Parsimonious Tensor Response Regression

Lexin Li and Xin Zhang

University of California at Berkeley; and Florida State University

Abstract

Aiming at abundant scientific and engineering data with not only high dimensionality but also complex structure, we study the regression problem with a multidimensional array (tensor) response and a vector predictor. Applications include, among others, comparing tensor images across groups after adjusting for additional covariates, which is of central interest in neuroimaging analysis. We propose parsimonious tensor response regression adopting a generalized sparsity principle. It models all voxels of the tensor response jointly, while accounting for the inherent structural information among the voxels. It effectively reduces the number of free parameters, leading to feasible computation and improved interpretation. We achieve model estimation through a nascent technique called the envelope method, which identifies the immaterial information and focuses the estimation based upon the material information in the tensor response. We demonstrate that the resulting estimator is asymptotically efficient, and it enjoys a competitive finite sample performance. We also illustrate the new method on two real neuroimaging studies.

Key Words: Envelope method; multidimensional array; multivariate linear regression; reduced rank regression; sparsity principle; tensor regression.

[1]Lexin Li is Associate Professor (Email: lexinli@berkeley.edu), Division of Biostatistics, University of California at Berkeley, Berkeley, CA 94720-3370. Xin Zhang is Assistant Professor (Email: henry@stat.fsu.edu), Department of Statistics, Florida State University, Tallahassee, FL 32306-4330. Li's research was partially supported by NSF grant DMS-1310319. Zhang's research was supported by the CRC-FYAP grant from Florida State University. The authors would like to thank the Editor, the Associate Editor and two referees for their insightful and constructive comments.

1 Introduction

For modern scientific data, an overarching feature that accompanies high or ultrahigh dimensionality is the complex structure of the data. For instance, in neuroimaging studies, electroencephalography (EEG) measures voltage value from numerous electrodes placed on scalp over time, and the resulting data is a two-dimensional matrix where the readings are both spatially and temporally correlated. Similarly, anatomical magnetic resonance imaging (MRI) takes the form of a three-dimensional array, where voxels correspond to brain locations and are spatially correlated. Multidimensional array data, also know as *tensor*, frequently arise in chemometrics, econometrics, psychometrics, and many other applications. Our motivation came from two neurological imaging studies: one is to compare the EEG scans between the alcoholic subjects and the general population, and the other is to compare the MRI scans of brains between the subjects with attention deficit hyperactivity disorder (ADHD) and the healthy controls after adjusting for age and gender of the subjects. This tensor comparison problem can be more generally formulated as a regression with the image tensor as response and the group indicator and other covariates as predictors. In this article, we aim to study this problem and term it *tensor response regression*. Although motivated by neuroimaging studies, the proposed methodology equally applies to a variety of scientific and engineering applications.

While there is an enormous body of literature tackling regression with high- or ultrahigh-dimensional predictors, there have been relatively few works on regression with multivariate response, and even fewer works on regression with tensor response. We review three major lines of related research. The first concerns multivariate *vector* response regression, and popular solutions include partial least squares (Helland, 1990, 1992; Chun and Keleş, 2010), canonical correlations (Zhou and He, 2008), reduced-rank regressions (Izenman, 1975; Reinsel and Velu, 1998; Yuan et al., 2007), sparse regressions with various penalties incorporating correlated response variables (Similä and Tikka, 2007; Turlach et al., 2005; Peng et al., 2010), and sparse reduced-rank regressions (Chen and Huang, 2012). Most of existing solutions adopt a linear association between the multivariate response and predictors, and they universally deal with the case where the

response variables are organized in the form of a vector. Our goal, however, is to model a tensor response, where the vector response can be viewed as a special case of a one-dimensional tensor. The second line of research directly models association between an image tensor and a vector of predictors in the context of brain imaging analysis. The dominating solution in the field regresses one response variable (voxel) at a time (Friston et al., 2007), which completely ignores underlying correlations among the voxels (Li et al., 2011). Li et al. (2011) and its follow-up works (Skup et al., 2012; Li et al., 2013) proposed a multiscale adaptive approach to smooth imaging response and to estimate parameters by building iteratively increasing neighbors around each voxel and combining observations within the neighbors with weights. Our approach differs in that we aim to model all the voxels in an image tensor *jointly* while incorporating the intrinsic spatial correlations among the voxels. Finally, there have been some recent developments regressing a scalar response on a tensor predictor (Reiss and Ogden, 2010; Zhou et al., 2013; Zhou and Li, 2014; Goldsmith et al., 2014; Wang et al., 2014). By contrast, our proposal *reverses* the role by treating the image tensor as response and the vector of covariates as predictors. The two treatments yield different interpretations. The former, the tensor predictor regression, focuses on understanding the change of a clinical outcome as the tensor image varies, so may be used for disease diagnosis and prognosis given image patterns. The latter, the tensor response regression, aims to study the change of the image as the predictors such as the disease status and age vary, and thus offers a more direct solution if the scientific interest is to identify brain regions exhibiting different activity patterns across different groups of subjects. In addition, the technique proposed in this article is completely different from the techniques used in tensor predictor regression, and as we will later show in the simulations, tensor response regression exhibits a competitive finite sample performance when the sample size is small.

In this article, we propose a parsimonious tensor response regression model and develop a novel estimation approach. Specifically, we continue to impose a linear association between the tensor response and the predictors. Meanwhile, we adopt a form of sparsity principle by assuming that part of the tensor response does not depend on the predictors and does not affect the rest of the response either. This principle is

biologically reasonable, as it is commonly believed that only a number of regions are usually affected by a particular disease. For instance, it is well known that hippocampus and entorhinal cortex are the most and earliest affected regions by the Alzheimer's disease (Du et al., 2001). Adopting this sparsity principle effectively reduces the number of free parameters, leads to a parsimonious model with improved interpretation, and yields a coefficient estimator that is asymptotically efficient. This principle can finds its natural counterpart in the sparsity principle in regression and variable selection with high-dimensional predictors, where only a subset of variables are believed to be relevant to the response. However, our proposal significantly differs from the popular sparse model estimation and selection in several ways. While the usual sparsity principle frequently adopted in variable selection assumes a subset of *individual* variables are irrelevant, we assume that the *linear combinations* are irrelevant to regression. Rather than using L_1 type penalty functions to induce sparsity, as is often done in variable selection, we employ a nascent technique called the *envelope* method (Cook et al., 2010) to estimate the unknown regression coefficient. Moreover, whereas most sparse modeling treats variable selection and parameter estimation separately, our envelope method essentially identifies and utilizes the material information in a joint estimation manner. We develop a fast estimation algorithm and study the asymptotic properties of the estimator. We demonstrate through both simulations and real data analyses that the new estimator improves dramatically over some alternative solutions.

The contributions of this article are multi-fold. First, it addresses a family of questions of substantial scientific interest but with relatively few solutions. A particular application is to compare tensor images across groups adjusting for potentially confounding covariates, which is of central interest in neuroimaging analysis. For instance, to evaluate effectiveness of a potential drug for Alzheimer's disease, it is crucial to compare the brain activity patterns between the drug and placebo groups while adjusting for age and gender effects. Our proposal offers a useful solution to this important problem, by systematically and jointly modeling all voxels of a tensor image given a vector of predictors. Second, while existing regularization solutions rely largely on penalty functions, our envelope based method provides an alternative way of introducing regularization

into estimation. It complements the usual penalty function based solutions, and usefully expands the realm of regularized estimation in general. Moreover, our method can be naturally coupled with penalty functions for further regularization. Third, our work advances the recent development of envelope method that was first proposed by Cook et al. (2010) then further developed in a series of papers (Su and Cook, 2011, 2012, 2013; Cook et al., 2013, 2015; Cook and Zhang, 2015a,b). Whilst all existing envelope methods concentrate on a scalar or vector response, our work differs obviously by tackling a tensor response. Such an extension is far from trivial, and new techniques are required throughout its development, even though to make our proposal easier to comprehend, we have chosen to present our method in a way that is parallel to that for a vector response. Furthermore, since the envelope methodology is new and sometimes uneasy to follow, we strive to connect it with the more familiar sparsity principle and clearly outline its assumptions, gains and limitations.

The rest of the article is organized as follows. Section 2 reviews key tensor notations and operations, and summarizes the envelope method for multivariate vector response regression. Section 3 proposes tensor response linear model, the generalized sparsity principle, then the concept of tensor envelope. Section 4 develops two estimators, and Section 5 studies their asymptotic properties. Simulations and real data analyses are presented in Sections 6 and 7, followed by a discussion in Section 8. All technical proofs are relegated to the Supplementary Materials.

2 Preparations

2.1 Tensor notations and operations

We begin with a quick review of some tensor notations and operations that are to be intensively used in this article. See also Kolda and Bader (2009) for an excellent review.

Multidimensional array $\boldsymbol{A} \in \mathbb{R}^{r_1 \times \cdots \times r_m}$ is called an *mth-order tensor*. The order of a tensor is also known as *dimension*, *way* or *mode*. A *fiber* is the higher order analogue of matrix row and column, and is defined by fixing every index of the tensor but one. A matrix column is a mode-1 fiber and a row is a mode-2 fiber.

The vec($\boldsymbol{A}$) *operator* stacks the entries of a tensor into a column vector, so that an entry $a_{i_1 \ldots i_m}$ of $\boldsymbol{A}$ maps to the j-th entry of vec($\boldsymbol{A}$), in which $j = 1 + \sum_{k=1}^{m}(i_k - 1)\prod_{k'=1}^{k-1} r_{k'}$. The *mode-k matricization*, $\boldsymbol{A}_{(k)}$, maps a tensor $\boldsymbol{A}$ into a matrix, denoted by $\boldsymbol{A}_{(k)} \in \mathbb{R}^{r_k \times (\prod_{j \neq k} r_j)}$, so that the $(i_1, \ldots, i_m)$ element of $\boldsymbol{A}$ maps to the (i_k, j) element of the matrix $\boldsymbol{A}_{(k)}$, where $j = 1 + \sum_{k' \neq k}(i_{k'} - 1)\prod_{k'' < k', k'' \neq k} r_{k''}$. The *k-mode product* of a tensor $\boldsymbol{A}$ and a matrix $\boldsymbol{C} \in \mathbb{R}^{s \times r_k}$ results in an mth-order tensor denoted as $\boldsymbol{A} \times_k \boldsymbol{C} \in \mathbb{R}^{r_1 \times \cdots \times r_{k-1} \times s \times r_{k+1} \times \cdots \times r_m}$, where each element is the product of mode-k fiber of $\boldsymbol{A}$ multiplied by $\boldsymbol{C}$. Similarly, the *k-mode vector product* of a tensor $\boldsymbol{A}$ and a vector $\boldsymbol{c} \in \mathbb{R}^{r_k}$ results in an $(m-1)$th-order tensor denoted as $\boldsymbol{A} \bar{\times}_k \boldsymbol{c} \in \mathbb{R}^{r_1 \times \cdots \times r_{k-1} \times r_{k+1} \times \cdots \times r_m}$, where each element is the inner product of each mode-k fiber of $\boldsymbol{A}$ with the vector $\boldsymbol{c}$.

The *Tucker decomposition* of a tensor is defined as $\boldsymbol{A} = \boldsymbol{C} \times_1 \boldsymbol{\Gamma}^{(1)} \times_2 \cdots \times_m \boldsymbol{\Gamma}^{(m)}$, where $\boldsymbol{C} \in \mathbb{R}^{u_1 \times \cdots \times u_m}$ is the *core tensor*, and $\boldsymbol{\Gamma}^{(k)} \in \mathbb{R}^{r_k \times u_k}$, $k = 1, \ldots, m$, are the factor matrices. It is a low-rank decomposition of the original tensor $\boldsymbol{A}$. For convenience, the Tucker decomposition is often represented by a shorthand, $[\![\boldsymbol{C}; \boldsymbol{\Gamma}^{(1)}, \ldots, \boldsymbol{\Gamma}^{(m)}]\!]$.

2.2 Multivariate response envelope model

Next we briefly review the multivariate linear model with vector-valued response, along with some key concepts of envelope, and with two goals in mind. First, it is to assist with a better understanding of the envelope methods in general, and second, to facilitate the construction of envelopes for tensor-valued response regression.

We start with the classical multivariate response linear model,

$$\boldsymbol{Y} = \boldsymbol{\beta}\boldsymbol{X} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{Y} \in \mathbb{R}^r$ is a response vector, $\boldsymbol{X} \in \mathbb{R}^p$ is a predictor vector, $\boldsymbol{\beta} \in \mathbb{R}^{r \times p}$ is the coefficient matrix, while the intercept is set to zero by centering the samples of $\boldsymbol{Y}$ and $\boldsymbol{X}$, and $\boldsymbol{\varepsilon} \in \mathbb{R}^r$ is the i.i.d. error that is independent of $\boldsymbol{X}$. It is often assumed that $\boldsymbol{\varepsilon}$ follows a multivariate normal distribution with mean zero and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{r \times r}$, $\boldsymbol{\Sigma} > 0$, though normality is not essential.

The envelope model (Cook et al., 2010) builds upon a key assumption that some aspects of the response vector are stochastically constant as the predictors vary. In

other words, part of the response is irrelevant to the regression. More rigorously, we assume there exists a subspace $\mathcal{S}$ of $\mathbb{R}^r$ such that

$$\boldsymbol{Q}_{\mathcal{S}}\boldsymbol{Y}|\boldsymbol{X} \sim \boldsymbol{Q}_{\mathcal{S}}\boldsymbol{Y}, \qquad \boldsymbol{Q}_{\mathcal{S}}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{P}_{\mathcal{S}}\boldsymbol{Y}|\boldsymbol{X}, \tag{2}$$

where $\boldsymbol{P}_{\mathcal{S}}$ is the projection matrix onto $\mathcal{S}$, $\boldsymbol{Q}_{\mathcal{S}} = \boldsymbol{I}_r - \boldsymbol{P}_{\mathcal{S}}$ is the projection onto the complement space $\mathcal{S}^{\perp}$, $\sim$ means identically distributed, and $\perp\!\!\!\perp$ means statistical independence. To better understand this assumption, we introduce a basis system of $\mathcal{S}$. Let $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times u}$ denote a basis matrix of $\mathcal{S}$, where u is the dimension of $\mathcal{S}$, $u \leq r$, and $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{r\times(r-u)}$ a basis of $\mathcal{S}^{\perp}$. Then (2) essentially states that the linear combinations $\boldsymbol{\Gamma}_0^{\intercal}\boldsymbol{Y} \in \mathbb{R}^{r-u}$ are immaterial to the estimation of $\boldsymbol{\beta}$ in that it responds neither to changes in the predictors nor to those in the rest of the response $\boldsymbol{\Gamma}^{\intercal}\boldsymbol{Y} \in \mathbb{R}^{u}$. Correspondingly, $\boldsymbol{\Gamma}^{\intercal}\boldsymbol{Y}$ carry all the material information in $\boldsymbol{Y}$, and intuitively, one can focus on $\boldsymbol{\Gamma}^{\intercal}\boldsymbol{Y}$ in subsequent regression modeling. We remark that, assumption (2), although looks somewhat unfamiliar, can find its natural counterpart in the well known and commonly adopted *sparsity principle* in classical variable selection, where only a subset of variables are assumed to be relevant. The two assumptions, at heart, share exactly the same spirit that only part of information is deemed useful for regressions and the rest irrelevant. However, they are also different in that, whereas the usual sparsity principle focuses on *individual* variables, (2) permits *linear combination* of the variables to be irrelevant. For this reason, we term assumption (2) as the *generalized sparsity principle*. Compared to the usual sparsity principle, it is more flexible, but could lose some interpretability.

To see how the generalized sparsity principle would facilitate estimation of $\boldsymbol{\beta}$ in model (1), we note that the following decompositions hold true under (2).

$$\text{span}(\boldsymbol{\beta}) \subseteq \mathcal{S} \;\text{ and }\; \boldsymbol{\Sigma} = \text{var}(\boldsymbol{P}_{\mathcal{S}}\boldsymbol{Y}) + \text{var}(\boldsymbol{Q}_{\mathcal{S}}\boldsymbol{Y}) = \boldsymbol{P}_{\mathcal{S}}\boldsymbol{\Sigma}\boldsymbol{P}_{\mathcal{S}} + \boldsymbol{Q}_{\mathcal{S}}\boldsymbol{\Sigma}\boldsymbol{Q}_{\mathcal{S}},$$

where $\text{span}(\boldsymbol{\beta})$ is the column subspace of $\boldsymbol{\beta}$, i.e. the subspace spanned by the columns of $\boldsymbol{\beta}$. Accordingly, we can rewrite the above decompositions in terms of the basis matrices,

$$\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\theta} \;\text{ and }\; \boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^{\intercal} + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^{\intercal}, \tag{3}$$

where $\boldsymbol{\theta} = \boldsymbol{\Gamma}^{\intercal}\boldsymbol{\beta} \in \mathbb{R}^{u\times p}$ denotes the coordinates of $\boldsymbol{\beta}$ relative to the basis $\boldsymbol{\Gamma}$, $\boldsymbol{\Omega} = \text{cov}(\boldsymbol{\Gamma}^{\intercal}\boldsymbol{Y} \mid \boldsymbol{X}) \in \mathbb{R}^{u\times u}$ is the material variation, and $\boldsymbol{\Omega}_0 = \text{cov}(\boldsymbol{\Gamma}_0^{\intercal}\boldsymbol{Y} \mid \boldsymbol{X}) = \text{cov}(\boldsymbol{\Gamma}_0^{\intercal}\boldsymbol{Y}) \in$

$\mathbb{R}^{(r-u)\times(r-u)}$ is the immaterial variation that contains no information about $\boldsymbol{Y}|\boldsymbol{X}$, but only brings extraneous variation in estimation.

Given the first result of (3), we note that model (1) can be rewritten as

$$\boldsymbol{\Gamma}^\intercal \boldsymbol{Y} = \boldsymbol{\theta}\boldsymbol{X} + \boldsymbol{\Gamma}^\intercal\boldsymbol{\varepsilon}, \text{ and } \boldsymbol{\Gamma}_0^\intercal \boldsymbol{Y} = \boldsymbol{\Gamma}_0^\intercal\boldsymbol{\varepsilon}. \tag{4}$$

In turn, (4) implies that regression modeling can now focus on the material part $\boldsymbol{\Gamma}^\intercal\boldsymbol{Y}$ only. The effective number of parameters in model (1) is reduced from $pr + r(r+1)/2$ without assumption (2), to $pu + (r-u)u + u(u+1)/2 + (r-u)(r-u+1)/2$ with the assumption, and the difference is $p(r-u)$.

Given the second result of (3), Cook et al. (2010) showed the gain in estimation efficiency for $\boldsymbol{\beta}$. Let $\widehat{\boldsymbol{\beta}}_{\text{ENV}}$ denote the estimator of $\boldsymbol{\beta}$ in (1) under (2), $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ the ordinary least squares estimator without imposing assumption (2), and $\boldsymbol{\beta}_{\text{TRUE}}$ the true value of $\boldsymbol{\beta}$. Then it is shown that, both $\sqrt{n}\left\{\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{OLS}}) - \text{vec}(\boldsymbol{\beta}_{\text{TRUE}})\right\}$ and $\sqrt{n}\left\{\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{ENV}}) - \text{vec}(\boldsymbol{\beta}_{\text{TRUE}})\right\}$ converge to a normal vector with mean zero and covariance matrix,

$$\text{avar}\left\{\sqrt{n}\,\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})\right\} = \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\otimes\boldsymbol{\Sigma}, \text{ and avar}\left\{\sqrt{n}\,\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{ENV}})\right\} = \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\otimes(\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^\intercal) + \boldsymbol{\Delta},$$

respectively, where $\boldsymbol{\Sigma}_{\boldsymbol{X}} = \text{cov}(\boldsymbol{X})$, the first term in $\text{avar}\{\sqrt{n}\,\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{ENV}})\}$ corresponds to the asymptotic variance of the estimator given that $\mathcal{S}$ is known, and the second term $\boldsymbol{\Delta}$ is the asymptotic cost of estimating $\mathcal{S}$. While Cook et al. (2010) showed that $\text{avar}\{\sqrt{n}\,\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{ENV}})\} \le \text{avar}\{\sqrt{n}\,\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})\}$ in general, in the light of decomposition of $\boldsymbol{\Sigma}$ in (3), it is straightforward to see that the first term in $\text{avar}\{\sqrt{n}\,\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{ENV}})\}$ is to be substantially smaller than $\text{avar}\{\sqrt{n}\,\text{vec}(\widehat{\boldsymbol{\beta}}_{\text{OLS}})\}$, if the immaterial variation $\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^\intercal$ dominates the material variation $\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^\intercal$.

Finally, we address the issue of existence and uniqueness of $\mathcal{S}$ in (2). The subspace $\mathcal{S}$ always exists, as it can trivially take the form of $\mathbb{R}^r$. However, $\mathcal{S}$ is not unique. Then the idea is to seek the *intersection* of all subspaces that satisfy (2), which is minimum and unique. Toward that end, Cook et al. (2010) gave two generic definitions.

Definition 1. *A subspace $\mathcal{R}\subseteq\mathbb{R}^p$ is said to be a reducing subspace of $\boldsymbol{M}\in\mathbb{R}^{p\times p}$ if $\mathcal{R}$ satisfies that $\boldsymbol{M} = \boldsymbol{P}_{\mathcal{R}}\boldsymbol{M}\boldsymbol{P}_{\mathcal{R}} + \boldsymbol{Q}_{\mathcal{R}}\boldsymbol{M}\boldsymbol{Q}_{\mathcal{R}}$.*

Definition 2. *Let $\boldsymbol{M}\in\mathbb{R}^{p\times p}$ and $\mathcal{B}\subseteq$ span$(\boldsymbol{M})$. Then the $\boldsymbol{M}$-envelope of $\mathcal{B}$, denoted by $\mathcal{E}_{\boldsymbol{M}}(\mathcal{B})$, is the intersection of all reducing subspaces of $\boldsymbol{M}$ that contain $\mathcal{B}$.*

Given those definitions, we see that any subspace $\mathcal{S}$ satisfying (2) under model (1) is a reducing subspace of $\boldsymbol{\Sigma}$, and the intersection of all such subspaces is the $\boldsymbol{\Sigma}$-envelope of $\mathcal{B} = \text{span}(\boldsymbol{\beta})$. This envelope $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ is also denoted by $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$, and uniquely exists. In our envelope based estimation, we seek the estimation of $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$ so to improve estimation of the coefficient matrix $\boldsymbol{\beta}$.

3 Models

3.1 Tensor response linear model

When facing a tensor response, we develop a model in analogous to the classical multivariate model (1). That is, for an mth order tensor response $\boldsymbol{Y} \in \mathbb{R}^{r_1 \times \cdots \times r_m}$, and a vector of predictors $\boldsymbol{X} \in \mathbb{R}^p$, consider the tensor response linear model

$$\boldsymbol{Y} = \boldsymbol{B} \bar{\times}_{(m+1)} \boldsymbol{X} + \boldsymbol{\varepsilon}. \tag{5}$$

In this model, the coefficient $\boldsymbol{B} \in \mathbb{R}^{r_1 \times \cdots \times r_m \times p}$ is an $(m+1)$th order tensor that captures the interrelationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$ and is the parameter of interest in our estimation. $\bar{\times}_{(m+1)}$ is the $(m+1)$-mode vector product. The intercept term is again omitted without losing any generality. The error $\boldsymbol{\varepsilon} \in \mathbb{R}^{r_1 \times \cdots \times r_m}$ is an mth order tensor that is independent of $\boldsymbol{X}$ and has mean zero. Furthermore, we assume that $\boldsymbol{\varepsilon}$ has a *separable Kronecker covariance structure* such that $\text{cov}\{\text{vec}(\boldsymbol{\varepsilon})\} = \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_m \otimes \cdots \otimes \boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_k > 0, k = 1, \ldots, m$. This separable covariance assumption is important to help reduce the number of free parameters in $\boldsymbol{\Sigma}$, which is part of our envelope estimation. Meanwhile, this separable structure has been fairly commonly imposed in the tensor literature; see for instance, Hoff (2011); Fosdick and Hoff (2014). Here to avoid notation proliferation, we continue to use $\boldsymbol{\Sigma}$ to denote the covariance matrix, as it should not cause any confusion in the context. The distribution of $\text{vec}(\boldsymbol{\varepsilon})$ is assumed to be normal, which enables likelihood estimation. However, normality is not essential, and moment based estimation can replace likelihood estimation when the normality assumption is in question.

Two special cases of model (5) are worth of brief mentioning. The first is when $\boldsymbol{X}$ is a scalar and takes the value of only 0 or 1. In this case, $\boldsymbol{B}$ reduces to an mth order tensor that can be interpreted as the mean difference of the tensor coefficients between

the two populations. The second case is when $m = 1$, where the response $\boldsymbol{Y}$ becomes a vector, and model (5) reduces to the classical multivariate linear model (1). In this case, $\boldsymbol{B}\bar{\times}_{(m+1)}\boldsymbol{X}$ becomes the inner product of each mode-2 fiber (i.e., row) of $\boldsymbol{B}$ with $\boldsymbol{X}$, which in turn is the usual matrix product of $\boldsymbol{B}$ and $\boldsymbol{X}$.

Next we consider an alternative tensor response linear model (5),

$$\mathrm{vec}(\boldsymbol{Y}) = \boldsymbol{B}_{(m+1)}^{\intercal}\boldsymbol{X} + \mathrm{vec}(\boldsymbol{\varepsilon}). \tag{6}$$

This model can be viewed as the *vectorized* form of model (5). However, the main difference is that *no* separable covariance structure is imposed on the error term $\boldsymbol{\varepsilon}$ in (6). The coefficient matrix $\boldsymbol{B}_{(m+1)} \in \mathbb{R}^{p\times\prod_{k=1}^{m} r_k}$ can be interpreted as the mode-$(m+1)$ matricization of the tensor coefficient $\boldsymbol{B}$ in (5). Each column of $\boldsymbol{B}_{(m+1)}$ is a $p \times 1$ coefficient vector that characterizes the linear relationship between each individual element of $\boldsymbol{Y}$ and the predictor vector $\boldsymbol{X}$. Therefore, estimating $\boldsymbol{B}_{(m+1)}$ in (6) is equivalent to estimating $\boldsymbol{B}$ in (5) by fitting individual elements of $\boldsymbol{Y}$ on $\boldsymbol{X}$ *one-at-a-time*. We call this estimator the ordinary least squares estimator of $\boldsymbol{B}$, and denote it by $\widehat{\boldsymbol{B}}_{\mathrm{OLS}}$. Given the data $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}_{i=1}^{n}$, it has an explicit form

$$\widehat{\boldsymbol{B}}_{\mathrm{OLS}} = \mathbb{Y} \times_{(m+1)} \{(\mathbb{X}\mathbb{X}^{\intercal})^{-1}\mathbb{X}\},$$

where $\mathbb{X} \in \mathbb{R}^{p\times n}$ and $\mathbb{Y} \in \mathbb{R}^{r_1\times\ldots\times r_m\times n}$ are the *stacked* predictor matrix and response array, respectively. If $\mathrm{vec}(\boldsymbol{\varepsilon})$ is further assumed to be normally distributed, then the above OLS estimator is also the maximum likelihood estimator based on model (6).

3.2 Generalized sparsity principle

For a tensor response, we expect a similar generalized sparsity principle like (2) to hold true. It is probably more so than the vector response case, as intuitively it is natural to expect certain regions of the tensor response to be immaterial. More specifically, suppose there exist a series of subspaces, $\mathcal{S}_k \subseteq \mathbb{R}^{r_k}$, $k = 1, \ldots, m$, such that

$$\boldsymbol{Y} \times_k \boldsymbol{Q}_k | \boldsymbol{X} \sim \boldsymbol{Y} \times_k \boldsymbol{Q}_k, \quad \boldsymbol{Y} \times_k \boldsymbol{Q}_k \perp\!\!\!\perp \boldsymbol{Y} \times_k \boldsymbol{P}_k | \boldsymbol{X}, \quad k = 1, \ldots, m, \tag{7}$$

where $\boldsymbol{P}_k \in \mathbb{R}^{r_k\times r_k}$ is the projection matrix onto $\mathcal{S}_k$, $\boldsymbol{Q}_k = \boldsymbol{I}_{r_k} - \boldsymbol{P}_k \in \mathbb{R}^{r_k\times r_k}$ is the projection onto the complement space $\mathcal{S}_k^{\perp}$, and $\times_k$ denotes the k-mode product. Then,

the first condition in (7) essentially states that $\boldsymbol{Y} \times_k \boldsymbol{Q}_k$ does not depend on $\boldsymbol{X}$, while the second condition in (7) says $\boldsymbol{Y} \times_k \boldsymbol{Q}_k$ does not affect the rest of the response, $\boldsymbol{Y} \times_k \boldsymbol{P}_k$, and there is no information leak between $\boldsymbol{Y} \times_k \boldsymbol{Q}_k$ and $\boldsymbol{Y} \times_k \boldsymbol{P}_k$. As such, we call $\boldsymbol{Y} \times_k \boldsymbol{Q}_k$ the immaterial information to the regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$, and call $\boldsymbol{Y} \times_k \boldsymbol{P}_k$ the material information. Combining the statements in (7) for all $k = 1, \ldots, m$, we arrive at a parsimonious representation:

$$\mathbb{Q}(\boldsymbol{Y})|\boldsymbol{X} \sim \mathbb{Q}(\boldsymbol{Y}), \qquad \mathbb{Q}(\boldsymbol{Y}) \perp\!\!\!\perp \mathbb{P}(\boldsymbol{Y})|\boldsymbol{X},$$

where $\mathbb{Q}(\boldsymbol{Y}) = \boldsymbol{Y} - \mathbb{P}(\boldsymbol{Y}) \in \mathbb{R}^{r_1 \times \cdots \times r_m}$, and $\mathbb{P}(\boldsymbol{Y}) = [\![\boldsymbol{Y}; \boldsymbol{P}_1, \ldots, \boldsymbol{P}_m]\!] \in \mathbb{R}^{r_1 \times \cdots \times r_m}$, i.e., a Tucker decomposition with $\boldsymbol{Y}$ as the core tensor, and $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_m$ as the factor matrices along each mode. This provides a decomposition of $\boldsymbol{Y}$, $\boldsymbol{Y} = \mathbb{P}(\boldsymbol{Y}) + \mathbb{Q}(\boldsymbol{Y})$, into the material part $\mathbb{P}(\boldsymbol{Y})$ and the immaterial part $\mathbb{Q}(\boldsymbol{Y})$.

Introducing (7) to the tensor response linear model (5), we have the following results.

Proposition 1. *Under the tensor response linear model* (5)*, the assumption* (7) *is true if and only if*

$$\boldsymbol{B} \times_k \boldsymbol{Q}_k = 0 \quad \textit{and} \quad \boldsymbol{\Sigma}_k = \boldsymbol{P}_k \boldsymbol{\Sigma}_k \boldsymbol{P}_k + \boldsymbol{Q}_k \boldsymbol{\Sigma}_k \boldsymbol{Q}_k, \quad k = 1, \ldots, m.$$

To turn the above decompositions of $\boldsymbol{B}$ and $\boldsymbol{\Sigma}_k$ into a basis representation, let $\boldsymbol{\Gamma}_k \in \mathbb{R}^{r_k \times u_k}$ denote a basis for $\mathcal{S}_k$, where u_k is the dimension of $\mathcal{S}_k$, and $\boldsymbol{\Gamma}_{0k} \in \mathbb{R}^{r_k \times (r_k - u_k)}$ denote the complement basis, $k = 1, \ldots, m$. Let $\boldsymbol{\Omega}_k \in \mathbb{R}^{u_k \times u_k}$ and $\boldsymbol{\Omega}_{0k} \in \mathbb{R}^{(r_k - u_k) \times (r_k - u_k)}$ denote two symmetric positive definite matrices. Then we have $\boldsymbol{P}_k = \boldsymbol{\Gamma}_k \boldsymbol{\Gamma}_k^\top$, $\boldsymbol{Q}_k = \boldsymbol{\Gamma}_{0k} \boldsymbol{\Gamma}_{0k}^\top$, plus the following parameterization for $\boldsymbol{B}$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_m \otimes \cdots \otimes \boldsymbol{\Sigma}_1$.

Proposition 2. *The parameterization in Proposition 1 is equivalent to the following coordinate representations:*

$$\begin{aligned} \boldsymbol{B} &= [\![\boldsymbol{\Theta}; \boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_m, \boldsymbol{I}_p]\!] \quad \textit{for some } \boldsymbol{\Theta} \in \mathbb{R}^{u_1 \times \ldots \times u_m \times p}, \\ \boldsymbol{\Sigma}_k &= \boldsymbol{\Gamma}_k \boldsymbol{\Omega}_k \boldsymbol{\Gamma}_k^\top + \boldsymbol{\Gamma}_{0k} \boldsymbol{\Omega}_{0k} \boldsymbol{\Gamma}_{0k}^\top, \quad k = 1, \ldots, m. \end{aligned}$$

Accordingly, one can rewrite the material response part $\mathbb{P}(\boldsymbol{Y})$ in the following way

$$\mathbb{P}(\boldsymbol{Y}) = [\![\boldsymbol{Y}; \boldsymbol{\Gamma}_1 \boldsymbol{\Gamma}_1^\top, \ldots, \boldsymbol{\Gamma}_m \boldsymbol{\Gamma}_m^\top]\!] = [\![[\![\boldsymbol{Y}; \boldsymbol{\Gamma}_1^\top, \ldots, \boldsymbol{\Gamma}_m^\top]\!]; \boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_m]\!],$$

where the *core tensor* is $\boldsymbol{Z} = [\![\boldsymbol{Y}; \boldsymbol{\Gamma}_1^\intercal, \ldots, \boldsymbol{\Gamma}_m^\intercal]\!] \in \mathbb{R}^{u_1 \times \cdots \times u_m}$. We see that, by recognizing and focusing on the material part of the tensor response $\mathbb{P}(\boldsymbol{Y})$, the regression modeling can now focus on the core tensor $\boldsymbol{Z}$ as the "surrogate response", which plays a similar role as $\boldsymbol{\Gamma}^\intercal \boldsymbol{Y}$ in (4) in the vector response case. Meanwhile, the key parameter to estimate becomes $\boldsymbol{\Theta} \in \mathbb{R}^{u_1 \times \cdots \times u_m \times p}$, along with $\{\boldsymbol{\Gamma}_k\}_{k=1}^m$, $\{\boldsymbol{\Omega}_k\}_{k=1}^m$ and $\{\boldsymbol{\Omega}_{0k}\}_{k=1}^m$. Consequently, the number for free parameters reduces from $p\prod_{k=1}^m r_k + \sum_{k=1}^m r_k(r_k+1)/2$ to $p\prod_{k=1}^m u_k + \sum_{k=1}^m \{u_k(r_k-u_k) + u_k(u_k+1)/2 + (r_k-u_k)(r_k-u_k+1)/2\}$, and in effect saving $p\big\{\prod_{k=1}^m r_k - \prod_{k=1}^m u_k\big\}$ parameters. With u_k usually being much smaller than r_k, substantial dimension reduction is achieved, which in turn improves the estimation.

3.3 Tensor envelope

Similar to the vector case, we next develop the notion of *tensor envelope* for tensor response model (5) to attain uniqueness of the subspaces $\mathcal{S}_k$ under the generalized sparsity principle (7). Unlike the vector case, however, there are two different ways to construct a tensor envelope. We will define the new concept in one way, then establish its equivalence with the other. Moreover, we will lay out the difference between the proposed tensor envelope and the vector envelope constructed based on the vectorized model (6).

One way to establish the tensor envelope for model (5) is to recognize that it should contain $\text{span}(\boldsymbol{B}_{(m+1)}^\intercal)$, meanwhile it should reduce the covariance $\boldsymbol{\Sigma}$ and respect the separable Kronecker covariance structure that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_m \otimes \cdots \otimes \boldsymbol{\Sigma}_1$. Then following Definitions 1 and 2, we come to the next definition of the tensor envelope.

Definition 3. *The tensor envelope for $\boldsymbol{B}$ in the tensor response linear model* (5)*, denoted by $\mathcal{T}_{\boldsymbol{\Sigma}}(\boldsymbol{B})$, is defined as the intersection of all reducing subspaces $\mathcal{E}$ of $\boldsymbol{\Sigma}$ that contains span$(\boldsymbol{B}_{(m+1)}^\intercal)$ and can be written as $\mathcal{E} = \mathcal{E}_m \otimes \cdots \otimes \mathcal{E}_1$, where $\mathcal{E}_k \subseteq \mathbb{R}^{r_k}$, $k = 1, \ldots, m$.*

Following this definition, we see that $\mathcal{T}_{\boldsymbol{\Sigma}}(\boldsymbol{B})$ is minimum and unique, and is central to our envelope based estimation of $\boldsymbol{B}$. Moreover, under the special case that $m = 1$ and the response is a vector, the tensor envelope $\mathcal{T}_{\boldsymbol{\Sigma}}(\boldsymbol{B})$ reduces to the envelope notion $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{B})$ for the vector response.

An alternative way to construct the tensor envelope is by noting that, due to the decomposition in Proposition 1, one can construct a series of envelopes, $\mathcal{E}_{\boldsymbol{\Sigma}_k}\left(\boldsymbol{B}_{(k)}\right)$, for $k = 1, \ldots, m$, that satisfy the generalized sparsity principle (7) under model (5). That is, $\mathcal{E}_{\boldsymbol{\Sigma}_k}\left(\boldsymbol{B}_{(k)}\right)$ is the smallest subspace $\mathcal{S}_k$ that contains $\text{span}(\boldsymbol{B}_{(k)})$ and reduces $\boldsymbol{\Sigma}_k$, $k = 1, \ldots, m$. Then one can construct a tensor envelope by combining $\left\{\mathcal{E}_{\boldsymbol{\Sigma}_k}\left(\boldsymbol{B}_{(k)}\right)\right\}_{k=1}^{m}$ to capture all the material information in the response. Naturally, the two ways of constructing the tensor envelope are well connected, due to the next equivalence property.

Proposition 3. *The tensor envelope as defined in Definition 3 satisfies that* $\mathcal{T}_{\boldsymbol{\Sigma}}(\boldsymbol{B}) = \mathcal{E}_{\boldsymbol{\Sigma}_m}\left(\boldsymbol{B}_{(m)}\right) \otimes \cdots \otimes \mathcal{E}_{\boldsymbol{\Sigma}_1}\left(\boldsymbol{B}_{(1)}\right)$.

Our estimation of the tensor envelope utilizes this result by first estimating a basis of the individual envelope $\mathcal{E}_{\boldsymbol{\Sigma}_k}\left(\boldsymbol{B}_{(k)}\right)$, then combining them by Kronekcer product to construct an estimate of the tensor envelope.

Finally, we remark that, in principle, one can develop an envelope for the ordinary least squares estimator in model (6) as well. By analogy to the envelope definition for a vector response, this envelope also contains $\text{span}(\boldsymbol{B}_{(m+1)}^{\intercal})$, and thus we denote it by $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{B}_{(m+1)})$. However, there are some important differences between $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{B}_{(m+1)})$ and the tensor envelope $\mathcal{T}_{\boldsymbol{\Sigma}}(\boldsymbol{B})$ in Definition 3. First, $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{B}_{(m+1)})$ does not take into account the separable covariance structure, nor can be decomposed into the Kronecker product of the individual envelopes $\mathcal{E}_{\boldsymbol{\Sigma}_k}\left(\boldsymbol{B}_{(k)}\right)$. Second, the computation involved in estimating $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{B}_{(m+1)})$ is prohibitive, as it replies on the estimation of the unstructured covariance matrix of $\text{vec}(\boldsymbol{\varepsilon}) \in \mathbb{R}^{\prod_{k=1}^{m} r_k \times \prod_{k=1}^{m} r_k}$. By contrast, the computation of $\mathcal{T}_{\boldsymbol{\Sigma}}(\boldsymbol{B})$ is much more feasible, as we demonstrate in the next section.

4 Estimation

Our primary target of estimation is $\boldsymbol{B}$ in the tensor response linear model (5). Our proposal is to estimate $\boldsymbol{B}$ through the tensor envelope $\mathcal{T}_{\boldsymbol{\Sigma}}(\boldsymbol{B})$, which also involves estimation of $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_m \otimes \cdots \otimes \boldsymbol{\Sigma}_1$. The objective function to minimize is of the form,

$$\ell(\boldsymbol{B}, \boldsymbol{\Sigma}) = \log|\boldsymbol{\Sigma}| + n^{-1} \sum_{i=1}^{n} \left\{\text{vec}(\boldsymbol{Y}_i) - \boldsymbol{B}_{(m+1)}^{\intercal} \boldsymbol{X}_i\right\}^{\intercal} \boldsymbol{\Sigma}^{-1} \left\{\text{vec}(\boldsymbol{Y}_i) - \boldsymbol{B}_{(m+1)}^{\intercal} \boldsymbol{X}_i\right\}. \quad (8)$$

Algorithm 1 Iterative optimization algorithm for minimizing $\ell(\boldsymbol{B}, \boldsymbol{\Sigma})$.

[1] Initialize $\boldsymbol{B}^{(0)}$ and $\boldsymbol{\Sigma}^{(0)} = \boldsymbol{\Sigma}_m^{(0)} \otimes \cdots \otimes \boldsymbol{\Sigma}_1^{(0)}$
repeat
  [2] Estimate envelope basis $\left\{\boldsymbol{\Gamma}_k^{(t+1)}\right\}_{k=1}^m$ given $\boldsymbol{B}^{(t)}$ and $\boldsymbol{\Sigma}^{(t)}$
  [3] Estimate parameters $\boldsymbol{\Theta}^{(t+1)}$, $\boldsymbol{\Omega}_k^{(t+1)}$ and $\boldsymbol{\Omega}_{0k}^{(t+1)}$ given $\left\{\boldsymbol{\Gamma}_k^{(t+1)}\right\}_{k=1}^m$.
  [4] Update $\boldsymbol{B}^{(t+1)}$ and $\boldsymbol{\Sigma}^{(t+1)} = \boldsymbol{\Sigma}_m^{(t+1)} \otimes \cdots \otimes \boldsymbol{\Sigma}_1^{(t+1)}$.
until the objective function converges

It is straightforward to verify that this objective function, aside from some constant, is the negative log-likelihood function of the model (5) if one assumes that the error follows a normal distribution. By adopting (7), then the parameter decompositions in Proposition 2, the minimization of $\ell(\boldsymbol{B}, \boldsymbol{\Sigma})$ becomes estimation of the envelope basis $\boldsymbol{\Gamma}_k \in \mathbb{R}^{r_k \times u_k}$, the reduced coefficient $\boldsymbol{\Theta} \in \mathbb{R}^{u_1 \times \ldots \times u_m \times p}$, and the matrices $\boldsymbol{\Omega}_k \in \mathbb{R}^{u_k \times u_k}$ and $\boldsymbol{\Omega}_{0k} \in \mathbb{R}^{(r_k - u_k) \times (r_k - u_k)}$, $k = 1, \ldots, m$. Here, with a slight abuse of notation, we continue to denote the dimension of the individual envelope $\mathcal{E}_{\boldsymbol{\Sigma}_k}\left(\boldsymbol{B}_{(k)}\right)$ as u_k.

We present two solutions, one an iterative estimator and the other a one-step estimator. The first solution alternates among steps of estimating one parameter given the rest fixed. It leads to a maximum likelihood estimator when the error follows a normal distribution and is a moment estimator otherwise. The second solution requires no iteration, and is essentially an approximate estimator, but it enjoys several appealing properties, both numerically and asymptotically.

4.1 Iterative estimator

We first summarize our iterative optimization of $\ell(\boldsymbol{B}, \boldsymbol{\Sigma})$ in Algorithm 1. We then give details for each individual step. Updating equations in each step are carefully derived as partial maximized likelihood estimators under the normal error assumption, with the detailed derivation given in the Supplementary Materials. As a result, the objective function is monotonically decreasing, guaranteeing the convergence of the algorithm.

The first step of Algorithm 1 is to initialize $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$. For $\boldsymbol{B}$, a natural initial estimator is the OLS estimator $\widehat{\boldsymbol{B}}_{\text{OLS}}$ in (7). That is, we fit each element of the tensor response $\boldsymbol{Y}$ on $\boldsymbol{X}$ one-at-a-time, and set the resulting estimator as the initial estimator $\boldsymbol{B}^{(0)} = \widehat{\boldsymbol{B}}_{\text{OLS}}$. For $\boldsymbol{\Sigma}$, we employ the covariance estimator of Dutilleul (1999) and

Manceur and Dutilleul (2013). That is, for $k = 1, \ldots, m$, in turn, we set

$$\boldsymbol{\Sigma}_k^{(0)} = \frac{1}{n \prod_{j \neq k} r_j} \sum_{i=1}^{n} \boldsymbol{e}_{i(k)} \left\{ (\boldsymbol{\Sigma}_m^{(0)})^{-1} \otimes \ldots \otimes (\boldsymbol{\Sigma}_{k+1}^{(0)})^{-1} \otimes (\boldsymbol{\Sigma}_{k-1}^{(0)})^{-1} \otimes \ldots \otimes (\boldsymbol{\Sigma}_1^{(0)})^{-1} \right\} \boldsymbol{e}_{i(k)}^{\intercal}, \quad (9)$$

where $\boldsymbol{e}_{i(k)}$ is the mode-k matricization of the residual, $\boldsymbol{e}_i = \boldsymbol{Y}_i - \boldsymbol{B}^{(0)} \times_{(m+1)} \boldsymbol{X}_i, i = 1, \ldots, n$, for n sample replications, and the iterative update of each covariance $\boldsymbol{\Sigma}_k^{(0)}$ given the rest starts with $\boldsymbol{\Sigma}_j^{(0)} = \boldsymbol{I}_{r_j}$, $j \neq k$. One can verify that the above estimator is the maximum likelihood estimator if the error in (5) follows a normal distribution. We also remark that, the individual component $\boldsymbol{\Sigma}_k$ is not identifiable. To address this issue, we normalize $\boldsymbol{\Sigma}_k^{(0)}$ by dividing the term by its Frobenius norm, and update the final estimator $\boldsymbol{\Sigma}^{(0)} = \tau \boldsymbol{\Sigma}_m^{(0)} \otimes \ldots \otimes \boldsymbol{\Sigma}_1^{(0)}$, where the scalar $\tau = (n \prod_j r_j)^{-1} \sum_{i=1}^{n} \mathrm{vec}^{\intercal}(\boldsymbol{e}_i) \big\{ (\boldsymbol{\Sigma}_m^{(0)})^{-1} \otimes \cdots \otimes (\boldsymbol{\Sigma}_1^{(0)})^{-1} \big\} \mathrm{vec}(\boldsymbol{e}_i)$.

The second step of Algorithm 1 is to estimate the envelope basis $\{\boldsymbol{\Gamma}_k\}_{k=1}^m$. Update of $\boldsymbol{\Gamma}_k^{(t+1)}$, given the current estimates $\boldsymbol{B}^{(t)}$ and $\boldsymbol{\Sigma}^{(t)}$, can be achieved by minimizing the following objective function, subject to the constraint that $\boldsymbol{G}_k^{\intercal} \boldsymbol{G}_k = \boldsymbol{I}_{u_k}$,

$$f_k^{(t)}(\boldsymbol{G}_k) = \log |\boldsymbol{G}_k^{\intercal} \boldsymbol{M}_k^{(t)} \boldsymbol{G}_k| + \log |\boldsymbol{G}_k^{\intercal} (\boldsymbol{N}_k^{(t)})^{-1} \boldsymbol{G}_k|, \quad (10)$$

where $\boldsymbol{M}_k^{(t)} = (n \prod_{j \neq k} r_j)^{-1} \sum_{i=1}^{n} \boldsymbol{\delta}_{i(k)}^{(t)} \{ (\boldsymbol{\Sigma}_m^{(t)})^{-1} \otimes \cdots \otimes (\boldsymbol{\Sigma}_{k+1}^{(t)})^{-1} \otimes (\boldsymbol{\Sigma}_{k-1}^{(t)})^{-1} \otimes \cdots \otimes (\boldsymbol{\Sigma}_1^{(t)})^{-1} \} (\boldsymbol{\delta}_{i(k)}^{(t)})^{\intercal}$, $\boldsymbol{N}_k^{(t)} = (n \prod_{j \neq k} r_j)^{-1} \sum_{i=1}^{n} \boldsymbol{Y}_{i(k)} \{ (\boldsymbol{\Sigma}_m^{(t)})^{-1} \otimes \cdots \otimes (\boldsymbol{\Sigma}_{k+1}^{(t)})^{-1} \otimes (\boldsymbol{\Sigma}_{k-1}^{(t)})^{-1} \otimes \cdots \otimes (\boldsymbol{\Sigma}_1^{(t)})^{-1} \} \boldsymbol{Y}_{i(k)}^{\intercal}$, and $\boldsymbol{\delta}_i^{(t)}$ is the fitted envelope model residual were the envelope basis $\{\boldsymbol{\Gamma}_j\}_{j=1, j \neq k}^m$ known: $\boldsymbol{\delta}_i^{(t)} = \boldsymbol{Y}_i - [\![\widehat{\boldsymbol{B}}_{\mathrm{OLS}}; \boldsymbol{P}_{\boldsymbol{\Gamma}_1}^{(t)}, \ldots, \boldsymbol{P}_{\boldsymbol{\Gamma}_{k-1}}^{(t)}, \boldsymbol{I}_{r_k}, \boldsymbol{P}_{\boldsymbol{\Gamma}_{k+1}}^{(t)}, \ldots, \boldsymbol{P}_{\boldsymbol{\Gamma}_m}^{(t)}, \boldsymbol{I}_p]\!] \times_{(m+1)} \boldsymbol{X}_i^{\intercal}$, where $\boldsymbol{P}_{\boldsymbol{\Gamma}_j}^{(t)} \equiv \boldsymbol{\Gamma}_j^{(t)} (\boldsymbol{\Gamma}_j^{(t)})^{\intercal}$ is the projection onto subspace $\mathrm{span}(\boldsymbol{\Gamma}_j^{(t)})$ and $\boldsymbol{\delta}_{i(j)}^{(t)}$ is the mode-j matricization of $\boldsymbol{\delta}_i^{(t)}$. Then we have $\boldsymbol{\Gamma}_k^{(t+1)} = \arg\min_{\boldsymbol{G}_k} f_k^{(t)}(\boldsymbol{G}_k)$ subject to the orthogonal constraint $\boldsymbol{G}_k^{\intercal} \boldsymbol{G}_k = \boldsymbol{I}_{u_k}$. This optimization is over all $r_k \times u_k$ dimensional Grassmann manifolds since $f_k^{(t)}(\boldsymbol{G}_k) = f_k^{(t)}(\boldsymbol{G}_k \boldsymbol{O})$ for any orthogonal matrix $\boldsymbol{O} \in \mathbb{R}^{u_k \times u_k}$.

The third step of Algorithm 1 is to update $\boldsymbol{\Theta}$ and $\boldsymbol{\Omega}_k$ and $\boldsymbol{\Omega}_{0k}$ given the current estimate of the envelope basis $\{\boldsymbol{\Gamma}_k\}_{k=1}^m$. We first note that $\boldsymbol{\Theta}$ can be estimated by regressing the core tensor, $\boldsymbol{Z} = [\![\boldsymbol{Y}; \boldsymbol{\Gamma}_1^{\intercal}, \ldots, \boldsymbol{\Gamma}_m^{\intercal}]\!] \in \mathbb{R}^{u_1 \times \ldots \times u_m}$, on the predictor $\boldsymbol{X}$ through ordinary least squares without any constraint. That is,

$$\boldsymbol{\Theta}^{(t+1)} = \mathbb{Z}^{(t)} \times_{(m+1)} \left\{ (\mathbb{X} \mathbb{X}^{\intercal})^{-1} \mathbb{X} \right\},$$

where $\boldsymbol{Z}_i^{(t)} = [\![\boldsymbol{Y}_i; (\boldsymbol{\Gamma}_1^{(t+1)})^\intercal, \ldots, (\boldsymbol{\Gamma}_m^{(t+1)})^\intercal]\!] \in \mathbb{R}^{u_1 \times \ldots \times u_m}$, and $\mathbb{Z}^{(t)} \in \mathbb{R}^{u_1 \times \ldots \times u_m \times n}$ is the array stacking $\boldsymbol{Z}_1^{(t)}$ to $\boldsymbol{Z}_n^{(t)}$. It is noteworthy that the dimension of the tensor response in this step is reduced from $\prod_{k=1}^m r_k$ of $\boldsymbol{Y}$ to $\prod_{k=1}^m u_k$ of $\boldsymbol{Z}$. Next we estimate $\boldsymbol{\Omega}_k$, again using the iterative approach of Dutilleul (1999) and Manceur and Dutilleul (2013).

$$\begin{aligned}\boldsymbol{\Omega}_k^{(t+1)} &= \frac{1}{n\prod_{j\neq k} r_j} \sum_{i=1}^n \boldsymbol{s}_{i(k)}^{(t)} \Big\{ (\boldsymbol{\Omega}_m^{(t+1)})^{-1} \otimes \ldots \otimes (\boldsymbol{\Omega}_{k+1}^{(t+1)})^{-1} \\ &\qquad \otimes (\boldsymbol{\Omega}_{k-1}^{(t+1)})^{-1} \otimes \ldots \otimes (\boldsymbol{\Omega}_1^{(t+1)})^{-1} \Big\} \boldsymbol{s}_{i(k)}^\intercal,\end{aligned}$$

where $\boldsymbol{s}_i^{(t)} = \boldsymbol{Z}_i^{(t)} - \boldsymbol{\Theta}^{(t+1)} \times_{(m+1)} \boldsymbol{X}_i$ is the residual from the regression of $\boldsymbol{Z}^{(t)}$ on $\boldsymbol{X}$, and $\boldsymbol{s}_{i(k)}^{(t)}$ is its mode-k matricization. We remark that the above estimation of $\boldsymbol{\Omega}_k$ is parallel to the iterative updating of $\boldsymbol{\Sigma}_k^{(0)}$ during the initialization. The only changes are to replace $\boldsymbol{e}_i$ with $\boldsymbol{s}_i^{(t)}$, to replace $\boldsymbol{\Sigma}_k^{(0)}$ with $\boldsymbol{\Omega}_k^{(t+1)}$, and to replace the starting of iteration $\boldsymbol{I}_{r_k}$ with $\boldsymbol{\Omega}_k^{(t)}$. Next we estimate $\boldsymbol{\Omega}_{0k}$ using the formula,

$$\begin{aligned}\boldsymbol{\Omega}_{0k}^{(t+1)} &= \frac{1}{n\prod_{j\neq k} r_j} \sum_{i=1}^n (\boldsymbol{\Gamma}_{0k}^{(t+1)})^\intercal \boldsymbol{Y}_{i(k)} \Big\{ (\boldsymbol{\Sigma}_m^{(t)})^{-1} \otimes \ldots \otimes (\boldsymbol{\Sigma}_{k+1}^{(t)})^{-1} \\ &\qquad \otimes (\boldsymbol{\Sigma}_{k-1}^{(t)})^{-1} \otimes \ldots \otimes (\boldsymbol{\Sigma}_1^{(t)})^{-1} \Big\} \boldsymbol{Y}_{i(k)}^\intercal \boldsymbol{\Gamma}_{0k}^{(t+1)},\end{aligned}$$

where $\boldsymbol{\Gamma}_{0k}^{(t+1)} \in \mathbb{R}^{r_k \times (r_k - u_k)}$ is the orthogonal completion of $\boldsymbol{\Gamma}_k^{(t+1)} \in \mathbb{R}^{r_k \times u_k}$ such that $(\boldsymbol{\Gamma}_k^{(t+1)}, \boldsymbol{\Gamma}_{0k}^{(t+1)})$ is an orthogonal basis of $\mathbb{R}^{r_k}$. We also note that, unlike $\{\boldsymbol{\Omega}_k\}_{k=1}^m$, the estimation of $\{\boldsymbol{\Omega}_{0k}\}_{k=1}^m$ requires no iteration, since it is only based on the current estimator $\boldsymbol{\Sigma}^{(t)}$ and $\boldsymbol{\Gamma}_k^{(t+1)}$.

Finally, we update $\boldsymbol{B}$ through its parameterization $\boldsymbol{B} = [\![\boldsymbol{\Theta}; \boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_m, \boldsymbol{I}_p]\!]$ in Proposition 2. We are able to obtain the explicit formulae for such an update, i.e.,

$$\boldsymbol{B}^{(t+1)} = \mathbb{Y} \times_1 \boldsymbol{P}_{\boldsymbol{\Gamma}_1}^{(t+1)} \times_2 \ldots \times_m \boldsymbol{P}_{\boldsymbol{\Gamma}_m}^{(t+1)} \times_{(m+1)} \left\{ (\mathbb{X}\mathbb{X}^\intercal)^{-1} \mathbb{X} \right\} = [\![\widehat{\boldsymbol{B}}_{\text{OLS}}; \boldsymbol{P}_{\boldsymbol{\Gamma}_1}^{(t+1)}, \ldots, \boldsymbol{P}_{\boldsymbol{\Gamma}_m}^{(t+1)}, \boldsymbol{I}_p]\!],$$

where $\mathbb{X}$ and $\mathbb{Y}$ are defined in $\widehat{\boldsymbol{B}}_{\text{OLS}}$ in (7). Then we update the covariance $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma}_k^{(t+1)} = \boldsymbol{\Gamma}_k^{(t+1)} \boldsymbol{\Omega}_k^{(t+1)} (\boldsymbol{\Gamma}_k^{(t+1)})^\intercal + \boldsymbol{\Gamma}_{0k}^{(t+1)} \boldsymbol{\Omega}_{0k}^{(t+1)} (\boldsymbol{\Gamma}_{0k}^{(t+1)})^\intercal, \text{ and } \boldsymbol{\Sigma}^{(t+1)} = \boldsymbol{\Sigma}_m^{(t+1)} \otimes \ldots \otimes \boldsymbol{\Sigma}_1^{(t+1)}.$$

4.2 One-step estimator

Algorithm 1 is iterative, and steps 2 to 4 are repeated until the objective function converges. Although our numerical experiences suggest that the estimate often does not

Algorithm 2 Moment-based algorithm for minimizing $f_k^{*(0)}(\boldsymbol{G}_k)$.

for $s = 0, \ldots, u_k - 1$ **do**

Set $\boldsymbol{G}_k^s = \boldsymbol{0}$ if $s = 0$ and $\boldsymbol{G}_k^s = (\boldsymbol{g}_{k1}, \ldots, \boldsymbol{g}_{ks})$ otherwise

Construct $\boldsymbol{G}_{0k}^s$ as an orthogonal basis complement to $\boldsymbol{G}_k^s$ in $\mathbb{R}^{r_k}$

Solve the objective function over $\boldsymbol{w} \in \mathbb{R}^{r-s}$ subject to $\boldsymbol{w}^\top \boldsymbol{w} = 1$:

$$\boldsymbol{w}_{k+1} = \arg\min_{\boldsymbol{w}} \log \left\{ \boldsymbol{w}^\top \left((\boldsymbol{G}_{0k}^s)^\top \boldsymbol{\Sigma}_k^{(0)} \boldsymbol{G}_{0k}^s \right) \boldsymbol{w} \right\} + \log \left\{ \boldsymbol{w}^\top \left((\boldsymbol{G}_{0k}^s)^\top \boldsymbol{N}_k^{(0)} \boldsymbol{G}_{0k}^s \right)^{-1} \boldsymbol{w} \right\}.$$

Set $\boldsymbol{g}_{k+1} = \boldsymbol{G}_{0k}^s \boldsymbol{w}_{k+1} \in \mathbb{R}^{r_k}$ and normalize to unit length

end for

vary significantly after only a few iterations, the computations involved can still be intensive. In this iterative procedure, we recognize that the major computational expense arises from Step 2 that estimates the envelope basis $\{\boldsymbol{\Gamma}_k^{(t+1)}\}_{k=1}^m$ by optimizing the objective functions $f_k^{(t)}(\boldsymbol{G}_k)$ in (10) over the Grassmann manifolds. This is partly because $f_k^{(t)}$, for $k = 1, \ldots, m$, depend on each others' minimizers, so that the step of estimation of the envelope basis requires iterations within itself. Moreover, the Grassmann optimization is non-convex and possesses multiple local minima, and as such the algorithm usually requires multiple starting values.

Next we propose an alternative approach, which adopts the same framework of Algorithm 1, but replaces Step 2 with an approximate solution by introducing a modified objective function than (10). We then restrain the new estimator to be non-iterative, in that it goes through the steps of Algorithm 1 *only once*. Specifically, the new objective function is of the form,

$$f_k^{*(t)}(\boldsymbol{G}_k) = \log |\boldsymbol{G}_k^\top \boldsymbol{\Sigma}_k^{(t)} \boldsymbol{G}_k| + \log |\boldsymbol{G}_k^\top (\boldsymbol{N}_k^{(t)})^{-1} \boldsymbol{G}_k|. \tag{11}$$

Comparing the two objective functions, the term $\boldsymbol{M}_k^{(t)}$ in (10) is replaced by $\boldsymbol{\Sigma}_k^{(t)}$ in (11). It is also interesting to note that, for the first round of iteration, the term $\boldsymbol{P}_{\boldsymbol{\Gamma}_j}^{(t)}$ would take the initial value $\boldsymbol{I}_{r_j}$, and as such the term $\boldsymbol{\delta}_i^{(t)}$ becomes the OLS residual $\boldsymbol{e}_i = \boldsymbol{Y}_i - \boldsymbol{B}^{(0)} \times_{(m+1)} \boldsymbol{X}_i$. Consequently, $\boldsymbol{M}_k^{(0)} = \boldsymbol{\Sigma}_k^{(0)}$, and thus $f_k^{*(0)}(\boldsymbol{G}_k) = f_k^{(0)}(\boldsymbol{G}_k)$. This new objective function (11) has some appealing features. First of all, estimation of the envelope basis $\boldsymbol{\Gamma}_k$ through $f_k^{*(t)}$ does not depend on the values of other envelope basis $\boldsymbol{\Gamma}_j$, $j \neq k$. Therefore, estimation of $\boldsymbol{\Gamma}_1$ to $\boldsymbol{\Gamma}_m$ becomes *separable*. This property

alone could increase computational efficiency substantially. Second, the optimization of $f_k^{*(t)}$ is through the 1D algorithm recently proposed by Cook and Zhang (2016), and is summarized in Algorithm 2. It is faster and more stable than the full Grassmannian optimization, and much less sensitive to the initial guess. Thanks to both the modified objective function and the non-iterative fashion of the optimization, the resulting one-step estimator is computationally much faster than the iterative estimator. Our numerical studies have found it has a competitive finite sample performance. Moreover, as we show in the next section, like its iterative counterpart, this one-step estimator remains a consistent estimator of the true parameters.

4.3 Dimension selection

Selecting envelope dimensions $\{u_k\}_{k=1}^m$ is both of practical importance and plays an integral role of our proposed method. We recognize that, the objective function (8), upon which the iterative estimator is based, is essentially a likelihood function when the error follows a normal distribution. As such, one can, in principle, select the envelope dimensions $\{u_k\}_{k=1}^m$ using standard likelihood-based methods, e.g., sequential likelihood-ratio tests, or information criteria such as Bayesian information criterion (BIC). As shown in Shao (1997); Yang (2005), BIC is consistent in that the probability of selecting the correct dimension approaches one as the sample size goes to infinity. On the other hand, we also note that, the computation of the standard BIC in our setup can be intensive, as one has to search all the combinations of $(u_1, \ldots, u_m)$, within $1 \leq u_k \leq r_k, k = 1, \ldots, m$. To reduce the computational cost, one may heuristically impose that $u_1 = \ldots = u_m$.

Alternatively, we propose a variant of BIC that selects the envelope dimensions $\{u_k\}_{k=1}^m$ *separately*. That is, for every $k = 1, \ldots, m$, we select u_k that minimizes,

$$\mathrm{BIC}_k(u_k) = -\frac{n}{2}\log|\boldsymbol{\Gamma}_k^\intercal \boldsymbol{\Sigma}_k^{(0)} \boldsymbol{\Gamma}_k| - \frac{n}{2}\log|\boldsymbol{\Gamma}_k^\intercal (\boldsymbol{N}_k^{(0)})^{-1} \boldsymbol{\Gamma}_k| + \log(n) p u_k. \tag{12}$$

We make a few remarks on why this criterion is a reasonable choice. First, the first two terms in (12) are $(-n/2)$ times the objective function (11) of our one-step estimator, with the initial estimators $\boldsymbol{\Sigma}_k^{(0)}$ and $\boldsymbol{N}_k^{(0)}$ plugged in. As such, they form a negative

log-pseudo-likelihood, and thus possess all the desirable properties, including having a unique minimizer on the Grassmann manifold, and being continuous with first and second order derivatives converging uniformly in probability to their population counterparts. We conjecture that, without requiring the error distribution being normal, as long as the initial estimators $\boldsymbol{\Sigma}_k^{(0)}$ and $\boldsymbol{N}_k^{(0)}$ are $\sqrt{n}$-consistent, each selected u_k is to converge to the true envelope dimension in probability as the sample size goes to infinity. Second, this separable BIC connects nicely with the one-step estimator, as well as the selection of simultaneous envelope dimensions of Cook and Zhang (2015b). In Cook and Zhang (2015b), envelope dimension reduction was proposed for both multivariate response and predictor vectors, where the envelope dimensions were selected separately for the response and predictor and were shown to work well empirically. Our separate BIC selection shares a similar spirit as Cook and Zhang (2015b). Last but not least, the computational cost involved in the separate BIC is much reduced, and the estimation also becomes more stable. Its effectiveness is demonstrated empirically in both simulations (Sections 6.3, 6.4) and real data analysis (Section 7).

5 Asymptotics

In this section, we study the asymptotic properties of the envelope based estimators. We investigate both the iterative estimator, denoted as $\widehat{\boldsymbol{B}}^{it}_{\text{ENV}}$, and the one-step estimator, denoted as $\widehat{\boldsymbol{B}}^{os}_{\text{ENV}}$, under two scenarios: the normality of the error distribution holds or does not hold. We comment that, following a common practice in the envelope literature, all asymptotic results are established under the known true envelope dimensions. We believe such results are interesting and help improve our understanding of the tensor envelope estimators in several ways. First, we show that the one-step estimator is consistent, and thus justifying this estimator. Second, we establish that the envelope estimator achieves a smaller asymptotic variance than OLS, and is asymptotically efficient.

5.1 Consistency

We first establish that, under fairly weak moment conditions, both the iterative estimator and the one-step estimator are $\sqrt{n}$-consistent, when the error term in the tensor response

linear model (5) does not necessarily follow a normal distribution. We note that the consistency is established in terms of the projection matrices, $\boldsymbol{P}^{it}_{\boldsymbol{\Gamma}_k}$ for the iterative estimator $\widehat{\boldsymbol{B}}^{it}_{\text{ENV}}$, and $\boldsymbol{P}^{os}_{\boldsymbol{\Gamma}_k}$ for the one-step $\widehat{\boldsymbol{B}}^{os}_{\text{ENV}}$, since a subspace can have infinitely many semi-orthogonal basis but only one unique projection matrix.

Before we state the consistency result, we point out that a critical assumption required is the $\sqrt{n}$-consistency of the initial estimator $\boldsymbol{\Sigma}^{(0)}_k$ used in Algorithms 1 and 2. We have used the covariance estimator of Manceur and Dutilleul (2013) as our initial estimator, so we first establish the $\sqrt{n}$-consistency of this estimator. This result has been conjectured in Manceur and Dutilleul (2013), but it had no proof in that paper.

Lemma 1. *Assuming $\boldsymbol{\varepsilon}_i$, $i = 1, \ldots, n$, in model (5) are i.i.d. with a tensor normal distribution, i.e.* $\text{vec}(\boldsymbol{\varepsilon}_i)$ *follows a normal distribution with covariance $\boldsymbol{\Sigma}_m \otimes \cdots \otimes \boldsymbol{\Sigma}_1 > 0$, then the initial estimator $\boldsymbol{\Sigma}^{(0)}_k$ in (9) is $\sqrt{n}$-consistent for $\boldsymbol{\Sigma}_k$, $k = 1, \ldots, m$.*

Next we establish the consistency of our enveloped based estimators.

Theorem 1. *Assuming* $\text{vec}(\boldsymbol{\varepsilon}_i)$, $i = 1, \ldots, n$, *in model (5) are i.i.d. with finite fourth moments and the initial covariance estimator $\boldsymbol{\Sigma}^{(0)}_k$ is $\sqrt{n}$-consistent, $k = 1, \ldots, m$, then the projection $\boldsymbol{P}^{it}_{\boldsymbol{\Gamma}_k}$ and $\boldsymbol{P}^{os}_{\boldsymbol{\Gamma}_k}$ are both $\sqrt{n}$-consistent estimators for the projection onto the envelope $\mathcal{E}_{\boldsymbol{\Sigma}_k}\left(\boldsymbol{B}_{(k)}\right)$. The corresponding estimators $\widehat{\boldsymbol{B}}^{it}_{ENV}$ and $\widehat{\boldsymbol{B}}^{os}_{ENV}$ both converge at rate-$\sqrt{n}$ to the true tensor coefficient $\boldsymbol{B}_{TRUE}$.*

We also briefly comment that, one does not have to use Manceur and Dutilleul (2013) as an initial estimator of $\boldsymbol{\Sigma}^{(0)}_k$. Any $\sqrt{n}$-consistent covariance estimator can serve to initialize Algorithms 1 and 2 to obtain the same asymptotic consistency property.

5.2 Asymptotic normality

We next establish the asymptotic normality of the iterative estimator $\widehat{\boldsymbol{B}}^{it}_{\text{ENV}}$ when the error term $\text{vec}(\boldsymbol{\varepsilon})$ follows a normal distribution. Since only the iterative estimator is to be examined, we abbreviate its notation simply as $\widehat{\boldsymbol{B}}_{\text{ENV}}$, and the corresponding projection $\boldsymbol{P}^{it}_{\boldsymbol{\Gamma}_k}$ as $\widehat{\boldsymbol{P}}_k$. Here the iterative estimator is not guaranteed to be necessarily the maximum likelihood estimator (MLE), due to the existence of multiple local minima. However, it is asymptotically equivalent to the MLE. This is because, under the tensor

normal distribution, the initialization of Algorithm 1 is built upon $\sqrt{n}$-consistent estimators, while each parameter in Algorithm 1 is iteratively obtained along the partial derivative of the log-likelihood. From the classical theory of point estimation, we know that one Newton-Raphson step from the starting value provides an estimator that is asymptotically equivalent to the MLE even in the presence of multiple local minima (Lehmann and Casella, 1998, p. 454).

Theorem 2. *Assuming* $\text{vec}(\boldsymbol{\varepsilon}_i)$, $i = 1, \ldots, n$, *in model* (5) *are i.i.d. with a normal distribution, then* $\widehat{\boldsymbol{P}}_k$ *is* $\sqrt{n}$*-consistent for the projection onto the envelope* $\mathcal{E}_{\boldsymbol{\Sigma}_k}\left(\boldsymbol{B}_{(k)}\right)$, *for* $k = 1, \ldots, m$. *Also, the envelope estimator* $\widehat{\boldsymbol{B}}_{ENV}$ *is asymptotically distributed as* $\sqrt{n}\text{vec}(\widehat{\boldsymbol{B}}_{ENV} - \boldsymbol{B}_{TRUE}) \to N(0, \boldsymbol{U}_{ENV})$, *while the OLS estimator* $\widehat{\boldsymbol{B}}_{OLS}$ *satisfies that* $\sqrt{n}\text{vec}(\widehat{\boldsymbol{B}}_{OLS} - \boldsymbol{B}_{TRUE}) \to N(0, \boldsymbol{U}_{OLS})$, *and* $\boldsymbol{U}_{ENV} \leq \boldsymbol{U}_{OLS}$.

In addition to the established asymptotic normality, an important finding of Theorem 2 is that the asymptotic variance of the envelope estimator $\widehat{\boldsymbol{B}}_{\text{ENV}}$ is no greater than that of the OLS estimator $\widehat{\boldsymbol{B}}_{\text{OLS}}$. Therefore, $\widehat{\boldsymbol{B}}_{\text{ENV}}$ is asymptotically more efficient than $\widehat{\boldsymbol{B}}_{\text{OLS}}$. One can conveniently obtain the asymptotic covariance of $\text{vec}(\widehat{\boldsymbol{B}}_{\text{OLS}})$,

$$\boldsymbol{U}_{\text{OLS}} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_m \otimes \cdots \otimes \boldsymbol{\Sigma}_1,$$

where $\boldsymbol{\Sigma}_{\mathbf{X}} = \text{cov}(\boldsymbol{X})$. Next we give two additional results to gain more insight of $\boldsymbol{U}_{\text{ENV}}$. One assumes that the envelope basis is known a priori, and the other obtains the asymptotic variance of the estimator for both $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ jointly.

We first assume the envelope basis is known, and denote the corresponding envelope estimator of $\boldsymbol{B}$ as $\widehat{\boldsymbol{B}}_{\boldsymbol{\Gamma}}$. We then compare its asymptotic variance with that of $\widehat{\boldsymbol{B}}_{\text{OLS}}$.

Theorem 3. *Under the same conditions as in Theorem 1,* $\widehat{\boldsymbol{B}}_{\boldsymbol{\Gamma}}$ *is* $\sqrt{n}$*-consistent and asymptotically normal. The asymptotic covariance of* $\text{vec}(\widehat{\boldsymbol{B}}_{\boldsymbol{\Gamma}})$ *is*

$$\boldsymbol{U}_{\boldsymbol{\Gamma}} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{P}_{\boldsymbol{\Gamma}_m} \boldsymbol{\Sigma}_m \boldsymbol{P}_{\boldsymbol{\Gamma}_m} \otimes \cdots \otimes \boldsymbol{P}_{\boldsymbol{\Gamma}_1} \boldsymbol{\Sigma}_1 \boldsymbol{P}_{\boldsymbol{\Gamma}_1} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Gamma}_m \boldsymbol{\Omega}_m \boldsymbol{\Gamma}_m^{\mathsf{T}} \otimes \cdots \otimes \boldsymbol{\Gamma}_1 \boldsymbol{\Omega}_1 \boldsymbol{\Gamma}_1^{\mathsf{T}}.$$

Recall the decomposition $\boldsymbol{\Sigma}_k = \boldsymbol{\Gamma}_k \boldsymbol{\Omega}_k \boldsymbol{\Gamma}_k^{\mathsf{T}} + \boldsymbol{\Gamma}_{0k} \boldsymbol{\Omega}_{0k} \boldsymbol{\Gamma}_{0k}^{\mathsf{T}}$, $k = 1, \ldots, m$, in Proposition 2. Then it is straightforward to see that $\boldsymbol{U}_{\boldsymbol{\Gamma}} \leq \boldsymbol{U}_{\text{OLS}}$, and the more dominating the immaterial variation $\boldsymbol{\Gamma}_{0k} \boldsymbol{\Omega}_{0k} \boldsymbol{\Gamma}_{0k}^{\mathsf{T}}$ compared to the material variation $\boldsymbol{\Gamma}_k \boldsymbol{\Omega}_k \boldsymbol{\Gamma}_k^{\mathsf{T}}$, the bigger the

difference is between $\boldsymbol{U}_{\boldsymbol{\Gamma}}$ and $\boldsymbol{U}_{\text{OLS}}$. This result agrees with the pattern we have observed and reviewed in Section 2.2 for the vector response case, and shows the explicit gain of the envelope estimator in terms of estimation efficiency.

We next compare the asymptotic covariance of the envelope estimator and the OLS estimator when the envelope basis is unknown. Intuitively, there is an extra term in the covariance of the envelope estimator as the cost of estimating the unknown envelope basis. Toward that end, we introduce the following notations.

$$\boldsymbol{h} = \begin{pmatrix} \boldsymbol{h}_1 \\ \boldsymbol{h}_2 \end{pmatrix} = \begin{pmatrix} \text{vec}(\boldsymbol{B}) \\ \text{vech}(\boldsymbol{\Sigma}) \end{pmatrix}, \quad \boldsymbol{\phi} = \begin{pmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \\ \vdots \\ \boldsymbol{\phi}_{m+1} \end{pmatrix} = \begin{pmatrix} \text{vec}(\boldsymbol{B}) \\ \text{vech}(\boldsymbol{\Sigma}_1) \\ \vdots \\ \text{vech}(\boldsymbol{\Sigma}_m) \end{pmatrix}, \quad \boldsymbol{\xi} = \begin{pmatrix} \boldsymbol{\xi}_1 \\ \vdots \\ \boldsymbol{\xi}_{3m+1} \end{pmatrix},$$

where the operator $\text{vech}(\cdot) : \mathbb{R}^{r\times r} \mapsto \mathbb{R}^{r(r+1)/2}$ stacks the unique entries of a symmetric matrix into a column vector, $\boldsymbol{\xi}_1 = \text{vec}(\boldsymbol{\Theta})$, $\{\boldsymbol{\xi}_j\}_{j=2}^{m+1} = \{\text{vec}(\boldsymbol{\Gamma}_k)\}_{k=1}^{m}$, $\{\boldsymbol{\xi}_j\}_{j=m+2}^{2m+1} = \{\text{vech}(\boldsymbol{\Omega}_k)\}_{k=1}^{m}$, and $\{\boldsymbol{\xi}_j\}_{j=2m+2}^{3m+1} = \{\text{vech}(\boldsymbol{\Omega}_{0k})\}_{k=1}^{m}$. It is interesting to note that the total number of free parameters is reduced from $\boldsymbol{h}$ to $\boldsymbol{\phi}$ by $\prod_{k=1}^{m} r_k(\prod_{k=1}^{m} r_k + 1)/2 - \sum_{k=1}^{m} r_k(r_k+1)/2$ because of the imposed separable Kronecker covariance structure, and is further reduced from $\boldsymbol{\phi}$ to $\boldsymbol{\xi}$ by $p(\prod_{k=1}^{m} r_k - \prod_{k=1}^{m} u_k)$. We also note that $\boldsymbol{h}$ is an estimable functions of $\boldsymbol{\phi}$ and $\boldsymbol{\xi}$, respectively, and thus we write $\boldsymbol{h} = \boldsymbol{h}(\boldsymbol{\phi}) = \boldsymbol{h}(\boldsymbol{\xi})$. Let $\boldsymbol{J}_{\boldsymbol{h}}$ denote the Fisher information matrix for $\boldsymbol{h}$, let $\boldsymbol{H} = \partial\boldsymbol{h}(\boldsymbol{\phi})/\partial\boldsymbol{\phi}$, and $\boldsymbol{K} = \partial\boldsymbol{h}(\boldsymbol{\xi})/\partial\boldsymbol{\xi}$. The explicit forms of $\boldsymbol{J}_{\boldsymbol{h}}, \boldsymbol{H}$ and $\boldsymbol{K}$ are given in the Supplementary Materials. Furthermore, let $\boldsymbol{h}_{\text{OLS}}$ denote the OLS estimator of $\boldsymbol{h}$, $\boldsymbol{h}_{\text{ENV}}$ be the envelope estimator, and $\boldsymbol{h}_{\text{TRUE}}$ be the true parameter. Then, we have the following result.

Theorem 4. *Under the same conditions as in Theorem 2, both* $\boldsymbol{h}_{OLS}$ *and* $\boldsymbol{h}_{ENV}$ *are* $\sqrt{n}$*-consistent and asymptotically normal, so that* $\sqrt{n}(\boldsymbol{h}_{OLS} - \boldsymbol{h}_{TRUE}) \to N(0, \boldsymbol{V}_{OLS})$*, where* $\boldsymbol{V}_{OLS} = \boldsymbol{H}(\boldsymbol{H}^\intercal \boldsymbol{J}_{\boldsymbol{h}} \boldsymbol{H})^\dagger \boldsymbol{H}^\intercal$*, and* $\sqrt{n}(\boldsymbol{h}_{ENV} - \boldsymbol{h}_{TRUE}) \to N(0, \boldsymbol{V}_{ENV})$*, where* $\boldsymbol{V}_{ENV} = \boldsymbol{K}(\boldsymbol{K}^\intercal \boldsymbol{J}_{\boldsymbol{h}} \boldsymbol{K})^\dagger \boldsymbol{K}^\intercal$*. Moreover,*

$$\boldsymbol{V}_{OLS} - \boldsymbol{V}_{ENV} = \boldsymbol{J}_{\boldsymbol{h}}^{-1/2} \left(\boldsymbol{P}_{\boldsymbol{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{H}} - \boldsymbol{P}_{\boldsymbol{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{K}} \right) \boldsymbol{J}_{\boldsymbol{h}}^{-1/2} = \boldsymbol{J}_{\boldsymbol{h}}^{-1/2} \boldsymbol{P}_{\boldsymbol{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{H}} \boldsymbol{Q}_{\boldsymbol{J}_{\boldsymbol{h}}^{1/2}\boldsymbol{K}} \boldsymbol{J}_{\boldsymbol{h}}^{-1/2} \geq 0.$$

Once again, the envelope estimator is asymptotically more efficient than the OLS estimator. On the other hand, the envelope estimator of $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ are asymptotically

correlated. As such, there is no explicit form for the asymptotic covariance of $\widehat{\boldsymbol{B}}_{\text{ENV}}$, except that it is the $p(\prod r_k) \times p(\prod r_k)$ upper-left block of $\boldsymbol{V}_{\text{ENV}}$, when the envelope basis is unknown. This is different from the vector response case.

In applications such as brain imaging analysis, it is often useful to produce a voxel-by-voxel p-value map, so one can visually identify subregions of brains that display distinctive patterns between disease and control groups. Given the results of Theorems 2 and 4, we can produce such a p-value map for our envelope based estimator $\widehat{\boldsymbol{B}}_{\text{ENV}}$. In principle, one can obtain its asymptotic covariance $\boldsymbol{U}_{\text{ENV}}$ by extracting the upper-left block of $\boldsymbol{V}_{\text{ENV}}$. In practice, however, we suggest to substitute $\boldsymbol{U}_{\text{ENV}}$ with $\boldsymbol{U}_{\text{OLS}}$, which is computationally much simpler, though it would lead to more conservative p-values. Once the p-values are obtained, one can further employ either simple thresholding or false discovery rate correction.

6 Simulations

In this section, we report simulations to study the finite-sample performance of our envelope based estimator. In particular, we compare with the dominating solution in the literature, the one-at-a-time OLS estimator (Section 6.1), investigate the effect of the immaterial variation magnitude (Section 6.2), study the performance of the proposed BIC criterion for envelope dimension selection (Section 6.3), and examine a model when the response is a three-way tensor (Section 6.4). We also study a case where the simulated data does *not* comply with the envelope structure and to compare with the tensor predictor regression of Zhou et al. (2013) (Section 6.5).

6.1 Comparison with OLS

We begin with a comparison with the alternative solution that dominates the literature, the OLS estimator. Specifically, we consider the model of the form (5),

$$\boldsymbol{Y}_i = \boldsymbol{B} X_i + \sigma \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n. \tag{13}$$

We set the sample size $n = 20$, a fairly small value, to mimic the common scenario of imaging studies with a small number of subjects. In this model, $\boldsymbol{Y}_i$ is a 64×64 matrix,

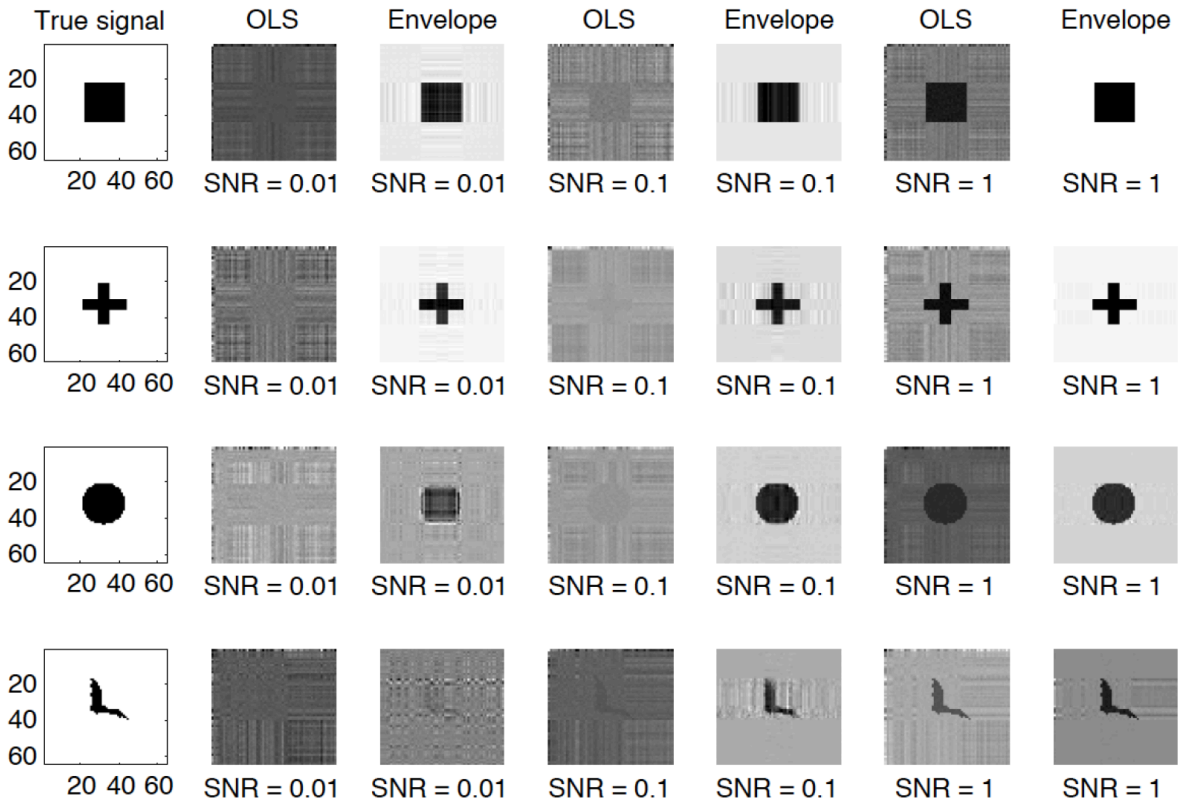


Figure 1: Comparison with OLS: The true and estimated regression coefficient tensors under various signal shapes and signal-to-noise ratios (SNR). The simulated data complies with the envelope structure.

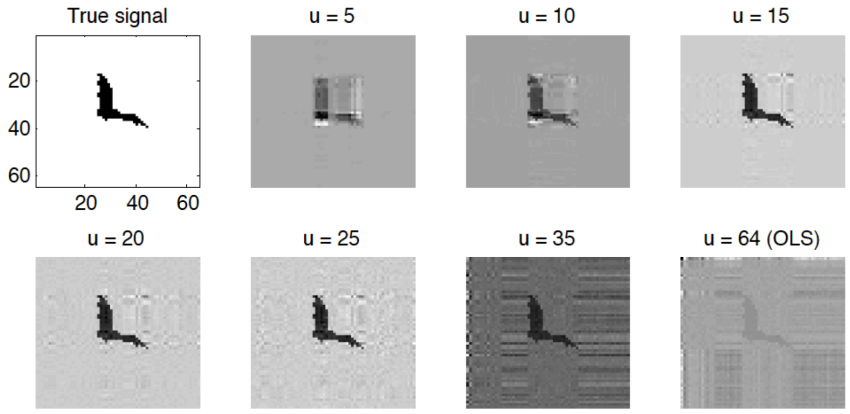
X_i is a scalar that only takes two values, 0 or 1, representing for instance the disease and control groups. The regression coefficient matrix $\boldsymbol{B} \in \mathbb{R}^{64\times 64}$ represents the mean change of the response between the two groups. Elements of $\boldsymbol{B}$ are either 0 or 1, and is plotted in the first column of Figure 1. We varied the shape of $\boldsymbol{B}$ among a square, a cross, a round disk and a bat shape. The constant σ in front of the error term $\boldsymbol{\varepsilon}$ was introduced to explicitly control the signal strength, and it took a value such that the signal-to-noise ratio (SNR) equals 0.01, 0.1, and 1, respectively. The error $\boldsymbol{\varepsilon}$ was generated from a matrix normal distribution, $\mathrm{vec}(\boldsymbol{\varepsilon}) \sim N(0, \boldsymbol{\Sigma}_2 \otimes \boldsymbol{\Sigma}_1)$. To make the model conform to the generalized sparsity principle (7), we generated $\boldsymbol{\Sigma}_k = \boldsymbol{\Gamma}_k\boldsymbol{\Omega}_k\boldsymbol{\Gamma}_k^{\mathrm{T}} + \boldsymbol{\Gamma}_{0k}\boldsymbol{\Omega}_{0k}\boldsymbol{\Gamma}_{0k}^{\mathrm{T}}$, $k = 1, 2$, then normalized it by its Frobenius norm. We set the working envelope dimension $u_1 = u_2$ equal to the numerical dimension of the true coefficient matrix $\boldsymbol{B}$, which is 1 for the square signal, 2 for the cross, 8 for the disk, and 14 for the bat-

shape. We eigen-decomposed the coefficient matrix as $\boldsymbol{B} = \boldsymbol{G}_1 \boldsymbol{D} \boldsymbol{G}_2^\intercal$. Then we generated $\boldsymbol{\Gamma}_k = \boldsymbol{G}_k \boldsymbol{O}_k$, for an orthogonal matrix $\boldsymbol{O}_k \in \mathbb{R}^{u_k \times u_k}$, whose elements were from a standard uniform distribution. This way, it is guaranteed that $\text{span}(\boldsymbol{B}) \subseteq \text{span}(\boldsymbol{\Gamma}_1)$ and $\text{span}(\boldsymbol{B}^\intercal) \subseteq \text{span}(\boldsymbol{\Gamma}_2)$. We then orthogonalized $\boldsymbol{\Gamma}_k$ and obtained $\boldsymbol{\Gamma}_{0k}$. The covariances $\boldsymbol{\Omega}_k \in \mathbb{R}^{u_k \times u_k}$ and $\boldsymbol{\Omega}_{0k} \in \mathbb{R}^{(r_k - u_k) \times (r_k - u_k)}$ were generated of the form $\boldsymbol{A}\boldsymbol{A}^\intercal$, where $\boldsymbol{A}$ is a square matrix with matching dimension and all its elements were from a standard uniform distribution. Figure 1 summarizes the estimated coefficient matrix $\boldsymbol{B}$, under varying image shapes and signal strengths. It is clearly seen that the envelope estimator $\widehat{\boldsymbol{B}}_{\text{ENV}}$ substantially outperforms the one-at-a-time OLS estimator $\widehat{\boldsymbol{B}}_{\text{OLS}}$. For instance, when the signal is weak (SNR is 0.01 or 0.1), the OLS estimator produced many small numbers in its estimator (as shown in dark color) and failed to identify any meaningful signal, whereas the envelope estimator produced a sound recovery. In addition, we emphasize that such a performance is achieved under a fairly small sample size ($n = 20$).

6.2 Envelope immaterial information

We next investigate the effect of the magnitude of the immaterial information on our proposed envelope based estimation. We generated $n = 100$ i.i.d. samples from model (13) (SNR=1) with the bat-shape signal, which is a natural shape and is relatively more complicated than the geometric shapes. We then varied the working envelope dimension $u_1 = u_2 = u$, with $u \in \{5, 10, 15, 20, 25, 35, 64\}$, while the numerical rank of the bat-shape signal equals 14 in this example. We also note that, if one sets the working envelope dimension u_k the same as the dimension of the tensor response r_k, then the envelope estimator degenerates to the OLS estimator. We introduced an additional scalar σ_0^2 in the covariance $\boldsymbol{\Sigma}_k = \boldsymbol{\Gamma}_k \boldsymbol{\Omega}_k \boldsymbol{\Gamma}_k^\intercal + \sigma_0^2 \, \boldsymbol{\Gamma}_{0k} \boldsymbol{\Omega}_{0k} \boldsymbol{\Gamma}_{0k}^\intercal$, where σ_0^2 controls the magnitude of the immaterial information. Figure 2(a) shows one snapshot of the results with $\sigma_0^2 = 1$ and Figure 2(b) shows the contrast snapshot of the results when $\sigma_0^2 = 1000$. Comparing the two figures, we first verify that, the more dominant of the immaterial information (i.e., the larger value of σ_0^2), the better performance of the envelope estimator. On the other hand, the OLS estimator ($u = 64$) continued to fail to identify any meaningful signal. This example thus shows the importance of recognizing the immaterial information to

(a) Immaterial-to-material variation ratio $\sigma_0^2 = 1$



(b) Immaterial-to-material variation ratio $\sigma_0^2 = 1000$

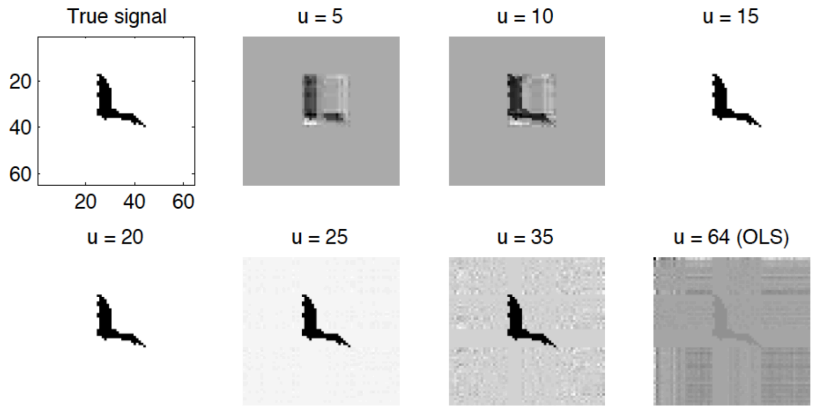


Figure 2: Effect of the strength of immaterial information on the envelope estimation.

improve the estimation. We also remark that, when the working envelope dimension is close to the true signal dimension ($u = 14$ in this example), the envelope estimator produced much refined recovery. When the working dimension is too small or too large than the truth, there is sign of under fitting or overfitting. Next we investigate the selection of envelope dimensions using the separable BIC criterion (12).

6.3 Envelope dimension selection

We first revisit the two-way matrix example in Section 6.1, then continue the investigation of dimension selection in the three-way tensor example in Section 6.4. Figure 3 reports the envelope based estimation under the estimated dimensions using (12) and

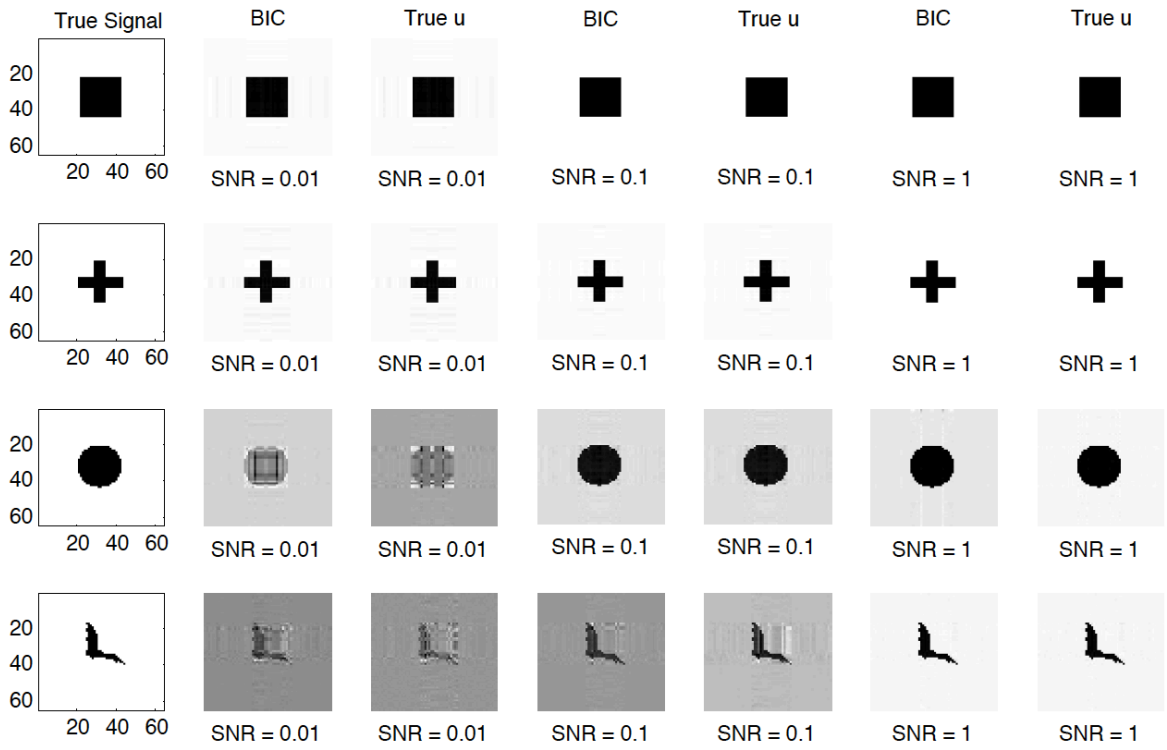


Figure 3: The envelope estimators with the true and estimated envelope dimensions.

under the true dimensions. The same simulated example of Section 6.1 was employed, while the sample size was set to $n = 100$. Our main observation is that, the recovered signals under BIC estimated dimensions are visually very similar to the ones under the true dimensions. This experiment, along with results in the next section, suggests that the BIC criterion (12) works reasonably well empirically.

6.4 Three-way tensor response

In the above simulations, we have primarily focused on the matrix-valued response (order-2 tensor) and a single predictor, since it enables a direct visualization of the resulting regression coefficient tensor (order-2). In this section, we consider a model where the response becomes an order-3 tensor with response dimensions $(r_1, r_2, r_3) = (20, 30, 40)$, and the predictor is a 5-dimensional vector. The resulting tensor coefficient is order-4. The rest of the setup was similar to that in Section 6.1. We simulated data with true envelope dimensions: $(u_1, u_2, u_3) = (2, 3, 4)$, $(8, 6, 4)$ and $(10, 10, 10)$, and sample

Dimension (u_1, u_2, u_3)		$(2,3,4)$		$(8,6,4)$		$(10,10,10)$	
Sample size n		100	400	100	400	100	400
OLS	Average	124.9478	30.1925	175.3339	40.3662	214.2976	52.5602
	(SE)	(4.2584)	(0.7922)	(3.7502)	(0.7610)	(3.4886)	(0.9056)
Env-True u	Average	0.0023	0.0007	0.0673	0.0232	1.9919	0.4849
	(SE)	(0.0006)	(0.0002)	(0.0024)	(0.0012)	(0.0511)	(0.0133)
Env-BIC	Average	0.0023	0.0006	0.0658	0.0220	1.9913	0.4894
	(SE)	(0.0006)	(0.0001)	(0.0021)	(0.0008)	(0.0512)	(0.0139)

Table 1: Average and SE (in parenthesis) of the estimation error $\|\boldsymbol{B} - \widehat{\boldsymbol{B}}\|^2$ for three estimators, OLS, the envelope estimator with the true envelope dimensions, and the envelope estimator with the estimated envelope dimensions based on BIC. The response is an order-3 tensor.

sizes: $n = 100$ and 400. We have chosen such settings to examine a variety of different true envelope dimensions. For each combination, we simulated 100 data replications, and report the average and the standard error of $\|\widehat{\boldsymbol{B}} - \boldsymbol{B}\|^2$ in Table 1. We compared three estimators: the OLS, the envelope estimator with true envelope dimensions, and the one with BIC estimated dimensions. A few observations are in order. First, both envelope estimators showed a clear improvement over the OLS in estimation accuracy. This agrees with the observed patterns in the previous simulations. Second, comparing the two envelope estimators, they essentially produced comparable results considering the standard errors, again providing evidence that the BIC selection works well.

6.5 Performance under model misspecification

Next we carry out a simulation using the tensor regression model of Zhou et al. (2013) that does *not* comply with the envelope structure, for two purposes. First, it allows us to study the empirical performance of the envelope method under model misspecification. Second, we compare with Zhou et al. (2013), which studies association between a scalar response and a *tensor predictor*. Even though both methods are motivated from neuroimaging analysis, and both involve a tensor variable in a regression analysis, the two

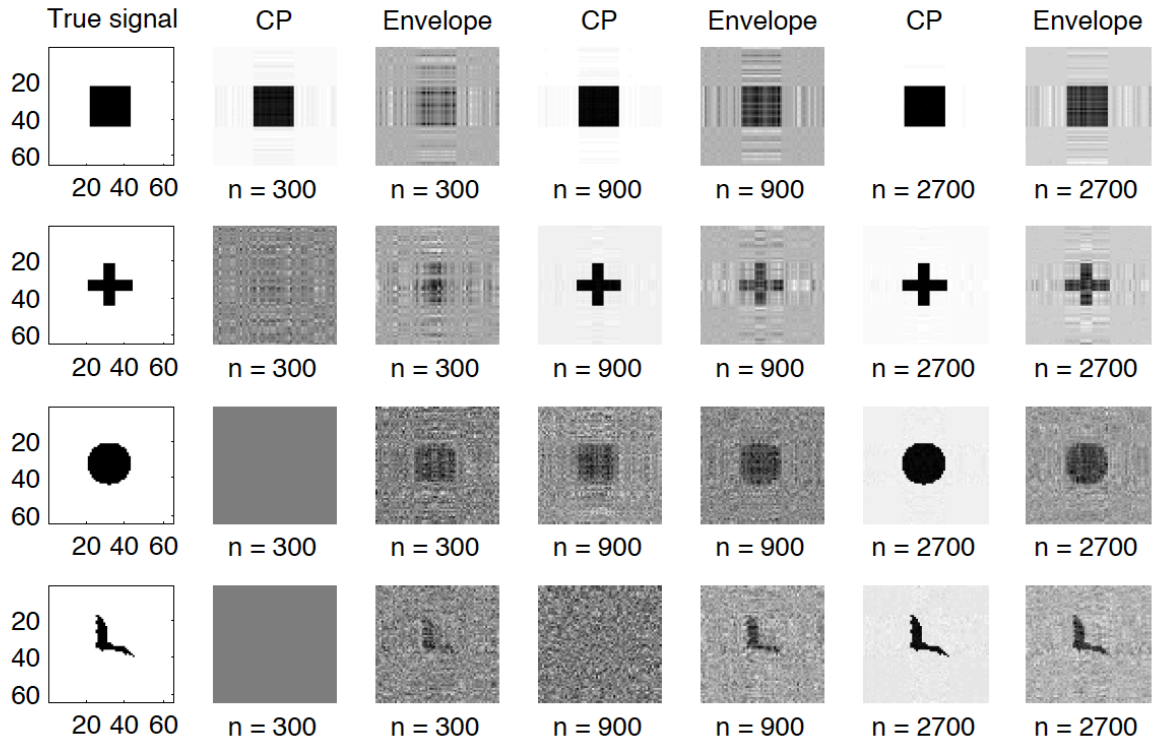


Figure 4: Comparison with tensor predictor regression: The true and estimated regression coefficient tensors under various signal shapes and sample sizes. The simulated data does *not* comply with the envelope structure.

clearly differ in the role the tensor plays in regression and the corresponding interpretation. Moreover, the techniques involved differ significantly, with our approach utilizing the generalized sparsity principle, whereas Zhou et al. (2013) employed a reduced-rank decomposition, the canonical decomposition or parallel factors (CP decomposition), of the coefficient tensor.

Specifically, we consider the model of Zhou et al. (2013),

$$Y_i = \langle \boldsymbol{B}, \boldsymbol{X}_i \rangle + \varepsilon_i, \quad i = 1, \ldots, n, \tag{14}$$

where Y_i is a scalar response, $\boldsymbol{X} \in \mathbb{R}^{64 \times 64}$ is an image whose elements were all drawn from a standard normal distribution, and the error ϵ is standard normal and independent of $\boldsymbol{X}$. The coefficient matrix $\boldsymbol{B} \in \mathbb{R}^{64 \times 64}$ was set in the same way as in Section 6.1. $\langle \boldsymbol{B}, \boldsymbol{X}_i \rangle = \langle \mathrm{vec}(\boldsymbol{B}), \mathrm{vec}(\boldsymbol{X}) \rangle$ is the tensor inner product. We examine three sample sizes, $n = 300, 900$ and 2700, respectively. Figure 4 summarizes the results. It is interesting

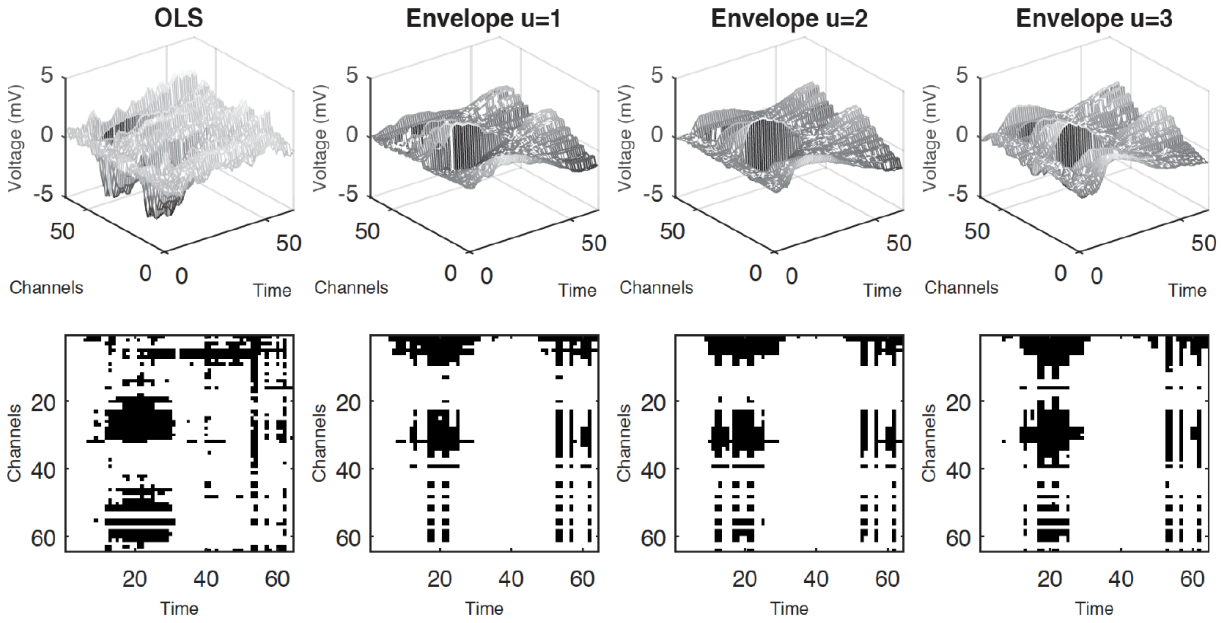


Figure 5: EEG data analysis: top panels are the estimated coefficient tensor using the OLS and envelope method with varying working dimensions; bottom panels are the corresponding p-value maps, thresholded at 0.05.

to observe that the envelope based estimator outperforms the CP estimator of Zhou et al. (2013) when the sample size is small ($n = 300$) to moderate ($n = 900$), and underperforms only when the sample size is fairly large ($n = 2700$), while still producing a reasonable signal recovery. It is also noteworthy that the data was generated from model (14), based on which that the CP estimator was built. As a result, it actually favors the CP estimator, but not the envelope estimator, since the generalized sparsity principle (7) is not guaranteed in this setup. Therefore this experiment shows the promise of our envelope estimator when the assumed envelope structure does not hold in the data.

7 Real Data Analysis

7.1 EEG data analysis

We first analyzed 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 Zhang et al. (1995) and Li et al. (2010). To facilitate the analysis, we downsized the data along the time domain by averaging four consecutive time points, yielding a 64×64 matrix response. We report the OLS estimator and the envelope based estimators in Figure 5, along with the corresponding p-value maps thresholded at 0.05. The BIC selected the envelope dimensions $u_1 = 1$ and $u_2 = 1$, while the results for $u_1 = u_2 = 2$ and 3 are shown in the plot too. We first note that, the envelope estimators under similar envelope dimensions produced a consistent pattern of findings. It is also interesting to observe that, the envelope estimators identified the channels between about 0 to 15, and between 25 to 40, at time range from 60 to 120, and from 200 to 240, in the original time scale, that are mostly relevant to distinguish the alcoholic group from the control. By contrast, the OLS estimator is much more variable, and the revealed signal regions are less clear.

7.2 ADHD data analysis

We next analyzed a magnetic resonance imaging (MRI) data for a study of attention deficit hyperactivity disorder (ADHD). The data was produced by the ADHD-200 Sample Initiative, then preprocessed by the Neuro Bereau and made available at `http://neurobureau.projects.nitrc.org/ADHD200/Data.html`. It consists of 776 subjects, among whom 285 are combined ADHD subjects and 491 are normal controls. We removed 47 subjects due to the missing observations or poor image quality, then downsized the MRI images from $256 \times 198 \times 256$ to $30 \times 36 \times 30$, which is to serve as our 3-way tensor response. This downsizing step is to facilitate envelope estimation, and results in a reduced resolution in images. This is a compromise given the limited sample size and large number of unknown parameters. The predictors include the group indicator (1 for ADHD and 0 for control), the subject's age and gender. Figure 6 show the OLS and envelope based estimators, and Figure 7 shows the corresponding p-value maps. Here we did not correct for multiple testing, as it is a nontrivial task when the tests on image voxels are spatially correlated. The BIC selected the envelope dimensions

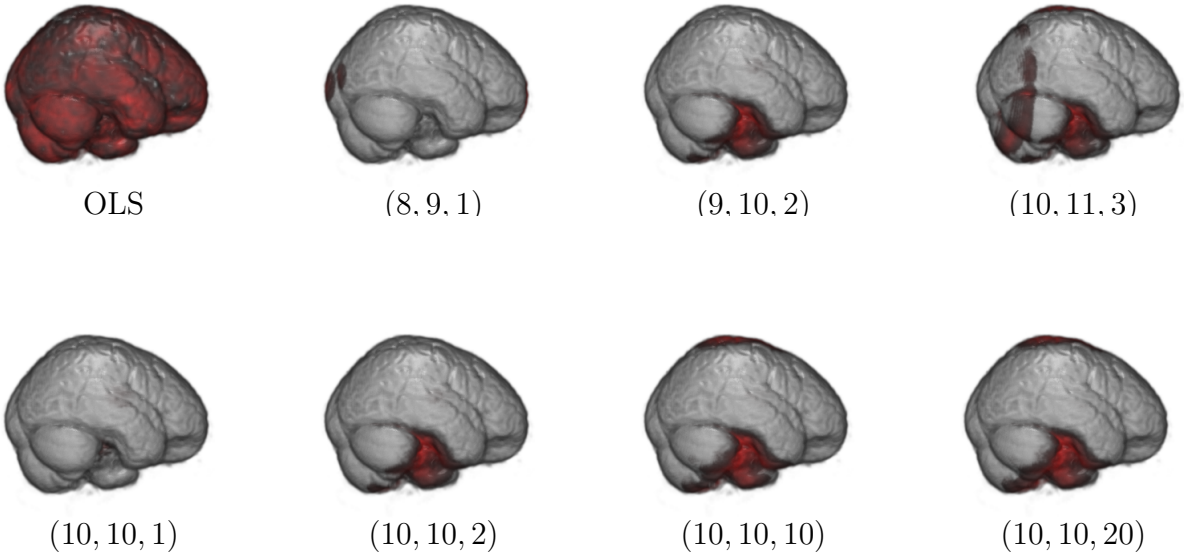


Figure 6: ADHD data analysis: Shown are the estimated coefficient tensor using the OLS and envelope method with varying working dimensions. The BIC selected envelope dimensions are $(9, 10, 2)$.

$(9, 10, 2)$, and for comparison, we also report the envelope estimators under a sequence of other working dimensions. We see from this example the importance of the choice of working dimensions in envelope estimation. Meanwhile, we note that the envelope estimators under the working dimensions that are close to the selected dimensions show a consistent pattern of relevant regions. Moreover, the OLS estimator reveals essentially no useful information, whereas the envelope estimators identify some regions of distinctive activity patterns between the ADHD and control subjects. In particular, the identified regions under $(9, 10, 2)$ correspond to superior temporal gyrus, and pyramid and uvula in cerebellum, and such findings are consistent with the literature (Bigler et al., 2007; Vaidya et al., 2014; Rubia et al., 2014; Cao et al., 2006; Rapin et al., 2014).

8 Discussion

In this article, we have proposed a parsimonious model for regression with a tensor response and a vector of predictors. Adopting a generalized sparsity principle, we have

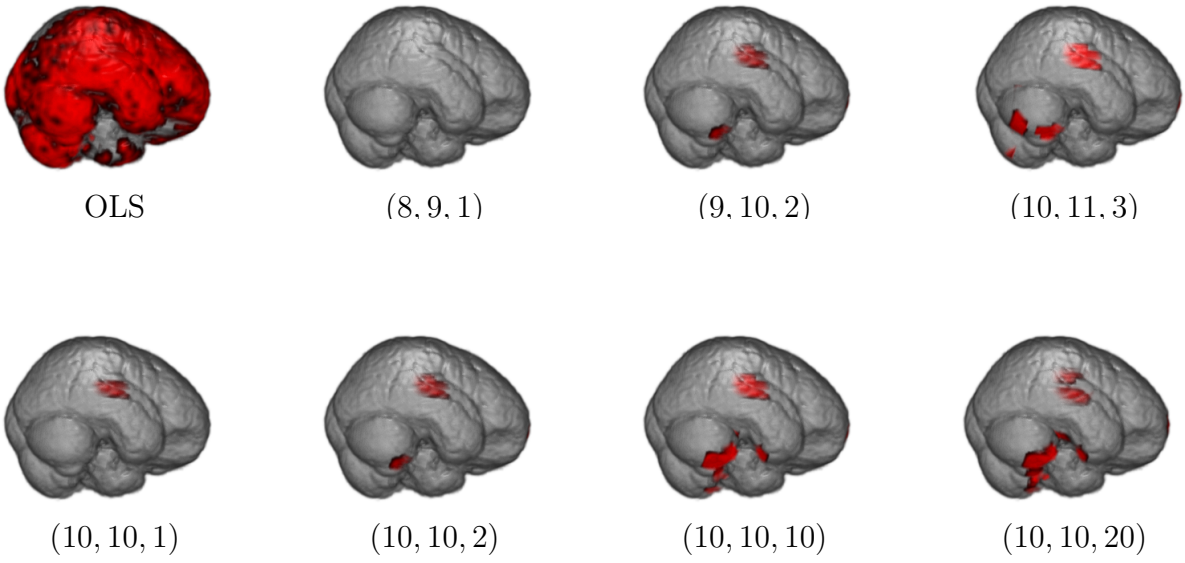


Figure 7: ADHD data analysis: Shown are the corresponding p-value map, thresholded at 0.05, using the OLS and envelope method with varying working dimensions. The BIC selected envelope dimensions are $(9, 10, 2)$.

developed an envelope estimator that can identify and focus on the material information of the tensor response. By doing so, the number of free parameters is effectively reduced, and the resulting estimator is asymptotically efficient. Both simulations and real data analysis have demonstrated effectiveness of the new estimator.

We make a few remarks regarding the envelope dimension selection. First, the selection of envelope dimensions reflects a bias-variance trade-off. When the estimated envelope dimensions are smaller than the truth, the corresponding envelope estimator is biased, whereas if the selected dimensions are larger than the truth, the resulting estimator remains consistent but can be more variable and lose some efficiency. Second, through our numerical analysis, we have frequently observed that the envelope estimator is not overly sensitive to the envelope dimensions as long as the values are close. In simulations, estimators based on the selected and true dimensions are visually similar. Third, theoretical properties of our BIC based selection are useful for further justification and understanding of this method. However, the task is far from trivial, and will

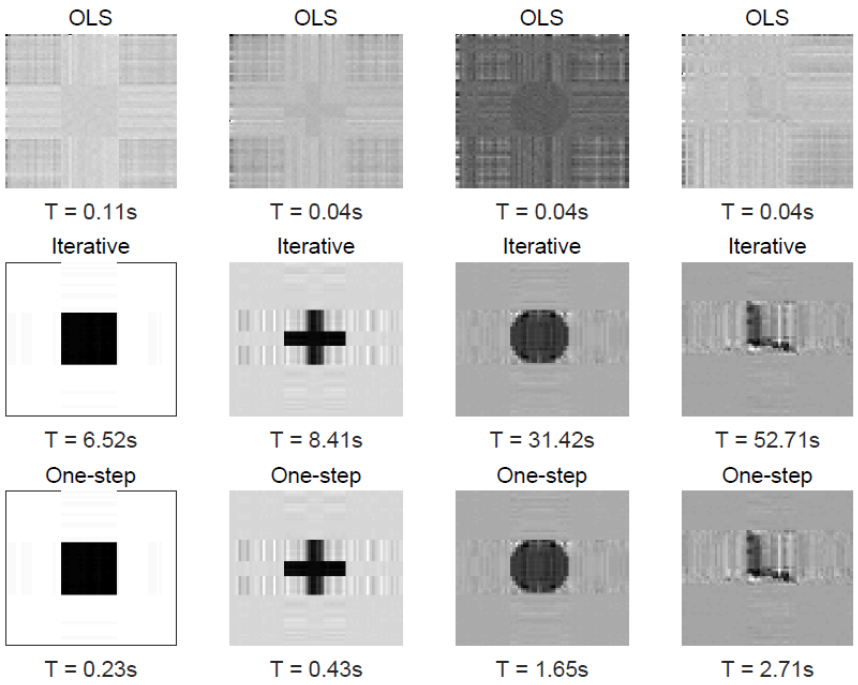


Figure 8: Computation time for the OLS, the iterative and the one-step envelope estimators. The simulation setup follows that in Section 6.1 with SNR$= 0.1$ and $n = 20$.

consist of our future research.

We also make some remarks on regularization. The core idea of our proposal is to recognize and focus the estimation based upon the relevant information in the tensor response, and the method can be viewed as a regularized estimation. Sparsity here is defined in a general sense and is achieved through the envelope method. This is different from the commonly used strategy in sparse modeling, which induces sparsity mostly through penalty functions. On the other hand, our envelope based estimation can be naturally *coupled* with penalty functions to attain further regularization. This line of work is currently under investigation.

Finally, we record and report the computation time of OLS, the iterative envelope estimator of Algorithm 1, and the one-step estimator using Algorithm 2. For the simulation example of Section 6.1, Figure 8 shows both the estimation results and the computation time for one data replication. All computation was done on a Windows 7 laptop computer with Intel(R) Core(TM) i5-5300U CPU@2.30GHz processor, 8.00 GB

installed memory (RAM), 64-bit Operating System. The running time of the iterative estimation was 6.52, 8.41, 31.42, and 52.71 seconds, respectively, for the four signal shapes. By comparison, the running time of the one-step estimation was 0.23, 0.43, 1.65 and 2.71 seconds, and that for OLS was 0.11, 0.04, 0.04, and 0.04 seconds, respectively. As we see, the one-step estimation is much faster than the iterative estimation, but produces similar signal recovery. Meanwhile, the running time of the one-step estimation is slightly longer than, but comparable to that of OLS, whereas the one-step estimator produces a much more accurate recovery of the coefficient signal than OLS. Given its appealing features in both computation and asymptotic consistency, we recommend the one-step estimation in practice.

References

Bigler, E. D., Mortensen, S., Neeley, E. S., Ozonoff, S., Krasny, L., Johnson, M., Lu, J., Provencal, S. L., McMahon, W., and Lainhart, J. E. (2007). Superior temporal gyrus, language function, and autism. *Developmental neuropsychology*, 31(2):217–238.

Cao, Q., Zang, Y., Sun, L., Sui, M., Long, X., Zou, Q., and Wang, Y. (2006). Abnormal neural activity in children with attention deficit hyperactivity disorder: a resting-state functional magnetic resonance imaging study. *Neuroreport*, 17(10):1033–1036.

Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107(500):1533–1545.

Chun, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B.*, 72(1):3–25.

Cook, R. D., Forzani, L., and Zhang, X. (2015). Envelopes and reduced-rank regression. *Biometrika*, 102(2):439–456.

Cook, R. D., Helland, I. S., and Su, Z. (2013). Envelopes and partial least squares regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 75(5):851–877.

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

Cook, R. D. and Zhang, X. (2015a). Foundations for envelope models and methods. *Journal of the American Statistical Association*, 109(510):599–611.

Cook, R. D. and Zhang, X. (2015b). Simultaneous envelopes for multivariate linear regression. *Technometrics*, 57(1):11–25.

Cook, R. D. and Zhang, X. (2016). Algorithms for envelope estimation. *Journal of Computational and Graphical Statistics*, In Press.

Du, A., Schuff, N. Amend, D., Laakso, M., Hsu, Y., Jagust, W., Yaffe, K., Kramer, J., Reed, B., Norman, D., Chui, H., and Weiner, W. (2001). Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and alzheimer's disease. *Journal of Neurology, Neurosurgery and Psychiatry*, 71:441–447.

Dutilleul, P. (1999). The mle algorithm for the matrix normal distribution. *J. Statist. Comput. Simul.*, 64:105–123.

Fosdick, B. K. and Hoff, P. D. (2014). Separable factor analysis with applications to mortality data. *Ann. Appl. Stat.*, 8(1):120–147.

Friston, K., Ashburner, J., Kiebel, S., Nichols, T., and Penny, W., editors (2007). *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Academic Press.

Goldsmith, J., Huang, L., and Crainiceanu, C. (2014). Smooth scalar-on-image regression via spatial bayesian variable selection. *Journal of Computational and Graphical Statistics*, 23:46–64.

Helland, I. S. (1990). Partial least squares regression and statistical models. *Scand. J. Statist.*, 17(2):97–114.

Helland, I. S. (1992). Maximum likelihood regression on relevant components. *J. Roy. Statist. Soc. Ser. B*, 54(2):637–647.

Hoff, P. D. (2011). Separable covariance arrays via the Tucker product, with applications to multivariate relational data. *Bayesian Anal.*, 6(2):179–196.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *J. Multivariate Anal.*, 5:248–264.

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

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*, volume 31. Springer Science & Business Media.

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

Li, Y., Gilmore, J. H., Shen, D., Styner, M., Lin, W., and Zhu, H. (2013). Multiscale adaptive generalized estimating equations for longitudinal neuroimaging data. *NeuroImage*, 72(0):91 – 105.

Li, Y., Zhu, H., Shen, D., Lin, W., Gilmore, J. H., and Ibrahim, J. G. (2011). Multiscale adaptive regression models for neuroimaging data. *Journal of the Royal Statistical Society: Series B*, 73:559–578.

Manceur, A. M. and 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.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R., and Wang, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics*, 4(1):53.

Rapin, L., Poissant, H., and Mendrek, A. (2014). Atypical activations of fronto-cerebellar regions during forethought in parents of children with ADHD. *Journal of attention disorders*, page 1087054714524983.

Reinsel, G. C. and Velu, R. P. (1998). *Multivariate reduced-rank regression*, volume 136 of *Lecture Notes in Statistics*. Springer-Verlag, New York. Theory and applications.

Reiss, P. and Ogden, R. (2010). Functional generalized linear models with images as predictors. *Biometrics*, 66:61–69.

Rubia, K., Smith, A. B., Brammer, M. J., Toone, B., and Taylor, E. (2014). Abnormal brain activation during inhibition and error detection in medication-naive adolescents with ADHD. *American Journal of Psychiatry*.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica sinica*, 7(2):221–242.

Similä, T. and Tikka, J. (2007). Input selection and shrinkage in multiresponse linear regression. *Computational Statistics & Data Analysis*, 52(1):406–422.

Skup, M., Zhu, H., and Zhang, H. (2012). Multiscale adaptive marginal analysis of longitudinal neuroimaging data with time-varying covariates. *Biometrics*, 68(4):1083–1092.

Su, Z. and Cook, D. (2012). Inner envelopes: efficient estimation in multivariate linear regression. *Biometrika*, 99(3):687–702.

Su, Z. and Cook, R. D. (2011). Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika*, 98(1):133–146.

Su, Z. and Cook, R. D. (2013). Estimation of multivariate means with heteroscedastic errors using envelope models. *Statist. Sinica*, 23(1):213–230.

Turlach, B. A., Venables, W. N., and Wright, S. J. (2005). Simultaneous variable selection. *Technometrics*, 47(3):349–363.

Vaidya, C. J., Bunge, S. A., Dudukovic, N. M., Zalecki, C. A., Elliott, G. R., and Gabrieli, J. D. (2014). Altered neural substrates of cognitive control in childhood ADHD: evidence from functional magnetic resonance imaging. *American Journal of Psychiatry*.

Wang, X., Nan, B., Zhu, J., and Koeppe, R. (2014). Regularized 3D functional regression for brain image data via haar wavelets. *The Annals of Applied Statistics*, 8:1045–1064.

Yang, Y. (2005). Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950.

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):329–346.

Zhang, X., Begleiter, H., Porjesz, B., Wang, W., and Litke, A. (1995). Event related potentials during object recognition tasks. *Brain Res. Bull.*, 38:531–538.

Zhou, H. and Li, L. (2014). Regularized matrix regression. *Journal of the Royal Statistical Society. Series B*, 76:463–483.

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

Zhou, J. and He, X. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. *Ann. Statist.*, 36(4):1649–1668.