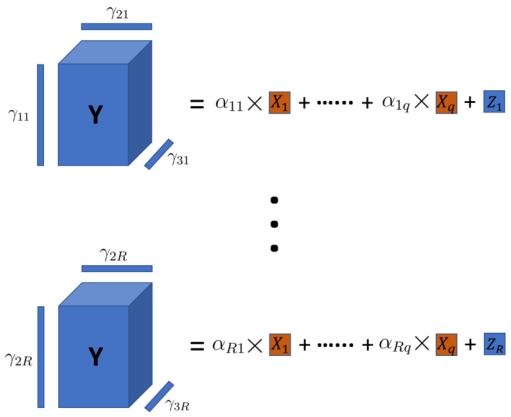

Figure 2. Dimension Reduction in Model (3).

the tensor response envelope method [11], which assumes that

$$\begin{aligned} &\mathbf{B}_k = [\![\mathbf{\Theta}_k; \mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_M]\!] \\ &\quad \text{for some } \mathbf{\Theta_k} \in \mathbb{R}^{d_1 \times \cdots \times d_M}, \ k = 1, \ldots, q, \\ &\mathbf{\Sigma}_m = \mathbf{\Gamma}_m \mathbf{\Omega}_m \mathbf{\Gamma}_m^T + \mathbf{\Gamma}_{m0} \mathbf{\Omega}_{m0} \mathbf{\Gamma}_{m0}^T, \ m = 1, \ldots, M, \end{aligned}$$

where $(\mathbf{\Gamma}_m, \mathbf{\Gamma}_{m0}) \in \mathbb{R}^{p_m \times p_m}$ is an orthogonal matrix, $d_m = \dim(\text{span}(\mathbf{\Gamma}_m))$, and $\mathbf{\Omega}_m$ and $\mathbf{\Omega}_{m0}$ are positive and symmetric matrices. The rank $(d_1, \ldots, d_M)$ is defined as the "tensor envelope rank". Note that $\mathbf{B}_k$ can also be written as

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

where $\theta^{(k)}_{j_1, \ldots, j_M}$ is the $(j_1, \ldots, j_m)$-th element of $\mathbf{\Theta}_k$ and $\mathbf{\Gamma}_{m j_m}$ is the j_m-th column of $\mathbf{\Gamma}_m$. Let $\mathcal{T}_{\mathbf{\Sigma}}(\mathbf{B}) = \text{span}(\otimes_{m=M}^{1} \mathbf{\Gamma}_m)$. It is obvious that $\mathcal{P}_{\mathbf{\Sigma}}(\mathbf{B}) \subseteq \mathcal{T}_{\mathbf{\Sigma}}(\mathbf{B})$ because $\theta^{(k)}_{j_1, \ldots, j_M}$ 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 2, are diagonal matrices, whose diagonal elements are all different. For this example, $\mathcal{P}_{\mathbf{\Sigma}}(\mathbf{B}) = \text{span}(\mathbf{e}_1 \otimes \mathbf{e}_1, \mathbf{e}_2 \otimes \mathbf{e}_2)$ 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. CATL can always identify a smaller or, at most, the same subspace as the tensor envelope for tensor response regression models.

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_{mr}, $r = 1, \ldots, R$, are the eigenvectors of $\mathbf{\Sigma}_m$. Hence, we first obtain the eigen-decomposition of $\mathbf{\Sigma}_m$, $m = 1, \ldots, M$, 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 $\mathbf{\Sigma}_m$: $\mathbf{v}_1^{(m)}, \ldots, \mathbf{v}_{p_m}^{(m)}$, with ordered eigenvalues $\lambda_1^{(m)} \geq \cdots \geq \lambda_{p_m}^{(m)}$.
2. Calculate the envelope scores: $\phi_{l_1,\ldots,l_M} = \|[\![\mathbf{B}; \mathbf{v}_{l_1}^{(1)}, \ldots, \mathbf{v}_{l_M}^{(M)}]\!]\|_2$, for $l_m = 1, \ldots, p_m$, $m = 1, \ldots, M$. 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 $(\mathbf{v}_{l_1^{(j)}}^{(1)}, \ldots, \mathbf{v}_{l_M^{(j)}}^{(M)})$ be the eigenvectors corresponding to $\phi_{(j)}$.
3. Output: $\mathcal{F}_{\mathbf{\Sigma}}(\mathbf{B}) = \text{span}(\otimes_{m=M}^1 \mathbf{v}_{l_m^{(1)}}^{(m)}, \otimes_{m=M}^1 \mathbf{v}_{l_m^{(2)}}^{(m)}, \ldots, \otimes_{m=M}^1 \mathbf{v}_{l_m^{(\widetilde{R})}}^{(m)})$, where $\widetilde{R}$ is the smallest number such that $\phi_{(\widetilde{R}+1)} = 0$.

Let $\boldsymbol{\gamma}_m = (\gamma_{m1}, \ldots, \gamma_{mR})$, $\mathbf{P}_{\boldsymbol{\gamma}_m}$ 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 motivation behind this algorithm is that if $\mathbf{v}_{l_1}^{(1)}, \ldots, \mathbf{v}_{l_M}^{(M)}$ are all belongs to $\mathbf{P}_{\boldsymbol{\gamma}_1}, \ldots, \mathbf{P}_{\boldsymbol{\gamma}_M}$, respectively, then the envelope score is non-zero. Otherwise, the envelope score is exactly zero.

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 $\mathbf{P}_{\boldsymbol{\gamma}_m}\mathbf{\Sigma}_m\mathbf{P}_{\boldsymbol{\gamma}_m}$ are all different and are distinct from those of $\mathbf{Q}_{\boldsymbol{\gamma}_m}\mathbf{\Sigma}_m\mathbf{Q}_{\boldsymbol{\gamma}_m}$, for $m = 1, \ldots, M$, then $\mathcal{F}_{\mathbf{\Sigma}}(\mathbf{B}) = \mathcal{P}_{\mathbf{\Sigma}}(\mathbf{B})$ and $\widetilde{R} = R$; If the eigenvalues of $\mathbf{P}_{\boldsymbol{\gamma}_m}\mathbf{\Sigma}_m\mathbf{P}_{\boldsymbol{\gamma}_m}$ are distinct from those of $\mathbf{Q}_{\boldsymbol{\gamma}_m}\mathbf{\Sigma}_m\mathbf{Q}_{\boldsymbol{\gamma}_m}$, then $\mathcal{P}_{\mathbf{\Sigma}}(\mathbf{B}) \subseteq \mathcal{F}_{\mathbf{\Sigma}}(\mathbf{B}) \subseteq \mathcal{T}_{\mathbf{\Sigma}}(\mathbf{B})$ and $R \leq \widetilde{R} \leq \prod_{m=1}^M d_m$. More specifically, let $\mathcal{V}_{l_m}^{(m)}$, $l_m = 1, \ldots, u_m$, where $u_m \leq d_m$, be the eigenspaces with non-zero eigenvalue of $\mathbf{P}_{\boldsymbol{\gamma}_m}\mathbf{\Sigma}_m\mathbf{P}_{\boldsymbol{\gamma}_m}$, and $\mathbf{V}_{l_m}^{(m)}$ be a basis matrix of $\mathcal{V}_{l_m}^{(m)}$. We have $\mathcal{F}_{\mathbf{\Sigma}}(\mathbf{B}) = \text{span}(\otimes_{m=M}^1 \widetilde{\boldsymbol{\gamma}}_{m1}, \ldots, \otimes_{m=M}^1 \widetilde{\boldsymbol{\gamma}}_{mR})$, 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 $\mathbf{P}_{\boldsymbol{\gamma}_m}\mathbf{\Sigma}_m\mathbf{P}_{\boldsymbol{\gamma}_m}$ are all different and are distinct from those of $\mathbf{Q}_{\boldsymbol{\gamma}_m}\mathbf{\Sigma}_m\mathbf{Q}_{\boldsymbol{\gamma}_m}$, for $m = 1, \ldots, M$, the subspace $\mathcal{F}_{\mathbf{\Sigma}}(\mathbf{B})$ obtained by the algorithm is exactly the same as $\mathcal{P}_{\mathbf{\Sigma}}(\mathbf{B})$. When the eigenvalues of $\mathbf{P}_{\boldsymbol{\gamma}_m}\mathbf{\Sigma}_m\mathbf{P}_{\boldsymbol{\gamma}_m}$ are distinct from those of $\mathbf{Q}_{\boldsymbol{\gamma}_m}\mathbf{\Sigma}_m\mathbf{Q}_{\boldsymbol{\gamma}_m}$ but are not all different, the algorithm can estimate a subspace that is larger than $\mathcal{P}_{\mathbf{\Sigma}}(\mathbf{B})$ but smaller than or equal to the tensor envelope subspace $\mathcal{T}_{\mathbf{\Sigma}}(\mathbf{B})$.

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 first consider the MLE for $\mathbf{B}$ and $\mathbf{\Sigma}_m$'s. Motivated by Proposition 1, we can construct the following EM algorithm for solving the MLE by treating G as the latent variable.

1. Let $\mathbf{B}^{(0)}$ be the OLS estimator, and $\mathbf{\Sigma}_m^{(0)} = \mathbf{I}_{p_m}$.
2. For $k = 0, 1, \ldots,$ repeat the following updates until convergence.
   (a) Update $\omega_i^{(k)} = (\nu + p)/(\nu + \|\mathbf{Y}_i - \mathbf{B}^{(k)} \bar{\times}_{M+1} \mathbf{X}_i\|^2_{\mathbf{\Xi}^{(k)}})$ and let $\mathbb{W}^{(k)} = \text{diag}(\omega_1^{(k)}, \ldots, \omega_n^{(k)})$.
   (b) Update $\mathbf{\Xi}^{(k+1)}$ by cyclically updating
   $$\widetilde{\mathbf{\Sigma}}_m = \frac{1}{np_{-m}} \sum_{i=1}^n \omega_i^{(k)} (\mathbf{Y}_i - \mathbf{B}^{(k)} \bar{\times}_{M+1} \mathbf{X}_i)_{(m)} \times \left(\bigotimes_{j \neq m} \widetilde{\mathbf{\Sigma}}_j^{-1} \right) (\mathbf{Y}_i - \mathbf{B}^{(k)} \bar{\times}_{M+1} \mathbf{X}_i)_{(m)}^T.$$
   (c) Update $\mathbf{B}^{(k+1)} = \mathbb{Y} \times_{M+1} (\mathbb{X}\mathbb{W}^{(k)}\mathbb{X}^T)^{-1}\mathbb{X}\mathbb{W}^{(k)}$.
3. At convergence, output $\widehat{\mathbf{B}}^{\text{MLE}}$ and $\widehat{\mathbf{\Sigma}}_m^{\text{MLE}}$, $m = 1, \ldots, M$.

To obtain the MLE, we need iterations between ω_i and $\mathbf{B}$ and $\mathbf{\Xi}$, and cyclically updates among $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_M$ within each iteration. Those iterations can be time-consuming in high dimensions. The EM algorithm motivates us to construct the following non-iterative robust estimates for the coefficient and the covariance matrices as an alternative approach.

Let $\widehat{\mathbf{B}}^{\text{OLS}} = \mathbb{Y} \times_{M+1} \{(\mathbb{X}\mathbb{X}^T)^{-1}\mathbb{X}\}$ be the OLS estimator, $\widehat{\omega}_i = p/\|\mathbf{Y}_i - \widehat{\mathbf{B}}^{\text{OLS}} \bar{\times}_{M+1} \mathbf{X}_i\|_F^2$, and $\widehat{\mathbb{W}} \in \mathbb{R}^{n\times n}$ be a diagonal matrix with the i-th diagonal elements to be ω_i. Define

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

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

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

In practice, the estimated envelope scores cannot be exactly zero. We use 5-fold cross validation to select the rank $\widetilde{R}$, which makes the cross-validation prediction error smallest. The full sample algorithm is summarized in Algorithm 1.

Algorithm 1 Robust decomposition-based algorithm for CATL.

1. Calculate $\widehat{\omega}_i = p/\|\mathbf{Y}_i - \widehat{\mathbf{B}}^{\mathrm{OLS}} \bar{\times}_{M+1} \mathbf{X}_i\|_F^2$, for $i = 1, \ldots, n$, and let $\widehat{\mathbb{W}}$ be a $n \times n$ matrix with its ith diagonal elements to be $\widehat{\omega}_i$ and other elements to be 0.
2. Calculate $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{\Sigma}}_m$, $m = 1, \ldots, M$, using (4).
3. Obtain the eigenvectors of $\widehat{\mathbf{\Sigma}}_m$: $\mathbf{v}_1^{(m)}, \ldots, \mathbf{v}_{p_m}^{(m)}$, with ordered eigenvalues $\lambda_1^{(m)} \geq \cdots \geq \lambda_{p_m}^{(m)}$.
4. Calculate the envelope scores: $\widehat{\phi}_{l_1,\ldots,l_M} = \|[\![\widehat{\mathbf{B}}; \mathbf{v}_{l_1}^{(1)}, \ldots, \mathbf{v}_{l_M}^{(M)}]\!]\|_2$, for $l_m = 1, \ldots, p_m$, $m = 1, \ldots, M$. Organize the envelope scores in the descending order $\widehat{\phi}_{(1)} \geq \widehat{\phi}_{(2)} \geq \cdots \geq \widehat{\phi}_{(\prod_{m=1}^M p_m)}$, and let $(\mathbf{v}_{l_1^{(j)}}^{(1)}, \ldots, \mathbf{v}_{l_M^{(j)}}^{(M)})$ be the eigenvectors corresponding to $\widehat{\phi}_{(j)}$.
5. Output: $\widehat{\mathbf{B}}^{\mathrm{CATL}} = \sum_{j=1}^{\widetilde{R}} [\![\widehat{\mathbf{B}}; \mathbf{v}_{l_1^{(j)}}^{(1)} (\mathbf{v}_{l_1^{(j)}}^{(1)})^T, \ldots, \mathbf{v}_{l_M^{(j)}}^{(M)} (\mathbf{v}_{l_M^{(j)}}^{(M)})^T]\!]$.

We make several remarks for estimations $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{\Sigma}}_m$. Firstly, in $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{\Sigma}}_m$, the weights $\widehat{\omega}_i$ are different from those used in the MLE. We replace the tensor Mahalanobius distance $\|\mathbf{Y}_i - \widehat{\mathbf{B}}^{\mathrm{MLE}} \bar{\times}_{M+1} \mathbf{X}_i\|_{\mathbf{\Xi}}$ by the Euclidean distance $\|\mathbf{Y}_i - \widehat{\mathbf{B}}^{\mathrm{OLS}} \bar{\times}_{M+1} \mathbf{X}_i\|_F$. 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, they measure 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. It guarantees the consistency of $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{\Sigma}}_m$ and enhances the robustness of the estimation. Secondly, we estimate $\widehat{\mathbf{\Sigma}}_m$ with an explicit formula. As a comparison, to obtain the MLE for $\mathbf{\Sigma}_m$, $m = 1, \ldots, M$, 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. Because the dimension p is large in most tensor datasets, the weights are very insensitive to the choice of ν. The $\widehat{\omega}_i$ we use can circumvent the problem of selecting the degree of freedom ν, which is also a challanging problem in high dimensions.

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 $\widehat{\mathbf{\Sigma}}_m$, $m = 1, \ldots, M$. The data splitting is a compromise to the limitation of the current techniques. In high-dimensional theoretical studies of iterative algorithms [e.g. 5, 1], such an assumption is often used to provide theoretical insights. In our algorithm, we use data splitting in theoretical analysis to make $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{\Sigma}}_m$ independent. 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 R 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 $\mathbf{\Sigma}_m$, $m = 1, \ldots, M$, are all bounded between positive constants c_1 and c_2. And the eigenvalues of $\mathbf{P}_{\boldsymbol{\gamma}_m} \mathbf{\Sigma}_m \mathbf{P}_{\boldsymbol{\gamma}_m}$ are distinct from those of $\mathbf{Q}_{\boldsymbol{\gamma}_m} \mathbf{\Sigma}_m \mathbf{Q}_{\boldsymbol{\gamma}_m}$.

(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 i-th 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 above 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 $m = 1, \ldots, M$.

(A6) The degree of freedom $\nu > 4$.

(A7) $\sqrt{\log(n)/p} \to 0$.

The first requirement of Assumption (A1) implies that the population parameter $\mathbf{\Sigma}_m$ is well-conditioned regardless of how p_m grows. By Lemma 1, the second requirement of Assumption (A1) ensures the proposed algorithm can estimate a good subspace. 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, $m = 1, \ldots, M$. 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. Assumption (A7) is used to show convergence of the weights $\widehat{\omega}_i$ to G_i, $i = 1, \ldots, n$. The weight $\widehat{\omega}_i$ can be viewed as an imputation for the latent variable G_i. By re-weighting the observations with $\widehat{\omega}_i = p/\|\mathbf{Y}_i - \widehat{\mathbf{B}}^{\mathrm{OLS}} \bar{\times}_{(M+1)} \mathbf{X}_i\|_F^2$, we can reduce the impact of the outliers. An especially interesting phenomenon is that, in evaluating ω_i, the high dimensionality is beneficial. Since all the elements in $\mathbf{Y}_i$ share the same latent variable G_i, having more of such elements gives us more information about G_i, and improves the accuracy in our imputation of G_i.

Theorem 1. *Under Assumptions (A1)–(A7),*

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

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)–(A7), when* $p_m \to \infty$ *and* $n \to \infty$, $\widehat{\mathbf{B}}^{\mathrm{CATL}} \to \mathbf{B}$ *in probability.*

Note that our model's distribution of the tensor response 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, $m = 1, \ldots, M$, 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 a faster convergence rate than the vector-based approaches.

6. COMPARISON

In this section, we compare the proposed method with several related models and methods. In particular, we discuss connections and new contributions to [11] and [24], from model assumptions, estimation methods, and theoretical studies.

6.1 Tensor response envelope regression

[11] first proposed the concept of tensor envelope for the tensor response regression model. Our solution differs from [11] in the following aspects.

Firstly, the proposed method finds a potentially smaller subspace than the tensor response envelope. As is discussed in the last two paragraphs of Section 3, the tensor response envelope structure can also be expressed using the proposed structure. However, the tensor envelope subspace can be much larger than the targeted subspace in this article. The analogy is similar to, but not exactly the same as, how tensor CP decomposition is different from tensor Tucker decomposition.

Secondly, the estimation method is completely different from that in [11], which considers the maximum likelihood estimation and its variants that involve solving a complex and non-convex objective function. In contrast, the proposed estimation method is decomposition-based and avoids the local optima issues.

Thirdly, besides the computational advantages, the new estimation approach allows us to establish the non-asymptotic convergence results and allows both p and n to diverge. To the best of our knowledge, the theoretical properties of the tensor envelope MLE is still unclear when p diverges. This is because of the iterative process in [11].

Finally, the proposed estimation is based on the newly proposed tensor t-distribution [24], which includes the tensor normal distributional assumption in [11] as a special case. The t-distributional assumption is robust against outliers and a useful non-trivial extension to tensor envelope methods.

6.2 Robust tensor response regression and other related methods

[24] introduced a tensor t-distribution and applied it to the tensor response regression model to achieve robust estimation. The method is relying solely on the sparsity assumption without incorporating any low-rank or low-dimensionality in their model assumptions. The method in this paper is based on matrix decomposition and can be much faster when the signal is dense than sparse.

In addition to the two most closely related articles, we summarize the connections between the proposed method and two other related articles in Table 1. [26] developed a straightforward way of envelope modeling from a principal components regression perspective and decomposition-based algorithms for the envelope method, but their methods are limited to vector data and are not robust against outliers. [22] considered a parsimonious tensor linear discriminant model and both iterative and non-iterative algorithms, but their methods do not work for regression models and are also not robust.

7. 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

Table 1. Comparisons with some related articles

	Tensor method	Low-rank	Robust	Non-iterative	Regression
[11]	✓	✓			✓
[24]	✓		✓		✓
[26]		✓		✓	✓
[22]	✓	✓		✓	
This paper	✓	✓	✓	✓	✓

Table 2. The averaged estimation error for $\mathbf{B}$ *(in Frobenius norm) and the associated standard errors (in parentheses) over 100 replicates*

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)

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 [11], 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 $\mathbf{B}$.

7.1 Simulation settings

We generate data from (2) with $\boldsymbol{\Xi} = \{\sigma^2\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M\}$. 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 $r = 1, \ldots, R$, to be specified later. The basics $\boldsymbol{\gamma}_{mr}$, $m = 1, \ldots, M$, $r = 1, \ldots, R$, unless otherwise specified, are randomly generated, which are orthogonal to each other for $r = 1, \ldots, R$. To make $\boldsymbol{\gamma}_{mr}$, $r = 1, \ldots, R$, be the eigenvectors for $\boldsymbol{\Sigma}_m$, we assume that $\boldsymbol{\Sigma}_m = \sum_{r=1}^{R} \lambda_{mr}\boldsymbol{\gamma}_{mr}\boldsymbol{\gamma}_{mr}^T + \boldsymbol{\gamma}_{m0}\boldsymbol{\Omega}_{m0}\boldsymbol{\gamma}_{m0}^T$, where $\boldsymbol{\gamma}_{m0}$ is a basis matrix of the complement subspace for $\mathrm{span}(\boldsymbol{\gamma}_{m1}, \ldots, \boldsymbol{\gamma}_{mr})$ and $\boldsymbol{\Omega}_0$ is a diagonal matrix, for $m = 1, \ldots, M$. The sample size $n = 100$ and the dimension of the predictor $q = 5$ unless otherwise 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 $r = 1, \ldots, R$, and $\boldsymbol{\Omega}_{0m} = 0.5\mathbf{I}_{p_m - u_m}$, where $u_m = \dim\big(\mathrm{span}(\boldsymbol{\gamma}_{m1}, \ldots, \boldsymbol{\gamma}_{mr})\big)$.

(C2) Each element of $(\lambda_{m1}, \ldots, \lambda_{mR})$ and $\mathrm{diag}(\boldsymbol{\Omega}_{m0})$ is randomly generated from Uniform$(0.5, 2)$ independently.

(C3) $\lambda_{mr} = 2^r$, for $r = 1, \ldots, R$, and $\mathrm{diag}(\boldsymbol{\Omega}_{m0}) = \exp(k_{m,1}, \ldots, k_{m,p_m - u_m})$, 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 Uniform$(0, 1)$ independently, for $k = 1, \ldots, q$. The parameter σ^2 is 6, 12, and 15 for covariances (C1)–(C3), respectively. For this model, envelope rank (d_1, d_2) for $\mathcal{T}_{\boldsymbol{\Sigma}}(\mathbf{B})$ is $(2, 2)$.

(M2) Similar to (M1) but $R = 4$. For basics $\boldsymbol{\gamma}_{m1}, \ldots, \boldsymbol{\gamma}_{m4}$, 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) and true envelope rank $(d_1, \ldots, d_M)$ for TE unless

Table 3. Dimension selection accuracy for M1–M4 with covariance (C1). For each model setup, we repeated 100 simulations and reported the number of cases (out of 100) where the rank R is correctly selected

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

otherwise specified.

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 2, 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 3, 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, a 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 $\mathcal{P}_{\mathbf{\Sigma}}(\mathbf{B})$ are all selected. Thus, we can use a slightly larger rank than that selected by BIC in practice.

We further use the simulation models M3 and M4, which have high dimensions, as examples to show the computational advantage of the proposed algorithm over Tensor envelope estimator (TE). Please see Table 4 for the results. Note that TE is implemented by a one-step estimator [11], which is an approximation of the MLE. Nevertheless, the proposed algorithm is more than two times faster.

7.2 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 3 visualizes the estimated coefficient matrix $\mathbf{B}$ by different methods. It is clearly seen that CATL performs much better than the OLS and WOLS estimators for square and cross shapes. For the bat shape, the coefficient is of relatively large rank, CATL performs similarly to 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.

Table 4. Average computational time for M3 and M4 over 100 replicates

Model	M1			M2		
	(C1)	(C2)	(C3)	(C1)	(C2)	(C3)
CATL	0.83 s	0.84 s	0.84 s	0.86 s	1.41 s	1.41 s
TE	1.80 s	1.68 s	2.16 s	2.13 s	3.87 s	4.31 s

7.3 Model mis-specification

In this section, we consider the model mis-specification issue. In particular, we examine the scenarios where the mean subspace and the covariance subspace are different. The simulation example is the same as M1 except that we replace the covariance $\mathbf{\Sigma}_m$ as $(1-\alpha)\mathbf{\Sigma}_m + \alpha\mathbf{\Phi}_m$ for each $m = 1, 2$. The additional covariance $\mathbf{\Phi}$ is constructed to induce model mis-specification and $\alpha \in (0, 1)$ indicates the magnitude of model mis-specification. Specifically, we set the diagonal elements of $\mathbf{\Phi}$ to be the same as those of $\mathbf{\Sigma}_m$, but the off-diagonal elements are $\mathbf{\Phi}_{ij} = 0.3\sqrt{(\mathbf{\Sigma}_m)_{ii}(\mathbf{\Sigma}_m)_{jj}}$. This covariance setting keeps the regression error term's magnitude at the same level for different α. When $\alpha = 0$, the model is the same as the original model M1. But when we increase the value of α, the covariance subspace becomes more and more unaligned from the mean subspace.

The average estimation error of the proposed method is shown in Figure 4, where we include all three covariance setting (C1)–(C3) (given in Section 7.1). As expected, the estimation error increases with α but slowly. In comparison, the averaged estimation errors for WOLS are 7.6, 16.0, and 14.9 for the three covariance models (C1)–(C3), respectively. The performance of the proposed method is still much better than WOLS even when $\alpha = 0.8$.

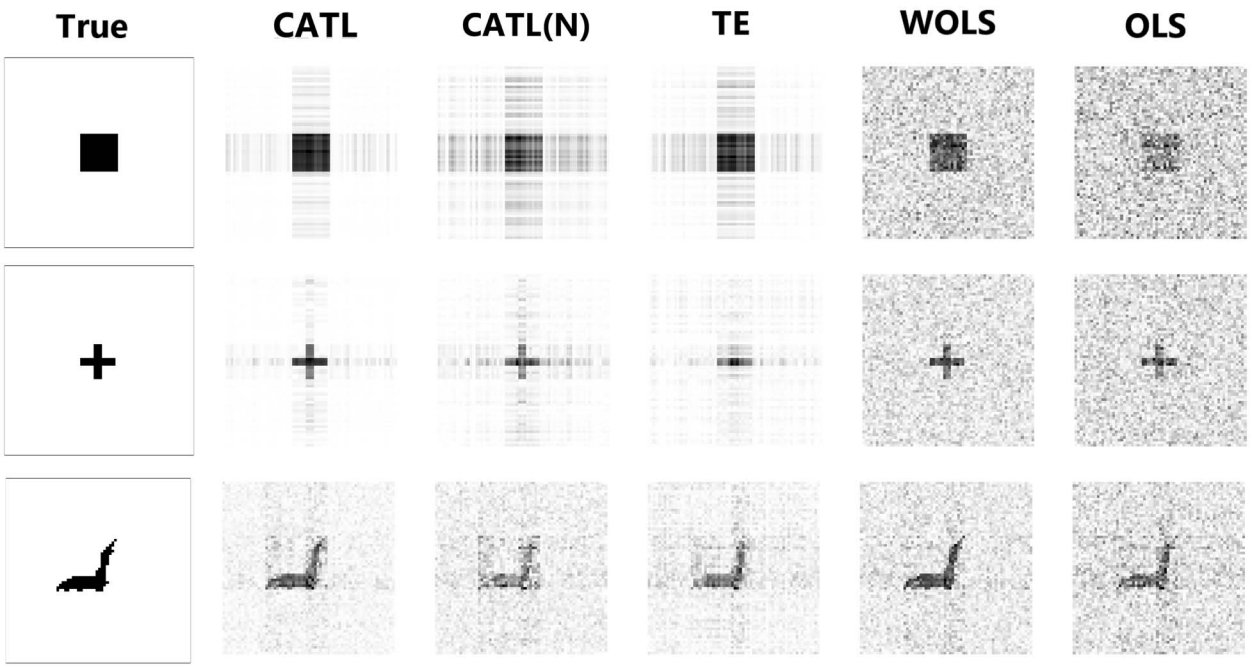

Figure 3. Pattern recovery results. Reported values are the absolute values of the estimated coefficients. The larger the absolute value is, the darker the pixel is.

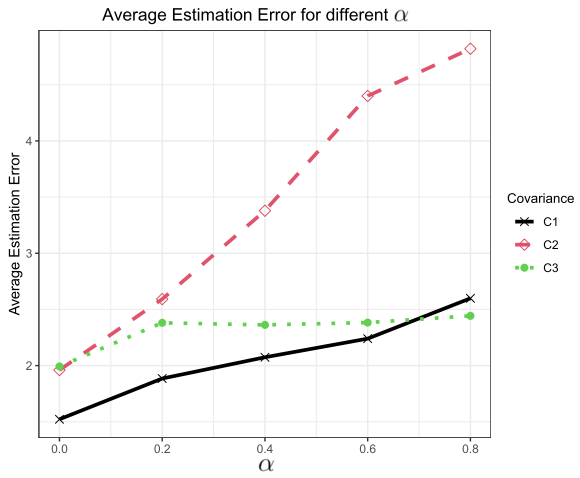

Figure 4. Average estimation error of the propose CATL method for varying magnitude of model mis-specification given in α.

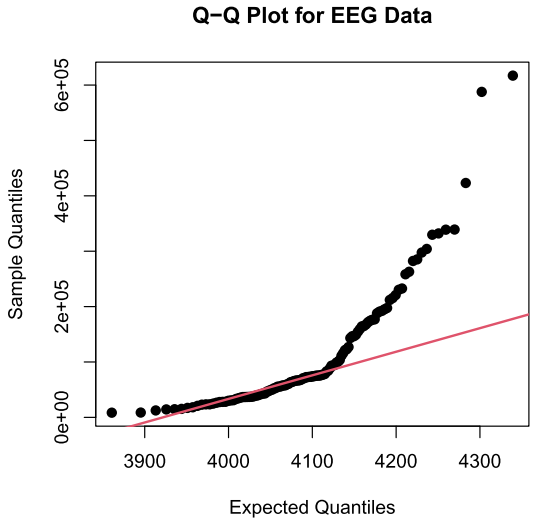

Figure 5. Quantile-Quantile (Q-Q) plot for EEG data.

7.4 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 in an EEG image of 64 channels by 256 time points. More information about data collection and some analysis can be found in [10] and [11]. 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 on 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

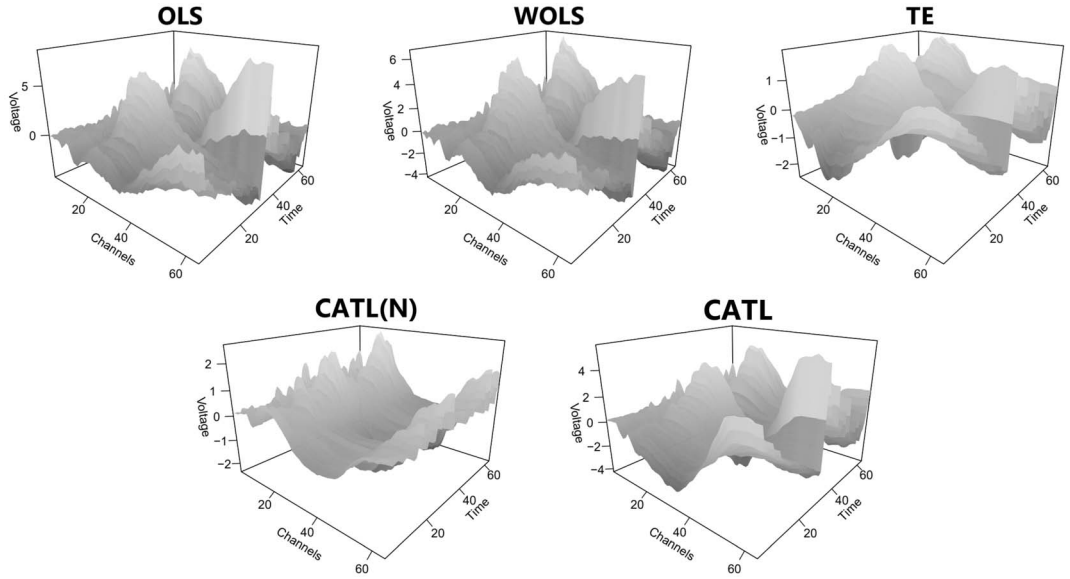

Figure 6. EEG data analysis: The five panels show the estimated coefficient tensor using different methods.

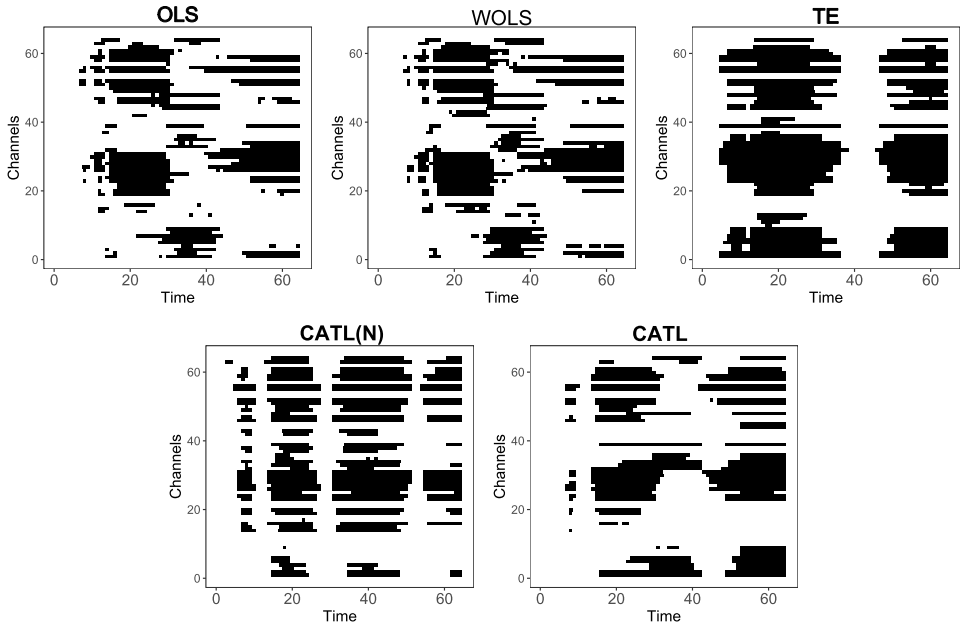

Figure 7. EEG data analysis: the five panels are the truncated tensor coefficient estimated by different methods at level 0.2.

standardized residuals with those of a χ^2 distribution with degree of freedom 64×64. From Figure 5, the heavy-tailed behavior is clear, and the potential outliers are possibly due to poor scan quality or problematic scan registration.

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 [11], we use envelope dimension $(1, 1)$. In Figures 6 and 7, we report the estimated tensor coefficient and its truncated version. To get truncated coefficients, we first calculate the maximum 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, which 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.

8. DISCUSSION

We propose a covariance-assisted robust tensor response regression model and develop a fast decomposition-based estimation method. The idea is closely related to the tensor envelope methods but is built on tensor t-distribution instead of normality assumptions. Moreover, the low-rank structural shared by mean and covariance parameters is essentially a simpler version of envelope subspace structural. The difference between our model assumptions and the tensor envelope models is analogous to the difference between the CP and Tucker tensor decompositions. Although this article focuses on the tensor regression model, the proposed robust decomposition-based estimation method can be extended naturally to tensor classification, tensor clustering, and other tensor envelope models.

The proposed method gains advantages over existing low-rank tensor methods mainly from two aspects. The first is the tensor t-distribution, a very recently proposed modeling strategy. The second is the covariance-assisted low-rank decompositions. For the covariance-assisted method to be effective, the key assumption is that the eigenvectors of $\mathbf{\Sigma}_m$ are identifiable and useful for constructing the regression mean parameter. If the covariance provides no useful eigenvector information, e.g., $\mathbf{\Sigma}_m = \sigma_m^2 \mathbf{I}_{p_m}$, then the proposed method would fail. Nevertheless, we note that the proposed method works reasonably well even when the covariance subspace is not perfectly aligned with the mean subspace (see Section 7.3). In our experience, this proposal provides a good alternative to the least square-type estimation methods in low-rank tensor regression, even when the mode-wise correlation structure is weak. In practice, one way to perform model diagnostics is by visualizing the projected response in Figure 2 versus the predictors (or a linear combination of the predictors). When the model is valid, the projected response and the predictors will have a non-trivial linear relationship.

ACKNOWLEDGEMENTS

The authors are grateful to the Editor, Associate Editor and referees for insightful comments that have led to significant improvements of this paper. Research for this paper was supported in part by grants CCF-1908969 and DMS-2053697 from the U.S. National Science Foundation.

Received 29 September 2022

REFERENCES

[1] Balakrishnan, S., Wainwright, M. J. and Yu, B. (2017). Statistical guarantees for the em algorithm: From population to sample-based analysis. *Ann. Statist.* **45**(1), 77–120. MR3611487

[2] Chen, R., Yang, D. and Zhang, C.-H. (2022). Factor models for high-dimensional tensor time series. *J. Am. Stat. Assoc.* **117**(537) 94–116. MR4399070

[3] Chi, E. C. and Kolda, T. G. (2012). On tensors, sparsity, and nonnegative factorizations. *SIAM J. Matrix Anal. Appl.* **33**(4) 1272–1299. MR3023474

[4] Cook, R., Li, B. and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Stat. Sin.* **20**(3) 927–960. MR2729839

[5] Gao, C., Ma, Z. and Zhou, H. H. (2017). Sparse CCA: Adaptive estimation and computational barriers. *Ann. Statist.* **45**(5) 2074–2101. MR3718162

[6] Hoff, P. D. (2015). Multilinear tensor regression for longitudinal relational data. *Ann. Appl. Stat.* **9**(3) 1169. MR3418719

[7] Hore, V., Viñuela, A., Buil, A., Knight, J., McCarthy, M. I., Small, K. and Marchini, J. (2016). Tensor decomposition for multiple-tissue gene expression experiments. *Nat. Genet.* **48**(9) 1094.

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

[9] Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Rev.* **51**(3) 455–500. MR2535056

[10] Li, B., Kim, M. K. and Altman, N. (2010). On dimension folding of matrix-or array-valued statistical objects. *Ann. Stat.* 1094–1121. MR2604706

[11] Li, L. and Zhang, X. (2017). Parsimonious tensor response regression. *J. Am. Stat. Assoc.* **112**(519) 1131–1146. MR3735365

[12] Lock, E. F. (2018). Tensor-on-tensor regression. *J. Comput. Graph. Stat.* **27**(3) 638–647. MR3863764

[13] Lyu, T., Lock, E. F. and Eberly, L. E. (2017). Discriminating sample groups with multi-way data. *Biostatistics* **18**(3) 434–450. MR3824759

[14] Mai, Q., Zhang, X., Pan, Y. and Deng, K. (2021). A doubly enhanced em algorithm for model-based tensor clustering. *J. Am. Stat. Assoc.* 1–15. MR4528493

[15] Manceur, A. M. and Dutilleul, P. (2013). Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *J. Comput. Appl. Math.* **239** 37–49. MR2991957

[16] Pan, Y., Mai, Q. and Zhang, X. (2019). Covariate-adjusted tensor classification in high dimensions. *J. Am. Stat. Assoc.* **114**(527) 1305–1319. MR4011781

[17] Rabusseau, G. and Kadri, H. (2016). Low-rank regression with tensor responses. In *NIPS*.

[18] Raskutti, G., Yuan, M., Chen, H. et al. (2019). Convex regularization for high-dimensional multiresponse tensor regression. *Ann. Stat.* **47**(3) 1554–1584. MR3911122

[19] Sidiropoulos, N. D., De Lathauwer, L., Fu, X., Huang, K., Papalexakis, E. E. and Faloutsos, C. (2017). Tensor decomposition for signal processing and machine learning. *IEEE Trans. Signal Process.* **65**(13) 3551–3582. MR3666587

[20] Sun, W. W. and Li, L. (2017). Store: sparse tensor response regression and neuroimaging analysis. *J. Mach. Learn. Res.* **18**(1) 4908–4944. MR3763769

[21] Sun, W. W., Lu, J., Liu, H. and Cheng, G. (2017). Provable sparse tensor decomposition. *J. R. Stat. Soc., Ser. B Stat. Methodol.* **79**(3) 899–916. MR3641413

[22] Wang, N., Wang, W. and Zhang, X. (2023). Parsimonious tensor discriminant analysis. *Stat. Sin.* (accepted).

[23] Wang, N., Zhang, X. and Li, B. (2022). Likelihood-based dimension folding on tensor data. *Stat. Sin.* **32** 2405–2430. MR4485089

[24] Wang, N., Zhang, X. and Mai, Q. (2023). High-dimensional tensor response regression using the t-distribution. *arXiv preprint arXiv:2306.12125*.

[25] Zhang, A. and Han, R. (2019). Optimal sparse singular value decomposition for high-dimensional high-order data. *J. Am. Stat. Assoc.* **114**(528) 1708–1725. MR4047294

[26] Zhang, X., Deng, K. and Mai, Q. (2023). Envelopes and principal component regression. *Electron. J. Stat.* **17**(2) 2447–2484. MR4652861

[27] Zhang, X. and Li, L. (2017). Tensor envelope partial least-squares regression. *Technometrics* **59**(4) 426–436. MR3740960

[28] Zhou, H., Li, L. and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *J. Am. Stat. Assoc.* **108**(502) 540–552. MR3174640

Ning Wang
Center of Statistics and Data Science
Beijing Normal University
Zhuhai, 519807
China
E-mail address: ningwangbnu@bnu.edu.cn

Xin Zhang
Department of Statistics
Florida State University
Tallahassee, 32312, Florida
USA
E-mail address: xzhang8@fsu.edu