Tensor Regression with Applications in Neuroimaging Data Analysis

Hua Zhou, Lexin Li, and Hongtu Zhu

Abstract

Classical regression methods treat covariates as a vector and estimate a corresponding vector of regression coefficients. Modern applications in medical imaging generate covariates of more complex form such as multidimensional arrays (tensors). Traditional statistical and computational methods are proving insufficient for analysis of these high-throughput data due to their ultrahigh dimensionality as well as complex structure. In this article, we propose a new family of tensor regression models that efficiently exploit the special structure of tensor covariates. Under this framework, ultrahigh dimensionality is reduced to a manageable level, resulting in efficient estimation and prediction. A fast and highly scalable estimation algorithm is proposed for maximum likelihood estimation and its associated asymptotic properties are studied. Effectiveness of the new methods is demonstrated on both synthetic and real MRI imaging data.

Key Words: Brain imaging; dimension reduction; generalized linear model (GLM); magnetic resonance imaging (MRI); multidimensional array; tensor regression.

[1] Hua Zhou is Assistant Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203 (Email: hua_zhou@ncsu.edu). Lexin Li is the corresponding author and Associate Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203 (Email: li@stat.ncsu.edu). Hongtu Zhu is Professor, Department of Biostatistics and Biomedical Research Imaging Center, University of North Carolina, Chapel Hill, NC 27599-7420 (E-mail: hzhu@bios.unc.edu).

1 Introduction

Understanding the inner workings of the human brains and their connection with neuropsychiatric and neurodegenerative disorders is one of the most intriguing scientific questions. Studies in neuroscience are greatly facilitated by a variety of neuroimaging technologies, including anatomical magnetic resonance imaging (MRI), functional magnetic resonance imaging (fMRI), electroencephalography (EEG), diffusion tensor imaging (DTI), and positron emission tomography (PET), among others. The sheer size and complexity of medical imaging data, however, pose unprecedented challenge to many classical statistical methods and have received increasing interest in recent years (Lindquist, 2008; Lazar, 2008; Martino et al., 2008; Friston, 2009; Ryali et al., 2010; Hinrichs et al., 2009; Kang et al., 2012).

In the literature, there have been roughly three categories of statistical methods for establishing association between brain images and clinical traits. The first is the voxel-based methods, which take each voxel as responses and clinical variables such as age and gender as predictors. They generate a statistical parametric map of test statistics or p-values across all voxels (Lazar, 2008; Worsley et al., 2004). A major drawback is that all voxels are treated as independent units and important spatially correlation is ignored (Li et al., 2011; Yue et al., 2010; Polzehl et al., 2010). The second type of solutions adopts the functional data analysis approach. Reiss and Ogden (2010) notably extended functional regression model to incorporate two-dimensional images as predictors. Generalizations to 3D and higher dimensional images, however, is far from trivial and requires substantial research. The third category employs a two-stage strategy. These methods first carry out a dimension reduction step, often by principal component analysis (PCA), and then fit a regression model based on the top principal components (Caffo et al., 2010). This strategy is intuitive and easy to implement. However, it is well known that PCA is an unsupervised dimension reduction technique and the extracted principal components can be irrelevant to the response.

In this article, we formulate a regression framework that treats clinical outcome as response, and images, in the form of *multi-dimensional array*, as covariates. Most classical regression methods take vectors as covariates. Naively turning an image array into a vector yields an unsatisfactory solution. For instance, typical anatomical MRI images of size 256-by-256-by-256 implicitly require $256^3 = 16,777,216$ regression parameters. Both computability and theoretical guarantee of the classical regression analysis are compromised by this ultra-high dimensionality. More seriously, vectorizing an array destroys the inherent spatial structure of the image that possesses wealth of information.

Exploiting the array structure in imaging data, our new regression method substantially reduces the dimensionality of imaging data, leading to efficient estimation and prediction. The method works for general array-valued covariates and/or any combination of them, and thus it is applicable to a variety of imaging modalities, e.g., EEG, MRI and fMRI. It is embedded in the generalized linear model (GLM) framework, so it works for both continuous and discrete responses. We develop a highly scalable algorithm for maximum likelihood estimation, as well as statistical inferential tools. Regularized tensor regression is also investigated to identify regions of interest in brains that are relevant to a particular response. This *region selection* problem corresponds to *variable selection* in the usual vector-valued regression.

The contributions of this article are two-fold. First, from a brain imaging analysis point of view, our proposal timely responds to a number of growing needs of neuroimaging analysis. In the review article, Lindquist (2008) noted the increasing trend and demands of using brain images for disease diagnosis and prediction, for characterization of subjective human experience, and for understanding association between brain regions and cognitive outcomes. Our tensor regression framework offers a systematic solution to this family of problems. Moreover, the framework warrants potential solutions to address questions such as multi-modality imaging analysis, multi-phenotype analysis and imaging genetics (Friston, 2009; Casey et al., 2010), which largely remain as open challenges. Second, from a statistical methodology point of view, our proposal develops a general statistical framework for *regression with array covariates.* A large number of models and extensions, e.g., quasi-likelihood models (McCullagh and Nelder, 1983), are potential outcomes within this framework. It can also be viewed as a logic extension from the classical vector-valued covariate regression to functional covariate regression and then to array-valued covariate regression.

The rest of the article is organized as follows. Section 2 begins with a review of matrix/array properties, and then develops the tensor regression models. Section 3 presents an efficient algorithm for maximum likelihood estimation. Section 4 provides theoretical results such as identifiability, consistency, and asymptotic normality. Section 5 discusses regularization including region selection. Section 6 presents numerical results. Section 7 concludes with a discussion of future extensions. Technical proofs are delegated to the Appendix.

2 Model

2.1 Preliminaries

Multidimensional array, also called *tensor*, plays a central role in our approach and we start with a brief summary of notation and a few results for matrix/array operations. Extensive references can be found in the survey paper (Kolda and Bader, 2009). In this article we use the terms multidimensional array and tensor interchangeably.

Given two matrices $\boldsymbol{A} = [\boldsymbol{a}_1 \ldots \boldsymbol{a}_n] \in \mathbb{R}^{m \times n}$ and $\boldsymbol{B} = [\boldsymbol{b}_1 \ldots \boldsymbol{b}_q] \in \mathbb{R}^{p \times q}$, the *Kronecker product* is the *mp*-by-*nq* matrix $\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} \boldsymbol{a}_1 \otimes \boldsymbol{B} & \boldsymbol{a}_1 \otimes \boldsymbol{B} & \ldots & \boldsymbol{a}_n \otimes \boldsymbol{B} \end{bmatrix}$. If $\boldsymbol{A}$ and $\boldsymbol{B}$ have the same number of columns $n = q$, then the *Khatri-Rao* product (Rao and Mitra, 1971) is defined as the *mp*-by-*n* columnwise Kronecker product $\boldsymbol{A} \odot \boldsymbol{B} = \begin{bmatrix} \boldsymbol{a}_1 \otimes \boldsymbol{b}_1 & \boldsymbol{a}_2 \otimes \boldsymbol{b}_2 & \ldots & \boldsymbol{a}_n \otimes \boldsymbol{b}_n \end{bmatrix}$. If $n = q = 1$, then $\boldsymbol{A} \odot \boldsymbol{B} = \boldsymbol{A} \otimes \boldsymbol{B}$. Some useful operations transform a tensor into a matrix/vector. The vec$(\boldsymbol{B})$ *operator* stacks the entries of a D-dimensional tensor $\boldsymbol{B} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ into a column vector. Specifically, an entry $b_{i_1 \ldots i_D}$ maps to the j-th entry of vec $\boldsymbol{B}$, in which $j = 1 + \sum_{d=1}^{D} (i_d - 1) \prod_{d'=1}^{d-1} p_{d'}$. For instance, when $D = 2$, the matrix entry $x_{i_1 i_2}$ maps to position $j = 1 + i_1 - 1 + (i_2 - 1)p_1 = i_1 + (i_2 - 1)p_1$, which is consistent with the more familiar vec operation on a matrix. The *mode-d matricization*, $\boldsymbol{B}_{(d)}$, maps a tensor $\boldsymbol{B}$ into a $p_d \times \prod_{d' \neq d} p_{d'}$ matrix such that the $(i_1, \ldots, i_D)$ element of the array $\boldsymbol{B}$ maps to the (i_d, j) element of the matrix $\boldsymbol{B}_{(d)}$, where $j = 1 + \sum_{d' \neq d} (i_{d'} - 1) \prod_{d'' < d', d'' \neq d} p_{d''}$. With $d = 1$, we observe that vec $\boldsymbol{B}$ is the same as vectorizing the mode-1 matricization $\boldsymbol{B}_{(1)}$. The *mode-(d, d')*

matricization $\boldsymbol{B}_{(dd')} \in \mathbb{R}^{p_d p_{d'} \times \prod_{d'' \neq d,d'} p_{d''}}$ is defined in a similar fashion (Kolda, 2006). We also introduce an operator that turns vectors into an array. Specifically, an *outer product*, $\boldsymbol{b}_1 \circ \boldsymbol{b}_2 \circ \cdots \circ \boldsymbol{b}_D$, of D vectors $\boldsymbol{b}_d \in \mathbb{R}^{p_d}$ is a $p_1 \times \cdots \times p_D$ array with entries $(\boldsymbol{b}_1 \circ \boldsymbol{b}_2 \circ \cdots \circ \boldsymbol{b}_D)_{i_1 \cdots i_D} = \prod_{d=1}^{D} b_{di_d}$.

Tensor decomposition plays a central role in our proposed tensor regression in Section 2.3. An array $\boldsymbol{B} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ admits a *rank-R decomposition* if

$$\boldsymbol{B} = \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} \circ \cdots \circ \boldsymbol{\beta}_D^{(r)}, \tag{1}$$

where $\boldsymbol{\beta}_d^{(r)} \in \mathbb{R}^{p_d}, d = 1, \ldots, D, r = 1, \ldots, R$, are all column vectors, and $\boldsymbol{B}$ cannot be written as a sum of less than R outer products. For convenience, the decomposition is often represented by a shorthand, $\boldsymbol{B} = [\![\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D]\!]$, where $\boldsymbol{B}_d = [\boldsymbol{\beta}_d^{(1)}, \ldots, \boldsymbol{\beta}_d^{(R)}] \in \mathbb{R}^{p_d \times R}$ (Kolda, 2006; Kolda and Bader, 2009). The following well-known result relates the mode-d matricization and the vec operator of an array to its rank-R decomposition.

Lemma 1. *If a tensor $\boldsymbol{B} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ admits a rank-R decomposition (1), then*

$$\boldsymbol{B}_{(d)} = \boldsymbol{B}_d(\boldsymbol{B}_D \odot \cdots \odot \boldsymbol{B}_{d+1} \odot \boldsymbol{B}_{d-1} \odot \cdots \odot \boldsymbol{B}_1)^{\mathsf{T}} \;\; \textit{and} \;\; \operatorname{vec} \boldsymbol{B} = (\boldsymbol{B}_D \odot \cdots \odot \boldsymbol{B}_1)\mathbf{1}_R.$$

Throughout the article, we adopt the following notations. Y is a univariate response variable, $\boldsymbol{Z} \in \mathbb{R}^{p_0}$ denotes a p_0-dimensional vector of covariates, such as age and sex, and $\boldsymbol{X} \in \mathbb{R}^{p_1 \times \ldots \times p_D}$ is a D-dimensional array-valued predictor. For instance, for MRI, $D = 3$, representing the 3D structure of an image, whereas for fMRI, $D = 4$, with an additional time dimension. The lower-case triplets $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$, $i = 1, \ldots, n$, denote the independent, observed sample instances of $(Y, \boldsymbol{X}, \boldsymbol{Z})$.

2.2 Motivation and Basic Model

To motivate our model, we first start with a vector-valued $\boldsymbol{X}$ and absorb $\boldsymbol{Z}$ into $\boldsymbol{X}$. In the classical GLM (McCullagh and Nelder, 1983) setting, Y belongs to an exponential family with probability mass function or density

$$p(y|\theta, \phi) = \exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}, \tag{2}$$

where θ and $\phi > 0$ denote the natural and dispersion parameters. The classical GLM relates a vector-valued $\boldsymbol{X} \in \mathbb{R}^p$ to the mean $\mu = E(Y|\boldsymbol{X})$ via $g(\mu) = \eta = \alpha + \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}$, where $g(\cdot)$ is a strictly increasing link function, and η denotes the linear systematic part with intercept α and the coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^p$.

Next, for a matrix-valued covariate $\boldsymbol{X} \in \mathbb{R}^{p_1 \times p_2}$ $(D = 2)$, it is intuitive to consider a GLM model with the systematic part given by

$$g(\mu) = \alpha + \boldsymbol{\beta}_1^{\mathsf{T}}\boldsymbol{X}\boldsymbol{\beta}_2,$$

where $\boldsymbol{\beta}_1 \in \mathbb{R}^{p_1}$ and $\boldsymbol{\beta}_2 \in \mathbb{R}^{p_2}$, respectively. The bilinear form $\boldsymbol{\beta}_1^{\mathsf{T}}\boldsymbol{X}\boldsymbol{\beta}_2$ is a natural extension of the linear term $\beta^{\mathsf{T}}\boldsymbol{X}$ in the classical GLM with a vector covariate $\boldsymbol{X}$. It is interesting to note that, this bilinear form was first proposed by Li et al. (2010) in the context of dimension reduction, and then employed by Hung and Wang (2011) in the logistic regression with matrix-valued covariates $(D = 2)$. Moreover, note that $\boldsymbol{\beta}_1^{\mathsf{T}}\boldsymbol{X}\boldsymbol{\beta}_2 = (\boldsymbol{\beta}_2 \otimes \boldsymbol{\beta}_1)^{\mathsf{T}}\text{vec}(\boldsymbol{X})$.

Now for a conventional vector-valued covariate $\boldsymbol{Z}$ and a general array-valued $\boldsymbol{X} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$, we propose a GLM with the systematic part given by

$$g(\mu) = \alpha + \boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{Z} + (\boldsymbol{\beta}_D \otimes \ldots \otimes \boldsymbol{\beta}_1)^{\mathsf{T}}\text{vec}(\boldsymbol{X}), \tag{3}$$

where $\boldsymbol{\gamma} \in \mathbb{R}^{p_0}$ and $\boldsymbol{\beta}_d \in \mathbb{R}^{p_d}$ for $d = 1, \ldots, D$. This is our *basic model* for regression with array covariates. The key advantage of model (3) is that it dramatically reduces the dimensionality of the tensor component, from the order of $\prod_d p_d$ to the order of $\sum_d p_d$. Take MRI imaging as an example, the size of a typical image is $256^3 = 16,777,216$. If we simply turn $\boldsymbol{X}$ into a vector and fit a GLM, this brutal force solution is over 16 million-dimensional, and the computation is practically infeasible. In contrast, the multilinear model (3) is only $256 + 256 + 256 = 768$-dimensional. The reduction in dimension, and consequently in computational cost, is substantial.

A critical question then is whether such a massive reduction in the number of parameters would limit the capacity of model (3) to capture regions of interest with specific shapes. The illustrative example in Figure 1 provides some clues. In Figure 1, we present several two-dimensional images $\boldsymbol{B} \in \mathbb{R}^{64 \times 64}$ (shown in the first column), along with the

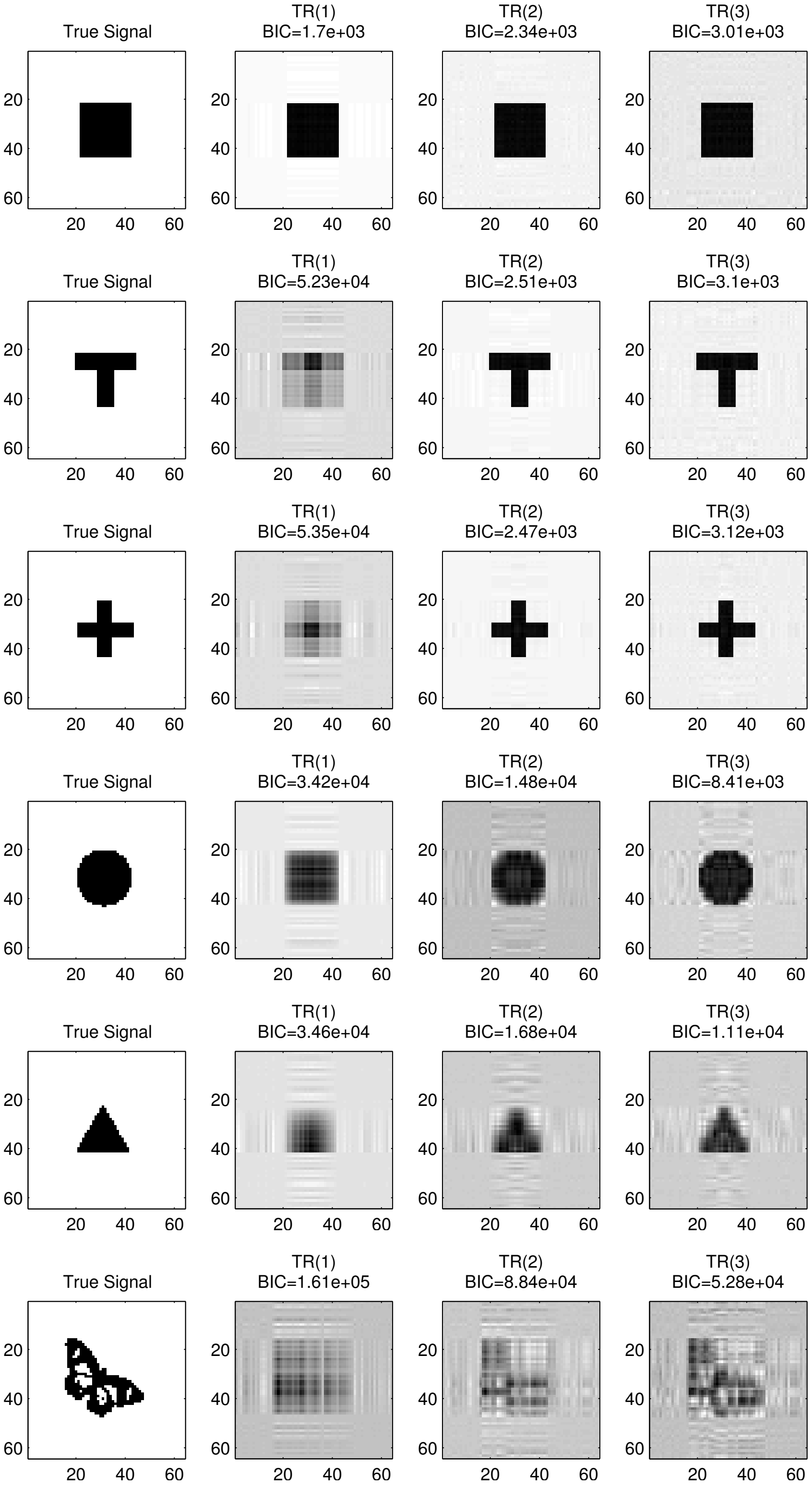


Figure 1: True and recovered image signals by tensor regression. The matrix variate has size 64 by 64 with entries generated as independent standard normals. The regression coefficient for each entry is either 0 (white) or 1 (black). The sample size is 1000. TR(R) means estimate from the rank-R tensor regression.

estimated images by model (3) (in the second column labeled by TR(1)). Specifically, we simulated 1,000 univariate responses y_i according to a normal model with mean $\mu_i = \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{z}_i + \langle \boldsymbol{B}, \boldsymbol{x}_i \rangle$, where $\boldsymbol{\gamma} = \mathbf{1}_5$. The inner product between two arrays is defined as $\langle \boldsymbol{B}, \boldsymbol{X} \rangle = \langle \mathrm{vec} \boldsymbol{B}, \mathrm{vec} \boldsymbol{X} \rangle = \sum_{i_1,\ldots,i_D} \beta_{i_1 \ldots i_D} x_{i_1 \ldots i_D}$. The coefficient array $\boldsymbol{B}$ is binary, with the true signal region equal to one and the rest zero. The regular covariate $\boldsymbol{z}_i$ and image covariate $\boldsymbol{x}_i$ are randomly generated with all elements being independent standard normals. Our goal is to see if model (3) can identify the true signal region in $\boldsymbol{B}$ using data $(y_i, \boldsymbol{z}_i, \boldsymbol{x}_i)$. Before examining the outcome, we make two remarks about this illustration. First, our problem differs from the usual edge detection or object recognition in imaging processing (Qiu, 2005, 2007). In our setup, all elements of the image $\boldsymbol{X}$ follow the same distribution. The signal region is defined through the coefficient image $\boldsymbol{B}$ and needs to be inferred from the association between Y and $\boldsymbol{X}$ after adjusting for $\boldsymbol{Z}$. Second, the classical GLM is difficult to apply in this example if we simply treat vec($\boldsymbol{X}$) as a covariate vector, since the sample size $n = 1,000$ is much less than the number of parameters $p = 5 + 64 \times 64 = 4,101$. Back to Figure 1, the second column clearly demonstrates the ability of model (3) in identifying the rectangular (square) type region (parallel to the image edges). On the other hand, since the parameter vector $\boldsymbol{\beta}_d$ in a rank-1 model is only able to capture the accumulative signal along the d-th dimension of the array variate $\boldsymbol{X}$, it is unsurprising that it does not perform well for signals that are far away from rectangle, such as triangle, disk, T-shape and butterfly. This motivates us to develop a more flexible tensor regression model in the next section.

2.3 Tensor Regression Model

We start with an alternative view of the basic model (3), which will lead to its generalization. Consider a D-dimensional array variate $\boldsymbol{X} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$, and a full coefficient array $\boldsymbol{B}$ of same size that captures the effects of each array element. Then the most flexible GLM suggests a linear systematic part

$$g(\mu) = \alpha + \gamma^{\mathsf{T}} \boldsymbol{Z} + \langle \boldsymbol{B}, \boldsymbol{X} \rangle.$$

The issue with this model is that $\boldsymbol{B}$ has the same number of parameters, $\prod_{d=1}^{D} p_d$, as $\boldsymbol{X}$, which is ultrahigh dimensional and far exceeds the usual sample size. Then a natural idea is to approximate $\boldsymbol{B}$ with less parameters. If $\boldsymbol{B}$ admits a rank-1 decomposition (1), i.e., $\boldsymbol{B} = \boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_2 \circ \cdots \circ \boldsymbol{\beta}_D$, where $\beta_d \in \mathbb{R}^{p_d}$, then by Lemma 1, we have

$$\operatorname{vec} \boldsymbol{B} = \operatorname{vec} (\boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_2 \circ \cdots \circ \boldsymbol{\beta}_D) = \boldsymbol{\beta}_D \odot \cdots \odot \boldsymbol{\beta}_1 = \boldsymbol{\beta}_D \otimes \cdots \otimes \boldsymbol{\beta}_1.$$

In other words, model (3) is indeed a *data-driven* model with a rank-1 approximation to the general signal array $\boldsymbol{B}$. This observation motivates us to consider a more flexible tensor regression model.

Specifically, we propose a family of *rank-R generalized linear tensor regression models*, in which the systematic part of GLM is of the form

$$\begin{aligned} g(\mu) &= \alpha + \gamma^{\mathsf{T}} \boldsymbol{Z} + \langle \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \cdots \circ \boldsymbol{\beta}_D^{(r)}, \boldsymbol{X} \rangle \\ &= \alpha + \gamma^{\mathsf{T}} \boldsymbol{Z} + \langle (\boldsymbol{B}_D \odot \cdots \odot \boldsymbol{B}_1) \mathbf{1}_R, \operatorname{vec} \boldsymbol{X} \rangle, \end{aligned} \tag{4}$$

where $\boldsymbol{B}_d = [\boldsymbol{\beta}_d^{(1)}, \ldots, \boldsymbol{\beta}_d^{(R)}] \in \mathbb{R}^{p_d \times R}$, $\boldsymbol{B}_D \odot \cdots \odot \boldsymbol{B}_1 \in \mathbb{R}^{\prod_d p_d \times R}$ is the Khatri-Rao product and $\mathbf{1}_R$ is the vector of R ones.. Equivalently we assume that the tensor regression parameter admits a rank-R decomposition $\boldsymbol{B} = [\![\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D]\!]$. When $R = 1$, it reduces to model (3). A few remarks on (4) are in order. First, since our formulation only deals with the linear predictor part of the model, it easily extends to the quasi-likelihood models (McCullagh and Nelder, 1983) where more general mean-variance relation is assumed. Second, for simplicity, we only discuss exponential family with a univariate response. Extension to multivariate exponential family, such as multinomial logit model, is straightforward. Third, due to the GLM setup (2), we call (4) a generalized linear tensor regression model. However, we should bear in mind that the systematic component η is a polynomial rather than linear in the parameters $\boldsymbol{B}_d$. Finally, the rank-R tensor decomposition (1) is called canonical decomposition or parallel factors (CANDECOMP/PARAFAC, or CP) in psychometrics (Kolda and Bader, 2009). In that sense, model (4) can be viewed as a *supervised* version of the classical CP decomposition for multi-dimensional arrays.

The number of parameters in model (4) is $p_0 + R \sum_d p_d$, which is still substantially smaller than $p_0 + \prod_d p_d$. With such a massive reduction in dimensionality, however, it provides a reasonable approximation to many low rank signals. Returning to the previous illustration, in Figure 1, images TR(R) are the recovered signals by the rank-R tensor regression (in third and fourth columns). The square signal can be perfectly recovered by a rank-1 model, whereas rank-2 and 3 regressions show signs of overfitting. The T-shape and cross signals can be perfectly recovered by a rank-2 regression. Triangle, disk, and butterfly shapes cannot be exactly recovered by any low rank approximations; however, a rank 3 tensor regression already yields a fairly informative recovery. Clearly, the general tensor regression model (4) is able to capture significantly more tensor signals than the basic model (3).

3 Estimation

We pursue the maximum likelihood (ML) route for parameter estimation in model (4). Given n i.i.d. data $\{(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i), i = 1, \ldots, n\}$, the log-likelihood function for (2) is

$$\ell(\alpha, \boldsymbol{\gamma}, \boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi), \tag{5}$$

where θ_i is related to regression parameters $(\alpha, \boldsymbol{\gamma}, \boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)$ through (4). We propose an efficient algorithm for maximizing $\ell(\alpha, \boldsymbol{\gamma}, \boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)$. A key observation is that although $g(\mu)$ in (4) is not linear in $(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)$ jointly, it is linear in $\boldsymbol{B}_d$ individually. This suggests alternately updating $(\alpha, \boldsymbol{\gamma})$ and $\boldsymbol{B}_d$, $d = 1, \ldots, D$, while keeping other components fixed. It yields a so-called *block relaxation algorithm* (de Leeuw, 1994; Lange, 2010). An appealing feature of this algorithm is that at each iteration, updating a block $\boldsymbol{B}_d$ is simply a classical GLM problem. To see this, when updating $\boldsymbol{B}_d \in \mathbb{R}^{p_d \times R}$, we rewrite the array inner product in (4) as

$$\langle \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \cdots \circ \boldsymbol{\beta}_D^{(r)}, \boldsymbol{X} \rangle = \langle \boldsymbol{B}_d, \boldsymbol{X}_{(d)}(\boldsymbol{B}_D \odot \cdots \odot \boldsymbol{B}_{d+1} \odot \boldsymbol{B}_{d-1} \odot \cdots \odot \boldsymbol{B}_1) \rangle.$$

Consequently the problem turns into a traditional GLM regression with Rp_d parameters, and the estimation procedure breaks into a sequence of low dimensional GLM optimizations and is extremely easy to implement using ready statistical softwares such as R, S+, SAS, and Matlab. The full estimation procedure is summarized in Algorithm 1. For the Gaussian models, it reduces to the alternating least squares (ALS) procedure (de Leeuw et al., 1976).

Algorithm 1 Block relaxation algorithm for maximizing (5).

```
Initialize: $(\alpha^{(0)}, \boldsymbol{\gamma}^{(0)}) = \operatorname{argmax}_{\alpha, \boldsymbol{\gamma}} \ell(\alpha, \boldsymbol{\gamma}, \mathbf{0}, \ldots, \mathbf{0})$, $\boldsymbol{B}_d^{(0)} \in \mathbb{R}^{p_d \times R}$ a random matrix for $d = 1, \ldots, D$.
repeat
  for $d = 1, \ldots, D$ do
    $\boldsymbol{B}_d^{(t+1)} = \operatorname{argmax}_{\boldsymbol{B}_d} \ell(\alpha^{(t)}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{B}_1^{(t+1)}, \ldots, \boldsymbol{B}_{d-1}^{(t+1)}, \boldsymbol{B}_d, \boldsymbol{B}_{d+1}^{(t)}, \ldots, \boldsymbol{B}_D^{(t)})$
  end for
  $(\alpha^{(t+1)}, \boldsymbol{\gamma}^{(t+1)}) = \operatorname{argmax}_{\alpha, \boldsymbol{\gamma}} \ell(\alpha, \boldsymbol{\gamma}, \boldsymbol{B}_1^{(t+1)}, \ldots, \boldsymbol{B}_D^{(t+1)})$
until $\ell(\boldsymbol{\theta}^{(t+1)}) - \ell(\boldsymbol{\theta}^{(t)}) < \epsilon$
```

As the block relaxation algorithm monotonically increases the objective function, it is numerically stable and the convergence of objective values $\ell(\boldsymbol{\theta}^{(t)})$ is guaranteed whenever $\ell(\boldsymbol{\theta})$ is bounded from above. Therefore the stopping rule of Algorithm 1 is well-defined. We denote the algorithmic map by M, i.e., $M(\boldsymbol{\theta}^{(t)}) = \boldsymbol{\theta}^{(t+1)}$, with $\boldsymbol{\theta} = (\alpha, \boldsymbol{\gamma}, \boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)$ collecting all parameters. Convergence properties of Algorithm 1 are summarized in Proposition 1.

Proposition 1. *Assume (i) the log-likelihood function $\ell(\boldsymbol{\theta})$ is continuous, coercive, i.e., the set $\{\boldsymbol{\theta} : \ell(\boldsymbol{\theta}) \geq \ell(\boldsymbol{\theta}^{(0)})\}$ is compact, and bounded above, (ii) the objective function in each block update of Algorithm 1 is strictly concave, and (iii) the set of stationary points (modulo scaling and permutation indeterminancy) of $\ell(\boldsymbol{\theta})$ are isolated. We have the following results.*

1. *(Global Convergence) The sequence $\boldsymbol{\theta}^{(t)} = (\alpha^{(t)}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{B}_1^{(t)}, \ldots, \boldsymbol{B}_D^{(t)})$ generated by Algorithm 1 converges to a stationary point of $\ell(\boldsymbol{\theta})$.*

2. *(Local Convergence) Let $\boldsymbol{\theta}^{(\infty)} = (\alpha^{(\infty)}, \boldsymbol{\gamma}^{(\infty)}, \boldsymbol{B}_1^{(\infty)}, \ldots, \boldsymbol{B}_D^{(\infty)})$ be a strict local maximum of $\ell(\boldsymbol{\theta})$. The iterates generated by Algorithm 1 are locally attracted to $\boldsymbol{\theta}^{(\infty)}$ for $\boldsymbol{\theta}^{(0)}$ sufficiently close to $\boldsymbol{\theta}^{(\infty)}$.*

We make a few quick remarks. First, although a stationary point is not guaranteed to be even a local maximum (it can be a saddle point), in practice the block relaxation algorithm almost always converges to at least a local maximum. In general, the algorithm should be run from multiple initializations to locate an excellent local maximum, especially for higher rank models with limited sample size. Second, $\ell(\boldsymbol{\theta})$ is not required to be jointly concave in $\boldsymbol{\theta}$. Only the concavity in the blocks of variables is needed. This condition holds for all GLM with canonical link such as linear model, logistic model and Poisson log-linear model.

The above algorithm assumes a known rank when estimating $\boldsymbol{B}$. Estimating an appropriate rank for our tensor model (4) is of practical importance. It can be formulated as a model selection problem, and we adopt the usual model section criterion, e.g., Bayesian information criterion (BIC), $-2\ell(\boldsymbol{\theta})+\log(n)p_e$, where p_e is the effective number of parameters for model (4): $p_e = R(p_1+p_2)-R^2$ for $D=2$, and $p_e = R(\sum_d p_d - D + 1)$ for $D > 2$. Returning to the illustrative example in Section 2.2, we fitted a rank-1, 2 and 3 tensor models, respectively, to various signal shapes. The corresponding BIC values are shown in Figure 1. The criterion is seen correctly estimating the rank for square as 1, and the rank for T and cross as 2. The true ranks for disk, triangle and butterfly are above 3, and their BIC values at rank 3 are smallest compared to those at 1 and 2.

4 Theory

We study the statistical properties of maximum likelihood estimate (MLE) for the tensor regression model defined by (2) and (4). For simplicity, we omit the intercept α and the classical covariate part $\boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{Z}$, though the conclusions generalize to an arbitrary combination of covariates. We adopt the usual asymptotic setup with a fixed number of parameters p and a diverging sample size n, because this is an important first step toward a comprehensive understanding of the theoretical properties of the proposed model. The asymptotics with a diverging p is our future work and is pursued elsewhere.

4.1 Score and Information

We first derive the score and information for the tensor regression model, which are essential for statistical estimation and inference. The following standard calculus notations are used. For a scalar function f, ∇f is the (column) gradient vector, $df = [\nabla f]^{\mathsf{T}}$ is the differential, and $d^2 f$ is the Hessian matrix. For a multivariate function $g : \mathbb{R}^p \mapsto \mathbb{R}^q$, $Dg \in \mathbb{R}^{q \times p}$ denotes the Jacobian matrix holding partial derivatives $\partial g_i / \partial x_j$.

We start from the Jacobian and Hessian of the systematic part $\eta \equiv g(\mu)$ in (4). The proof is given in the Appendix.

Lemma 2. *1. The gradient $\nabla\eta(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) \in \mathbb{R}^{R\sum_{d=1}^D p_d}$ is*

$$\nabla\eta(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) = [\boldsymbol{J}_1 \ \boldsymbol{J}_2 \ \cdots \ \boldsymbol{J}_D]^{\mathsf{T}}(\text{vec}\boldsymbol{X}),$$

where $\boldsymbol{J}_d \in \mathbb{R}^{\prod_{d=1}^D p_d \times p_d R}$ is the Jacobian

$$\boldsymbol{J}_d = D\boldsymbol{B}(\boldsymbol{B}_d) = \boldsymbol{\Pi}_d[(\boldsymbol{B}_D \odot \cdots \odot \boldsymbol{B}_{d+1} \odot \boldsymbol{B}_{d-1} \odot \cdots \odot \boldsymbol{B}_1) \otimes \boldsymbol{I}_{p_d}] \tag{6}$$

and $\boldsymbol{\Pi}_d$ is the $(\prod_{d=1}^D p_d)$-by-$(\prod_{d=1}^D p_d)$ permutation matrix that reorders $\text{vec}\boldsymbol{B}_{(d)}$ to obtain $\text{vec}\boldsymbol{B}$, i.e., $\text{vec}\boldsymbol{B} = \boldsymbol{\Pi}_d \, \text{vec}\boldsymbol{B}_{(d)}$.

2. The Hessian $d^2\eta(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) \in \mathbb{R}^{R\sum_{d=1}^D p_d \times R\sum_{d=1}^D p_d}$ has entries

$$h_{(i_d,r),(i_{d'},r')} = 1_{\{r=r', d\neq d'\}} \sum_{j_d=i_d, j_{d'}=i_{d'}} x_{j_1,\ldots,j_D} \prod_{d''\neq d,d'} \beta_{j_{d''}}^{(r)},$$

and can be partitioned in D^2 blocks as

$$\begin{pmatrix} \boldsymbol{0} & * & * & * \\ \boldsymbol{H}_{21} & \boldsymbol{0} & * & * \\ \vdots & \vdots & \ddots & * \\ \boldsymbol{H}_{D1} & \boldsymbol{H}_{D2} & \cdots & \boldsymbol{0} \end{pmatrix}.$$

The block $\boldsymbol{H}_{dd'} \in \mathbb{R}^{p_d R \times p_{d'} R}$ has $p_d p_{d'} R$ nonzero elements which can be retrieved from the matrix $\boldsymbol{X}_{(dd')}(\boldsymbol{B}_D \odot \cdots \odot \boldsymbol{B}_{d+1} \odot \boldsymbol{B}_{d-1} \odot \cdots \odot \boldsymbol{B}_{d'+1} \odot \boldsymbol{B}_{d'-1} \odot \cdots \odot \boldsymbol{B}_1)$, where $\boldsymbol{X}_{(dd')}$ is the mode-(d,d') matricization of $\boldsymbol{X}$.

Remark 1: The Hessian $d^2\eta$ is highly sparse and structured. An entry in $d^2\eta(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)$ is nonzero only if it belongs to different directions d but the same outer product r.

Let $\ell(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D|y, \boldsymbol{x}) = \ln p(y|\boldsymbol{x}, \boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)$ be the log-density. Next result derives the score function, Hessian, and Fisher information of the tensor regression model.

Proposition 2. *Consider the tensor regression model defined by (2) and (4).*

1. The score function (or score vector) is

$$\nabla\ell(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) = \frac{(y-\mu)\mu'(\eta)}{\sigma^2}[\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]^{\mathsf{T}}(\text{vec}\boldsymbol{X}) \tag{7}$$

with $\boldsymbol{J}_d$, $d = 1, \ldots, D$, defined by (6).

2. The Hessian of the log-density ℓ is

$$\begin{aligned} H(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) = & -\frac{[\mu'(\eta)]^2}{\sigma^2}([\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]^{\mathsf{T}} \mathrm{vec} \boldsymbol{X})([\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]^{\mathsf{T}} \mathrm{vec} \boldsymbol{X})^{\mathsf{T}} \\ & + \frac{(y-\mu)\theta''(\eta)}{\sigma^2}([\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]^{\mathsf{T}} \mathrm{vec} \boldsymbol{X})([\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]^{\mathsf{T}} \mathrm{vec} \boldsymbol{X})^{\mathsf{T}} \\ & + \frac{(y-\mu)\theta'(\eta)}{\sigma^2} d^2\eta(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D), \qquad (8) \end{aligned}$$

with $d^2\eta$ defined in Lemma 2.

3. The Fisher information matrix is

$$\begin{aligned} \boldsymbol{I}(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) &= E[-H(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)] = \mathrm{Var}[\nabla \ell(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) d\ell(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)] \\ &= \frac{[\mu'(\eta)]^2}{\sigma^2}[\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]^{\mathsf{T}}(\mathrm{vec}\boldsymbol{X})(\mathrm{vec}\boldsymbol{X})^{\mathsf{T}}[\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]. \qquad (9) \end{aligned}$$

Remark 2: For canonical link, $\theta = \eta$, $\theta'(\eta) = 1$, $\theta''(\eta) = 0$, and the second term of Hessian vanishes. For the classical GLM with linear systematic part ($D = 1$), $d^2\eta(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)$ is zero and thus the third term of Hessian vanishes. For the classical GLM ($D = 1$) with canonical link, both the second and third terms of the Hessian vanish and thus the Hessian is non-stochastic, coinciding with the information matrix.

4.2 Identifiability

Before studying asymptotic property, we need to deal with the identifiability issue. The parameterization in the tensor model is nonidentifiable due to two complications. Consider a rank-R decomposition of an array, $\boldsymbol{B} = [\![\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D]\!]$. The first complication is the indeterminacy of $\boldsymbol{B}$ due to scaling and permutation:

- scaling: $\boldsymbol{B} = [\![\boldsymbol{B}_1\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{B}_D\boldsymbol{\Lambda}_D]\!]$ for any diagonal matrices $\boldsymbol{\Lambda}_d = \mathrm{diag}(\lambda_{d1}, \ldots, \lambda_{dR})$, $d = 1, \ldots, D$, such that $\prod_d \lambda_{dr} = 1$ for $r = 1, \ldots, R$.
- permutation: $\boldsymbol{B} = [\![\boldsymbol{B}_1\boldsymbol{\Pi}, \ldots, \boldsymbol{B}_D\boldsymbol{\Pi}]\!]$ for any R-by-R permutation matrix $\boldsymbol{\Pi}$.

For the matrix case ($D = 2$), a further complication is the nonsingular transformation indeterminancy: $\boldsymbol{B}_1\boldsymbol{B}_2^{\mathsf{T}} = \boldsymbol{B}_1\boldsymbol{O}\boldsymbol{O}^{-1}\boldsymbol{B}_2^{\mathsf{T}}$ for any R-by-R nonsingular matrix $\boldsymbol{O}$. Note the scaling and permutation indeterminancy is subsumed in the nonsingular transformation indeterminancy. The singular value decomposition (SVD) of a matrix is unique because it imposes orthonormality constraint on the columns of the factor matrices.

To deal with this complication, it is necessary to adopt a specific constrained parameterization to fix the scaling and permutation indeterminacy. For $D > 2$, we need to put $(D-1)R$ restrictions on the parameters $\boldsymbol{B}$ and apparently there is an infinite number of ways to do this. In this paper we adopt the following convention. $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_{D-1}$ are scaled such that $\beta_{d1}^{(r)} = 1$, i.e., the first rows are ones. This in turn determines entries in the first row of $\boldsymbol{B}_D$ and fixes scaling indeterminacy. To fix the permutation indeterminancy, we assume that the first row entries of $\boldsymbol{B}_D$ are distinct and arranged in descending order $\beta_{D1}^{(1)} > \cdots > \beta_{D1}^{(R)}$. The resulting parameter space is

$$\mathcal{B} = \{(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) : \beta_{d1}^{(r)} = 1, \text{ for } d = 1, \ldots, D, r = 1, \ldots, R \text{ and } \beta_{D1}^{(1)} > \cdots > \beta_{D1}^{(R)}\},$$

which is open and convex. The formulae for score, Hessian and information in Proposition 2 require changes accordingly, i.e., the entries in the first rows of $\boldsymbol{B}_d$, $d = 1, \ldots, D-1$, are fixed at ones and their corresponding entries, rows and columns in score, Hessian and information need to be deleted. Treatment for the $D = 2$ case is similar and omitted for brevity. We emphasize that our choice of the restricted space $\mathcal{B}$ is arbitrary and exclude many arrays that might be of interest, e.g., arrays with any entries in the first rows of $\boldsymbol{B}_d$, $d = 1, \ldots, D-1$, equal to zeros or with ties in the first row of $\boldsymbol{B}_D$. However the set of such exceptional arrays has Lebesgue measure zero. In specific applications, subject knowledge may suggest alternative constraints on the parameters.

The second complication comes from possible non-uniqueness of decomposition when $D > 2$ even after adjusting scaling and permutation indeterminacy. The next proposition collects some recent results that give easy-to-check conditions for the uniqueness (up to scaling and permutation) of decomposition. The first two are useful for checking uniqueness of a given tensor, while the latter two give general conditions for uniqueness almost everywhere in the $D = 3$ or 4 case.

Proposition 3. *Suppose that a D-dimensional array $\boldsymbol{B} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ has rank R.*

1. *(Sufficiency)(Sidiropoulos and Bro, 2000) The decomposition (1) is unique up to scaling and permutation if $\sum_{d=1}^{D} k_{\boldsymbol{B}_d} \geq 2R + (D-1)$, where $k_{\boldsymbol{A}}$ is the k-rank*

of a matrix $\boldsymbol{A}$, i.e., the maximum value k such that any k columns are linearly independent.

2. *(Necessity)(Liu and Sidiropoulos, 2001) If the decomposition (1) is unique up to scaling and permutation, then* $\min_{d=1,\ldots,D} \operatorname{rank}(\boldsymbol{B}_1 \odot \cdots \odot \boldsymbol{B}_{d-1} \odot \boldsymbol{B}_{d+1} \odot \cdots \odot \boldsymbol{B}_D) = R$*, which in turn implies that* $\min_{d=1,\ldots,D} \left(\prod_{d' \neq d} \operatorname{rank}(\boldsymbol{B}_{d'}) \right) \geq R$*.*

3. *(de Lathauwer, 2006) When $D = 3$, $R \leq p_3$ and $R(R-1) \leq p_1(p_1-1)p_2(p_2-1)/2$, the decomposition (1) is unique for almost all such tensors except on a set of Lebesgue measure zero.*

4. *(de Lathauwer, 2006) When $D = 4$, $R \leq p_4$ and $R(R-1) \leq p_1p_2p_3(3p_1p_2p_3 - p_1p_2 - p_1p_3 - p_2p_3 - p_1 - p_2 - p_3 + 3)/4$, the decomposition (1) is unique for almost all such tensors except on a set of Lebesgue measure zero.*

Next we give a sufficient and necessary condition for local identifiability. The proof follows from a classical result (Rothenberg, 1971) that relates local identifiability to the Fisher information matrix.

Proposition 4 (Identifiability)**.** *Given iid data points $\{(y_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$ from the tensor regression model. Let $\boldsymbol{B}_0 \in \mathcal{B}$ be a parameter point and assume there exists an open neighborhood of $\boldsymbol{B}_0$ in which the information matrix has a constant rank. Then $\boldsymbol{B}_0$ is locally identifiable up to permutation if and only if*

$$I(\boldsymbol{B}_0) = [\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]^{\mathsf{T}} \left[\sum_{i=1}^{n} \frac{\mu'(\eta_i)^2}{\sigma_i^2} (\operatorname{vec} \boldsymbol{x}_i)(\operatorname{vec} \boldsymbol{x}_i)^{\mathsf{T}} \right] [\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]$$

is nonsingular.

Remark 3.1: Proposition 4 explains the merit of tensor regression from another angle. For identifiability, the classical linear regression requires $\operatorname{vec} \boldsymbol{x}_i \in \mathbb{R}^{\prod_d p_d}$, $i = 1, \ldots, n$, to be linearly independent in order to estimate all parameters, which requires a sample size $n \geq \prod_d p_d$. The more parsimonious tensor regression only requires linearly independence of the "collapsed" vectors $[\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]^{\mathsf{T}} \operatorname{vec} \boldsymbol{x}_i \in \mathbb{R}^{R(\sum_d p_d - D + 1)}$, $i = 1, \ldots, n$. The requirement on sample size is greatly lessened by imposing structure on the arrays.

Remark 3.2: Although global identifiability is hard to check for a finite sample, a parameter point $\boldsymbol{B} \in \mathcal{B}$ is asymptotically and globally identifiable as far as it admits a unique decomposition up to scaling and permutation and $\sum_{i=1}^{n}(\mathrm{vec}\, \boldsymbol{x}_i)(\mathrm{vec}\, \boldsymbol{x}_i)^{\mathsf{T}}$ has full rank for $n \geq n_0$, or, when considered stochastically, $\mathbf{E}[(\mathrm{vec}\, \boldsymbol{X})(\mathrm{vec}\, \boldsymbol{X})^{\mathsf{T}}]$ has full rank. To see this, whenever $\sum_{i=1}^{n}(\mathrm{vec}\, \boldsymbol{x}_i)(\mathrm{vec}\, \boldsymbol{x}_i)^{\mathsf{T}}$ has full rank, the full coefficient array is globally identifiable and thus the decomposition is identifiable whenever it is unique.

Generalizing the concept of estimable functions for linear models, we call any linear combination of $\langle \boldsymbol{x}_i, \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} \circ \cdots \circ \boldsymbol{\beta}_D^{(r)} \rangle$, $i = 1, \ldots, n$, an estimable function. We can estimate estimable or collection of estimable functions even when the parameters are not identifiable.

4.3 Asymptotics

The asymptotics for tensor regression follow from those for MLE or M-estimation. The key observation is that the nonlinear part of tensor model (4) is a degree-D polynomial of parameters and the collection of polynomials $\{\langle \boldsymbol{B}, \boldsymbol{X} \rangle, \boldsymbol{B} \in \mathcal{B}\}$ form a Vapnik-Červonenkis (VC) class. Then standard uniform convergence theory for M-estimation (van der Vaart, 1998) applies.

Theorem 1 (Consistency)**.** *Assume $\boldsymbol{B}_0 = [\![\boldsymbol{B}_{01}, \ldots, \boldsymbol{B}_{0D}]\!] \in \mathcal{B}$ is (globally) identifiable up to permutation and the array covariates $\boldsymbol{X}_i$ are iid from a bounded distribution. The MLE is consistent, i.e., $\hat{\boldsymbol{B}}_n$ converges to $\boldsymbol{B}_0$ (modulo permutation) in probability, in the following models: (1) normal tensor regression with a compact parameter space $\mathcal{B}_0 \subset \mathcal{B}$; (2) binary tensor regression; and (3) poisson tensor regression with a compact parameter space $\mathcal{B}_0 \subset \mathcal{B}$.*

Remark 4: (Misspecified Rank) In practice it is rare that the true regression coefficient $\boldsymbol{B}_{\mathrm{true}} \in \mathrm{I\!R}^{p_1 \times \cdots \times p_D}$ is exactly a low rank tensor. However the MLE of the rank-R tensor model converges to the maximizer of function $M(\boldsymbol{B}) = \mathbb{P}_{\boldsymbol{B}_{\mathrm{true}}} \ln p_{\boldsymbol{B}}$ or equivalently $\mathbb{P}_{\boldsymbol{B}_{\mathrm{true}}} \ln(p_{\boldsymbol{B}}/p_{\boldsymbol{B}_{\mathrm{true}}})$. In other words, the MLE is consistently estimating the best rank-R approximation of $\boldsymbol{B}_{\mathrm{true}}$ in the sense of Kullback-Leibler distance.

To establish the asymptotic normality of $\hat{\boldsymbol{B}}_n$, we note that the log-likelihood function of tensor regression model is quadratic mean differentiable (q.m.d).

Lemma 3. *Tensor regression model is quadratic mean differentiable (q.m.d.).*

Theorem 2 (Asymptotic Normality)**.** *For an interior point* $\boldsymbol{B}_0 = [\![\boldsymbol{B}_{01}, \ldots, \boldsymbol{B}_{0D}]\!] \in \mathcal{B}$ *with nonsingular information matrix* $\boldsymbol{I}(\boldsymbol{B}_{01}, \ldots, \boldsymbol{B}_{0D})$ *(9) and* $\hat{\boldsymbol{B}}_n$ *is consistent,*

$$\sqrt{n}[\operatorname{vec}(\hat{\boldsymbol{B}}_{n1}, \ldots, \hat{\boldsymbol{B}}_{nD}) - \operatorname{vec}(\boldsymbol{B}_{01}, \ldots, \boldsymbol{B}_{0D})]$$

converges in distribution to a normal with mean zero and covariance $\boldsymbol{I}^{-1}(\boldsymbol{B}_{01}, \ldots, \boldsymbol{B}_{0D})$*.*

5 Regularized Estimation

The sample size in typical neuroimaging studies is quite small, and thus even for a rank-1 tensor regression (3), it is likely that the number of parameters exceeds the sample size. Therefore the $p \gg n$ challenge is a rule rather than an exception in neuroimaging analysis, and regularization becomes essential. Even when the sample size exceeds the number of parameters, regularization is still useful for stabilizing the estimates and improving their risk property. We emphasize that there are a large number of regularization techniques for different purposes. Here we illustrate with using *sparsity* regularization for identifying sub-regions that are associated with the response traits. This problem can be viewed as an analogue of variable selection in the traditional vector-valued covariates. Toward that end, we maximize a regularized log-likelihood function

$$\ell(\alpha, \boldsymbol{\gamma}, \boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) - \sum_{d=1}^{D} \sum_{r=1}^{R} \sum_{i=1}^{p_d} P_\lambda(|\beta_{di}^{(r)}|, \rho),$$

where $P_\lambda(|\beta|, \rho)$ is a scalar penalty function, ρ is the penalty tuning parameter, and λ is an index for the penalty family. Some widely used penalties include: power family (Frank and Friedman, 1993), in which $P_\lambda(|\beta|, \rho) = \rho|\beta|^\lambda$, $\lambda \in (0, 2]$, and in particular lasso (Tibshirani, 1996) ($\lambda = 1$) and ridge ($\lambda = 2$); elastic net (Zou and Hastie, 2005), in which $P_\lambda(|\beta|, \rho) = \rho[(\lambda - 1)\beta^2/2 + (2 - \lambda)|\beta|], \lambda \in [1, 2]$; and SCAD (Fan and Li, 2001), in which $\partial/\partial|\beta| P_\lambda(|\beta|, \rho) = \rho \left\{1_{\{|\beta| \le \rho\}} + (\lambda\rho - |\beta|)_+/(\lambda - 1)\rho 1_{\{|\beta| > \rho\}}\right\}$, $\lambda > 2$, among many others. Choice of penalty function and tuning parameters ρ and λ depends on particular purposes: prediction, unbiased estimation, or region selection.

Regularized estimation for tensor models incurs slight changes in Algorithm 1. When updating $\boldsymbol{B}_d$, we simply fit a penalized GLM regression problem,

$$\boldsymbol{B}_d^{(t+1)} = \\ \text{argmax}_{\boldsymbol{B}_d} \ell(\alpha^{(t)}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{B}_1^{(t+1)}, \ldots, \boldsymbol{B}_{d-1}^{(t+1)}, \boldsymbol{B}_d, \boldsymbol{B}_{d+1}^{(t)}, \ldots, \boldsymbol{B}_D^{(t)}) - \sum_{r=1}^{R} \sum_{i=1}^{p_d} P_\lambda(|\beta_{di}^{(r)}|, \rho),$$

for which many software packages exist. Same paradigm certainly applies to regularizations other than sparsity. The fitting procedure boils down to alternating regularized GLM regression. The monotone ascent property of Algorithm 1 is retained under the modified algorithm. Convex penalties, such as elastic net and power family with $\lambda \geq 1$, tend to convexify the objective function and alleviate the local maximum problem. On the other hand, concave penalty such as power family with $\lambda < 1$ and SCAD produces more unbiased estimates but the regularized objective function is more ruggy and in practice the algorithm should be initialized from multiple start points to increase the chance of finding a global maximum. Many methods are available to guide the choice of the tuning parameter ρ and/or λ for regularized GLM, notably AIC, BIC and cross validation. For instance the recent work (Zhou et al., 2011) derives BIC type criterion for GLM with possibly non-concave penalties such as power family, which can be applied to regularized tensor regression models in a straightforward way.

Two remarks are in order. First, it is conceptually possible to apply these regularization techniques directly to the full coefficient array $\boldsymbol{B} \in \mathbb{R}^{\prod_d p_d}$ without considering any structured decomposition as in our models. That is, one simply treats vec$\boldsymbol{X}$ as the predictor vector as employed in the classical total variation regularization in image denoising and recovery. However, for the brain imaging data, we should bear in mind the dimensionality of the imaging arrays. For instance, to the best of our knowledge, no software is able to deal with fused lasso or even simple lasso on $64^3 = 262,144$ or $256^3 = 16,777,216$ variables. This ultrahigh dimensionality certainly corrupts the statistical properties of the regularized estimates too. Second, penalization is only one form of regularization. In specific applications, prior knowledge often suggests various constraints among parameters, which may be exploited to regularize parameter estimate. For instance, for MRI imaging data, sometimes it may be reasonable to impose symmetry on the parameters along the coronal plane, which effectively reduces the dimensionality by $p_d R/2$. In many applications, nonnegativity of parameter values is also enforced.

6 Numerical Analysis

We have carried out an extensive numerical study to investigate the finite sample performance of the proposed methods. In this section, we report selected results from synthetic examples and an analysis of a real brain imaging data.

6.1 2D Shape Examples

We first elaborate on the illustrative example given in Section 2.2 with a collection of 2D shapes. We examine the performance of the tensor model under a variety of sample sizes and signal strengths, and compare the estimates with and without regularization. More specifically, the tensor model in Section 2.2 is employed, where the response is normally distributed with mean, $\eta = \gamma^{\mathsf{T}} \boldsymbol{Z} + \langle \boldsymbol{B}, \boldsymbol{X} \rangle$, and standard deviation σ. $\boldsymbol{X}$ is a 64×64 2D matrix, $\boldsymbol{Z}$ is a 5-dimensional covariate vector, both of which have standard normal entries, $\boldsymbol{\gamma} = (1,1,1,1,1)^{\mathsf{T}}$, and $\boldsymbol{B}$ is binary with the true signal region equal to one and the rest zero. We fit both a rank-3 tensor model without regularization, and one with a lasso regularization. For sample size, we examine $n = 200, 300, 400, 500$ and 750. Note that, for this example, the number of parameters of a rank-3 model is $380 = 5 + 3 \times (64 + 64) - 3^2$. As such, there are multiple solutions when $n = 200$ or 300, and we arbitrarily choose one estimate. For signal strength, we vary the noise level $\sigma = 50\%, 20\%, 10\%, 5\%$ and 1% of the standard deviation of the mean η, respectively.

We summarize the results in three plots: the snapshots of estimates with varying sample size, the snapshots with varying signal strength, and the line plot of the average root mean squared error (RMSE) for estimation of $\boldsymbol{B}$. For space consideration, only the first plot is presented in Figure 2 (with 10% noise level), and the rest in the supplementary appendix. We make the following observations. First, estimation accuracy steadily increases with the sample size, demonstrating consistency of the proposed method. This can be seen from both the snapshots with improved quality and the decreasing RMSE. Similar patterns are observed with increasing signal strength. Second, regularization clearly improves estimation, especially when the sample size is limited. In practice, when the number of imaging subjects is moderate, regularized tensor regression is recommended.

For this example, we also examined the recovered signals by regularized tensor regression with a fixed sample size $n = 500$ and varying penalty parameter. The results are reported in the appendix A.2.

6.2 Attention Deficit Hyperactivity Disorder Data Analysis

We applied our methods to the attention deficit hyperactivity disorder (ADHD) data from the ADHD-200 Sample Initiative (http://fcon_1000.projects.nitrc.org/indi/adhd200/). ADHD is a common childhood disorder and can continue through adolescence and adulthood. Symptoms include difficulty in staying focused and paying attention, difficulty in controlling behavior, and over-activity. The data set that we used is part of the ADHD-200 Global Competition data sets. It consists of 776 subjects, with 491 normal controls and 285 combined ADHD subjects. Among them, there are 442 males with mean age 12.0 years and standard deviation 3.1 years, and 287 females with mean age 11.9 years and standard deviation 3.5 years. We removed 47 subjects due to the missing observations or poor image quality. Resting state fMRIs and T1-weighted images were acquired for each subject. The T1-weighted images were preprocessed by standard steps including AC (anterior commissure) and PC (posterior commissure) correction, N2 bias field correction, skull-stripping, intensity inhomogeneity correction, cerebellum removal, segmentation, and registration. After segmentation, the brains were segmented into four different tissues: grey matter (GM), white matter (WM), ventricle (VN), and cerebrospinal fluid (CSF). We quantified the local volumetric group differences by generating RAVENS maps (Davatzikos et al., 2001) for the whole brain and each of the segmented tissue type (GM, WM, VN, and CSF) respectively, using the deformation field we obtained during registration. RAVENS methodology is based on a volume-preserving spatial transformation, which ensures that no volumetric information is lost during the process of spatial normalization, since this process changes an individual's brain morphology to conform it to the morphology of a template. In addition to im-

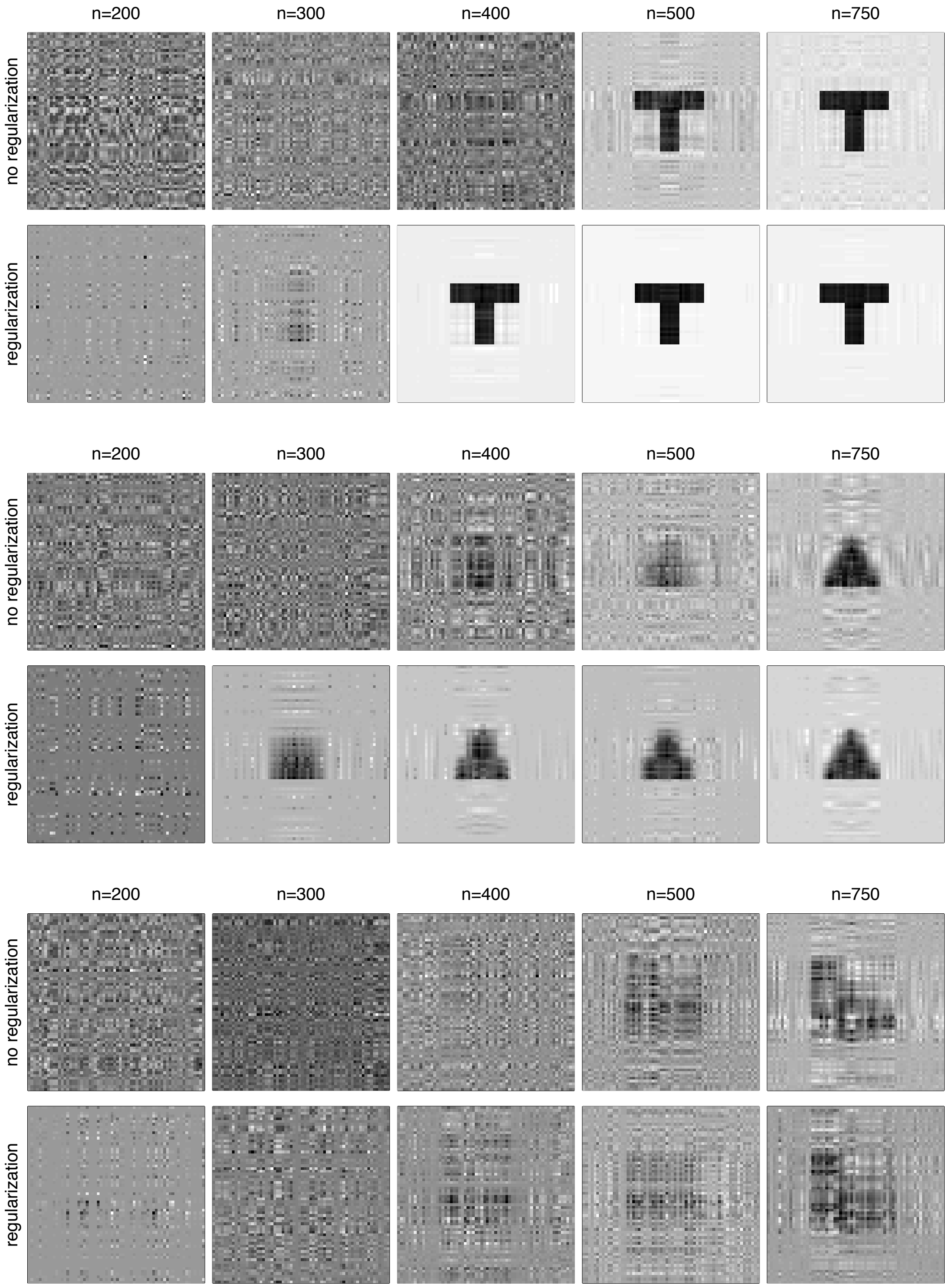
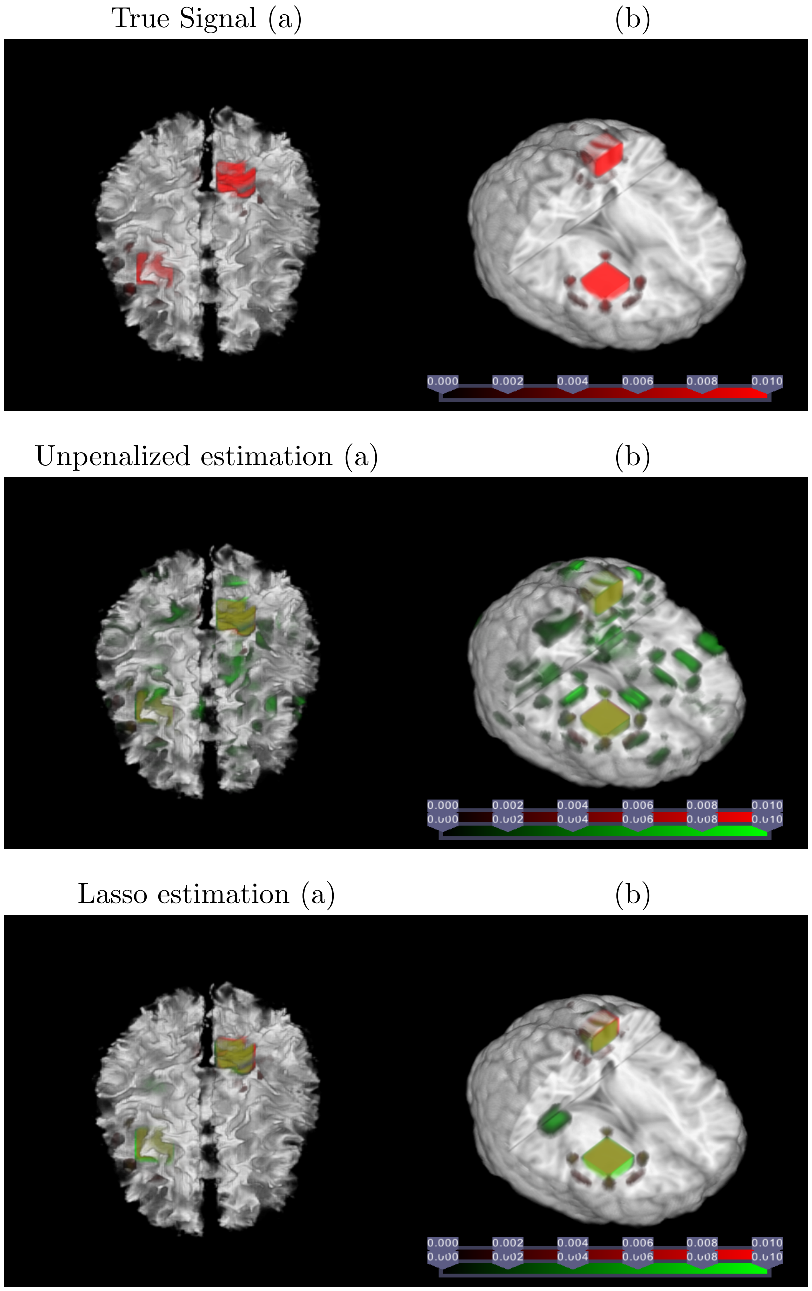


Figure 2: Snapshots of tensor estimation with varying sample size. The matrix variate has size 64 by 64 with entries generated as independent standard normals. The regression coefficient for each entry is either 0 (white) or 1 (black).

age covariates, we include the subjects' age, gender, and whole brain volume as regular covariates. One scientific question of interest is to understand association between the disease outcome and the brain image patterns after adjustment for the clinical and demographical variables. We first examined the case with real image covariates and simulated responses. The goal is to study the empirical performance of our methods under various response models. We then showed the performance of the regularized estimation in terms of region selection. Finally, we applied the method to the data with the true observed binary response.

6.2.1 Real Image Covariates and Simulated Response

We first consider a number of GLMs with the real brain image covariates, where $\eta = \boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{Z} + \langle \boldsymbol{B}, \boldsymbol{X} \rangle$, the signal tensor $\boldsymbol{B}$ admits a certain structure, $\boldsymbol{\gamma} = (1,1,1)^{\mathsf{T}}$, $\boldsymbol{X}$ denotes the 3D MRI image with dimension 256 × 256 × 198, and $\boldsymbol{Z}$ denotes the vector of age, gender and whole brain volume. We consider two structures for $\boldsymbol{B}$. The first admits a rank one decomposition, with $\boldsymbol{B}_1 \in \mathrm{I\!R}^{256\times 1}$, $\boldsymbol{B}_2 \in \mathrm{I\!R}^{256\times 1}$, and $\boldsymbol{B}_3 \in \mathrm{I\!R}^{198\times 1}$, and all of whose $(90+j)$th element equal to $\sin(j\pi/14)$ for $j = 0, 1, \ldots, 14$. This corresponds to a single-ball signal in a 3D space. The second admits a rank two decomposition, with $\boldsymbol{B}_1 \in \mathrm{I\!R}^{256\times 2}$, $\boldsymbol{B}_2 \in \mathrm{I\!R}^{256\times 2}$, and $\boldsymbol{B}_3 \in \mathrm{I\!R}^{198\times 2}$. All the first columns of $\boldsymbol{B}_d$ have their $(90+j)$th element equal to $\sin(j\pi/14)$, and the second columns of $\boldsymbol{B}_d$ have their $(140+j)$th element equal to $\sin(j\pi/14)$ for $j = 0, 1, \ldots, 14$. This mimics a two-ball signal in the 3D space. We then generate the response through the GLM models: for the normal model, $Y \sim \text{Normal}(\mu, 1)$, where $\mu = \eta$; for the binomial model, $Y \sim \text{Bernoulli}(p)$, with $p = 1/[1 + \exp(-0.1\eta)]$; and for the poisson model, $Y \sim \text{Poission}(\mu)$, with $\mu = \exp(0.01\eta)$. Table 1 summarizes the average RMSE and its standard deviation out of 100 data replications. We see that the normal and poisson responses both have competitive performance, whereas the binomial case is relatively more challenging. The two-ball signal is more challenging than a one-ball signal, and overall the tensor models work well across different response types and different signals.


True Signal (a)

(b)

Unpenalized estimation (a)

(b)

Lasso estimation (a)

(b)

Figure 3: Region selection. The true signal regions are colored in red, the estimated signal regions are in green, and the overlapped regions are in yellow. The left panel is the true or estimated signal overlaid on a randomly selected subject, and the right panel is a 3D rendering of the true or estimated signal overlaid on the template

Table 1: Tensor regression estimation for the ADHD data. Reported are mean RMSE and its standard deviation (in parenthesis) of evaluation criteria based on 100 data replications.

Signal	Param.	Normal	Binomial	Poisson
one-ball	$\boldsymbol{\gamma}$	0.0639 (0.0290)	0.2116 (0.0959)	0.0577 (0.0305)
	$\boldsymbol{B}$	0.0039 (0.0002)	0.0065 (0.0002)	0.0064 (0.0002)
two-ball	$\boldsymbol{\gamma}$	0.0711 (0.0310)	0.3119 (0.1586)	0.0711 (0.0307)
	$\boldsymbol{B}$	0.0058 (0.0002)	0.0082 (0.0003)	0.0083 (0.0003)

6.2.2 Regularized Estimation

Next we focus on the ability of the regularized tensor regression model to identify relevant regions in brain associated with the response. This is analogous to the variable selection problem in the traditional regression with vector-valued covariates. We employ the two-ball signal and the normal model in Section 6.2.1. Figure 3 shows images with the true signal, the un-regularized tensor regression estimate, and the regularized tensor regression estimates with a lasso penalty, respectively, overlaid on an image of an arbitrarily chosen subject, or on a 3D rendering of a template. The plots clearly show that the true sparse signal regions can be well recovered through regularization.

6.2.3 Real Data Analysis

Finally, we analyze the ADHD data with the observed binary diagnosis status as the response. We fitted a rank-3 tensor logistic regression model, since in practice it is rare that the true signal would follow an exact reduced rank formulation. We also applied the regularized estimation using a lasso penalty. Figure 4 shows the results. Inspecting Figure 4 reveals two regions of interest: left temporal lobe white matter and the splenium that connects parietal and occipital cortices across the midline in the corpus callosum. The anatomical disturbance in the temporal lobe has been consistently revealed and its interpretation would be consistent with a finer-grained analysis of the morphological features of the cortical surface, which reported prominent volume reductions in the temporal and frontal cortices in children with ADHD compared with matched controls

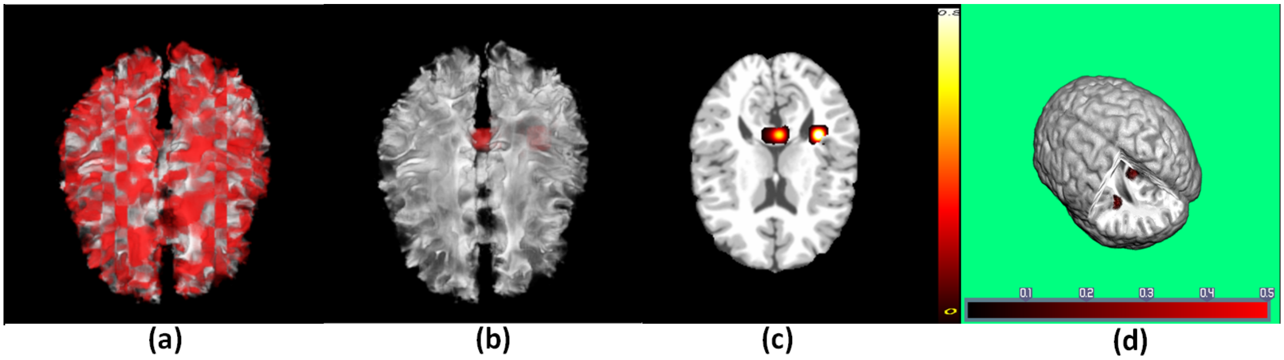


Figure 4: Application to the ADHD data. Panel (a) is the unpenalized estimate overlaid on a randomly selected subject; (b) is the regularized estimate overlaid on a randomly selected subject; (c) is a selected slice of the regularized estimate overlaid on the template; and (d) is a 3D rendering of the regularized estimate overlaid on the template.

(Sowell et al., 2003). Moreover, a reduced size of the splenium is the most reliable finding in the corpus callosum (Valera et al., 2007).

7 Discussion

We have proposed a tensor decomposition based approach for regression modeling with array covariates. The curse of dimensionality is lessened by imposing a low rank approximation to the extremely high-dimensional full coefficient array. This allows development of a fast estimation algorithm and regularization. Numerical analysis demonstrates that, despite its massive reduction, the method works well in recovering various geometric as well as natural shape images. Although there has been imaging studies utilizing tensor structure (Li et al., 2005; Park and Savvides, 2007), our proposal, to the best of our knowledge, is the first that integrates tensor decomposition within a statistical regression (*supervised learning*) paradigm.

A motivation of our work is the recent emergence of large-scale neuroimaging data. Early imaging studies usually have only a handful of subjects. More recently, however, a number of large-scale brain imaging studies are accumulating imaging data from a much larger number of subjects. For instance, the Attention Deficit Hyperactivity Disorder Sample Initiative (ADHD, 2012) consists of 776 participants from eight imaging centers with both MRI and fMRI images, as well as their clinical information. Another example is the Alzheimer's Disease Neuroimaging Initiative (ADNI, 2012) database, which includes 818 participants with MRI, fMRI and genomics data. The proposed tensor regression model aims to address the computational and modeling challenges of such large-scale imaging data. Meanwhile, our approach is equally applicable to smaller scale imaging data set, e.g., images acquired from a single lab with a moderate number of subjects. In this scenario, the regularization strategy outlined in Section 5 is expected to play a central role for scientific discovery.

The classical large n asymptotics in Section 4 may seem irrelevant for imaging data with a limited sample size. However, it outlines some basic properties of the proposed tensor regression model and has practical relevance in several aspects. For instance, by choosing a small rank such that the model size p_e is effectively smaller than n, we know that, under the specified conditions, the tensor regression model is consistently estimating the best rank-R approximation to the full model in the sense of Kullback-Liebler distance. A low rank estimate often provides a reasonable *approximation* to the true tensor regression parameter, even when the truth is of a high rank. This can be seen from our various numerical experiments, where a rank-3 model yields a good recovery of a butterfly shape in Figure 1 and a two-ball structure in Figure 3. Moreover, the regular asymptotics is useful for testing significance of a low rank sparse model in a replication study. Classical hypothesis testes such as likelihood ratio test can be formulated based on the asymptotic normality of the tensor estimates established in Section 4. The explicit formula for score and information in Section 4.1 and the identifiability issue discussed in Section 4.2 are not only useful for asymptotic theory but also for computation.

As the challenge of $p \gg n$ being a rule rather than exception in brain imaging analysis, regularization is to play a crucial role in practical applications. We consider the tensor regression, with or without regularization, presented in this paper an analog of the *hard thresholding* in classical regression, where the rank of the model is fixed. Respecting the array structure, the regularized tensor regression yields significantly better estimates than the classical regularization applied to the vectorized tensor covariates. Readers are referred to Appendix A.3 for a numerical comparison with the classical lasso. Currently, we are also investigating another line of regularization through "soft thresholding". That is, we estimate the tensor regression model *without* fixing the rank but instead subject to a convex regularization of the rank of the tensor parameter. Results along this line will be reported elsewhere.

The scale and complexity of neuroimaging data require the estimation algorithm to be highly scalable, efficient and stable. The methods in this paper are implemented in an efficient MATLAB toolbox. For instance, the median run time of fitting a rank-3 model to the 2D triangle shape in Figure 1 was about 5 seconds. Fitting a rank-3 logistic model to the 3D ADHD data in Section 6.2.3 took about 285 seconds for 10 runs from 10 random starting points, averaging < 30 seconds per run. Appendix A.4 contains further numerical results to study the algorithm stability with respect to starting values as well as computing time. All results were obtained on a standard laptop computer with a 2.6 GHz Intel i7 CPU.

We view the method of this article as a first step toward a more general area of array regression analysis, and the idea can be extended to a wide range of problems. We describe a few potential future directions here. First, although we only present results for models with a conventional covariate vector and an array covariate, the framework applies to arbitrary combination of array covariates. This provides a promising approach to the analysis of multi-modality data which becomes increasingly available in modern neuroimaging and medical studies. Second, we remark that our modeling approach and algorithm equally apply to many general loss functions occurring in classification and prediction. For example, for a binary response $Y \in \{0, 1\}$, the hinge loss takes the form

$$\sum_{i=1}^{n} [1 - y_i\{\alpha + \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{z}_i + \langle \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \cdots \circ \boldsymbol{\beta}_D^{(r)}, \boldsymbol{x}_i \rangle\}]_+$$

and should play an important role in support vector machines with array variates. Third, in this article rotation has not been explicitly considered in the modeling. When prior knowledge indicates, sometimes it is prudent to work in polar coordinates. For example, the 'disk' signal in Figure 1 can be effectively captured by a rank-1 outer product if the image is coded in polar coordinates. A diagonal signal array has full rank and cannot be approximated by any lower rank array, but if changed to polar coordinates, the rank reduces to one. Some of these extensions are currently under investigation. In summary, we believe that the proposed methodology timely answers calls in modern neuroimaging data analysis, whereas the general methodology of tensor regression is to play a useful role and also deserves more attention in statistical analysis of high-dimensional complex imaging data.

References

ADHD (2012), "The ADHD-200 sample," `http://fcon_1000.projects.nitrc.org/indi/adhd200/`, [Online; accessed 22-Jan-2012].

ADNI (2012), "Alzheimers Disease Neuroimaging Initiative," `http://adni.loni.ucla.edu`, [Online; accessed 22-Jan-2012].

Caffo, B., Crainiceanu, C., Verduzco, G., Joel, S., S.H., M., Bassett, S., and Pekar, J. (2010), "Two-stage decompositions for the analysis of functional connectivity for fMRI with application to Alzheimer's disease risk," *Neuroimage*, 51, 1140–1149.

Casey, B., Soliman, F., Bath, K. G., and Glatt, C. E. (2010), "Imaging genetics and development: Challenges and promises," *Human Brain Mapping*, 31, 838–851.

Davatzikos, C., Genc, A., Xu, D., and Resnick, S. (2001), "Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy." *NeuroImage*, 14, 1361–1369.

de Lathauwer, L. (2006), "A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization," *SIAM J. Matrix Analysis Applications*, 28, 642–666.

de Leeuw, J. (1994), "Block-relaxation algorithms in statistics," in *Information Systems and Data Analysis*, Springer, Berlin, pp. 308–325.

de Leeuw, J., Young, F., and Takane, Y. (1976), "Additive structure in qualitative data: an alternating least squares method with optimal scaling features," *Psychometrika*, 41, 471–503.

Fan, J. and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Statist. Assoc.*, 96, 1348–1360.

Frank, I. E. and Friedman, J. H. (1993), "A statistical view of some chemometrics regression tools," *Technometrics*, 35, 109–135.

Friston, K. J. (2009), "Modalities, modes, and models in functional neuroimaging," *Science*, 326, 399–403.

Hinrichs, C., Singh, V., Mukherjee, L., Xu, G., Chung, M. K., Johnson, S. C., and ADNI (2009), "Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset," *NeuroImage*, 48, 138–149.

Hung, H. and Wang, C.-C. (2011), “Matrix variate logistic regression analysis,” *arXiv:1105.2150v1.*

Kang, H., Ombao, H., Linkletter, C., Long, N., and Badre, D. (2012), “Spatio-spectral mixed effects model for functional magnetic resonance imaging data,” *Journal of American Statistical Association*, in press.

Kolda, T. G. (2006), “Multilinear operators for higher-order decompositions,” Tech. rep., Sandia National Laboratories.

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

Lange, K. (2004), *Optimization*, Springer Texts in Statistics, New York: Springer-Verlag.

— (2010), *Numerical Analysis for Statisticians*, Statistics and Computing, New York: Springer, 2nd ed.

Lazar, N. A. (2008), *The Statistical Analysis of Functional MRI Data*, New York: Springer.

Lehmann, E. L. and Romano, J. P. (2005), *Testing Statistical Hypotheses*, Springer Texts in Statistics, New York: Springer, 3rd ed.

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

Li, Y., Du, Y., and Lin, X. (2005), “Kernel-Based Multifactor Analysis for Image Synthesis and Recognition,” in *ICCV*, pp. 114–119.

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.

Lindquist, M. (2008), “The statistical analysis of fMRI data,” *Statistical Science*, 23, 439–464.

Liu, X. and Sidiropoulos, N. D. (2001), “Cramér-Rao lower bounds for low-rank decomposition of multidimensional arrays,” *IEEE Trans. on Signal Processing*, 49, 2074–2086.

Martino, F. D., Valente, G., Staeren, N., Ashburner, J., Goebel, R., and Formisano, E. (2008), “Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns,” *NeuroImage*, 43, 44–58.

McCullagh, P. and Nelder, J. A. (1983), *Generalized Linear Models*, Monographs on Statistics and Applied Probability, London: Chapman & Hall.

Ostrowski, A. M. (1960), *Solution of Equations and Systems of Equations*, Pure and Applied Mathematics, Vol. IX. Academic Press, New York-London.

Park, S. W. and Savvides, M. (2007), "Individual Kernel Tensor-Subspaces for Robust Face Recognition: A Computationally Efficient Tensor Framework Without Requiring Mode Factorization," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 37, 1156–1166.

Pollard, D. (1984), *Convergence of Stochastic Processes*, Springer Series in Statistics, New York: Springer-Verlag.

Polzehl, J., Voss, H. U., and Tabelow, K. (2010), "Structural adaptive segmentation for statistical parametric mapping," *NeuroImage*, 52, 515–523.

Qiu, P. (2005), *Image processing and jump regression analysis*, Wiley series in probability and statistics, John Wiley.

— (2007), "Jump surface estimation, edge detection, and image restoration," *Journal of the American Statistical Association*, 102, 745–756.

Rao, C. R. and Mitra, S. K. (1971), *Generalized Inverse of Matrices and its Applications*, John Wiley & Sons, Inc., New York-London-Sydney.

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

Rothenberg, T. J. (1971), "Identification in parametric models," *Econometrica*, 39, 577–91.

Ryali, S., Supekar, K., Abrams, D. A., and Menon, V. (2010), "Sparse logistic regression for whole-brain classification of fMRI data," *NeuroImage*, 51, 752–764.

Sidiropoulos, N. D. and Bro, R. (2000), "On the uniqueness of multilinear decomposition of N-way arrays," *Journal of Chemometrics*, 14, 229–239.

Sowell, E. R., Thompson, P. M., Welcome, S. E., Henkenius, A. L., Toga, A. W., and Peterson, B. S. (2003), "Cortical abnormalities in children and adolescents with attention-deficit hyperactivity disorder," *Lancet*, 362, 1699–1707.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. B*, 58, 267–288.

Valera, E. M., Faraone, S. V., Murray, K. E., and Seidman, L. J. (2007), "Meta-analysis of structural imaging findings in attention-deficit/hyperactivity disorder," *Biol Psychiatry*, 61, 1361–1369.

van der Vaart, A. and Wellner, J. (2000), *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer, corrected ed.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge: Cambridge University Press.

Worsley, K. J., Taylor, J. E., Tomaiuolo, F., and Lerch, J. (2004), "Unified univariate and multivariate random field theory," *NeuroImage*, 23, 189–195.

Yue, Y., Loh, J. M., and Lindquist, M. A. (2010), "Adaptive spatial smoothing of fMRI images." *Statistics and its Interface*, 3.

Zhou, H., Armagan, A., and Dunson, D. (2011), "Path following and empirical Bayes model selection for sparse regressions," *manuscript in preparation.*

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67, 301–320.

Supplementary Appendix

A1. 2D Shape Example with Varying Size and Strength

We report here the snapshots of estimates with varying signal strength at sample size $n = 500$ (Figure S.1), and the line plot of the average root mean squared error (RMSE) for estimation of $\boldsymbol{B}$ with both varying sample size and signal strength (Figure S.2). The findings were summarized in Section 6.1.

A2. 2D Shape Example with Regularization

We have run a numerical experiment to illustrate regularized tensor regression estimation. The setup is the same as that in Figure 1 except that the sample size is reduced to 500, which is only barely larger than the number of parameters $380 = 5+3\times(64+64)-9$ of a rank-3 tensor model. Figure S.3 shows the outcome of applying the lasso penalty to $\boldsymbol{B}_d$ in the rank-3 tensor regression model. Recovered signals at three different values of $\lambda = 0, 100, 1000$ are displayed. Without regularization ($\lambda = 0$), the rank-3 tensor regression is difficult to recover some signals such as triangle, disk and butterfly, mainly due to a very small sample size. On the other hand, excessive penalization compromises the quality of recovered signals too, as evidently in those shapes at $\lambda = 1000$. Regularized estimation with an appropriate amount of shrinkage improves estimation quality, as seen in triangle and disk at $\lambda = 100$ and in butterfly at $\lambda = 1000$. In practice the tuning parameter is chosen by certain model selection criterion such as BIC or cross validation. Moreover, we have experimented with the bridge and SCAD penalties for the same data and obtained similar results. The desirable unbiased (or nearly unbiased) estimates from these concave penalties are reflected by the improved contrast in the recovered signal. For the sake of space, we do not show those figures here.

A3. Comparison with Classical Lasso

We compare our regularized fixed rank tensor estimate with a classical regularized model, the lasso applied to vectorized image covariates. Our purpose is to investigate which method could provide a better estimate to the complicated true array signal with a

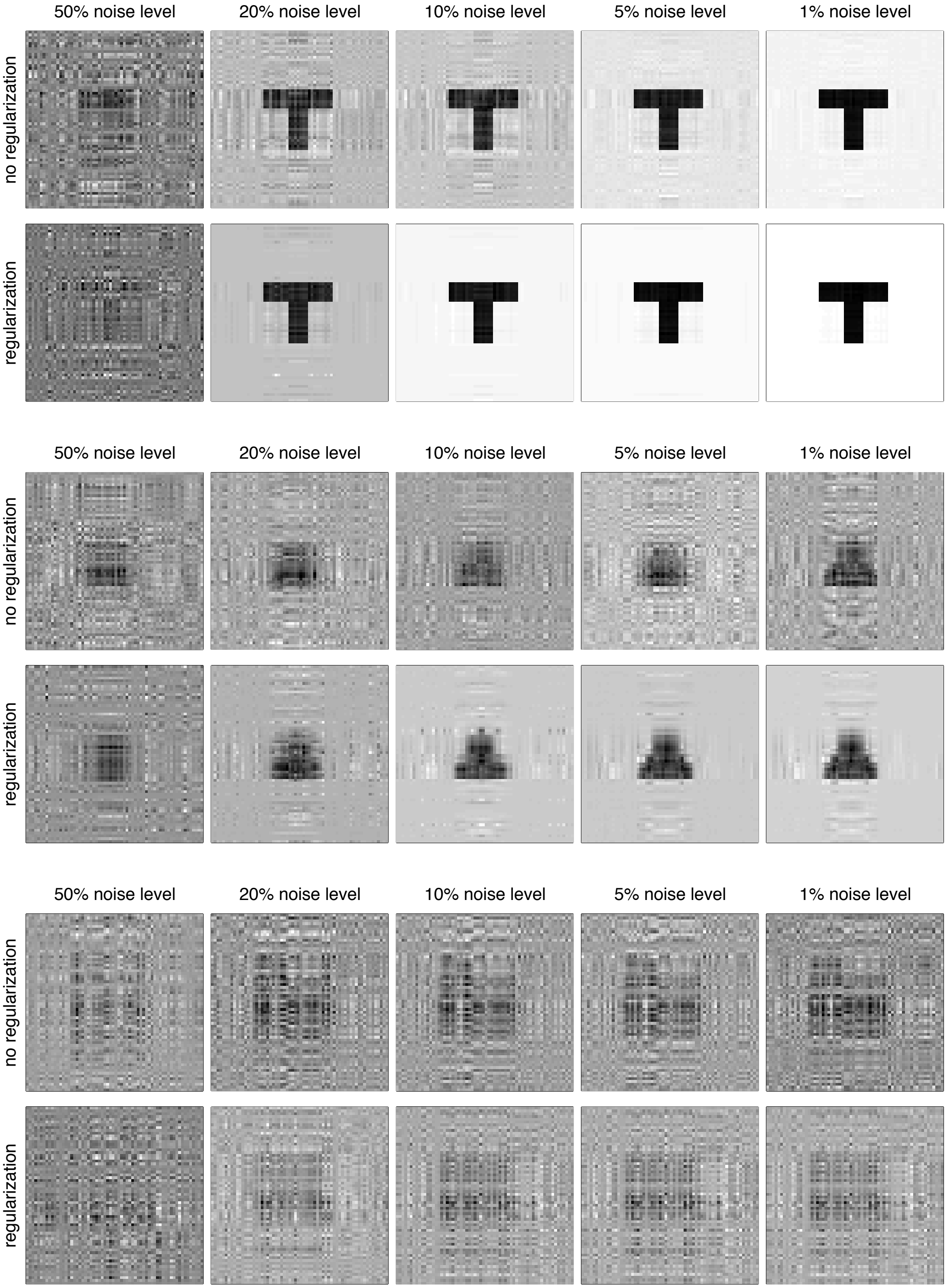


Figure S.1: Snapshots of tensor estimation with varying noise level. The matrix variate has size 64 by 64 with entries generated as independent standard normals. The regression coefficient for each entry is either 0 (white) or 1 (black).

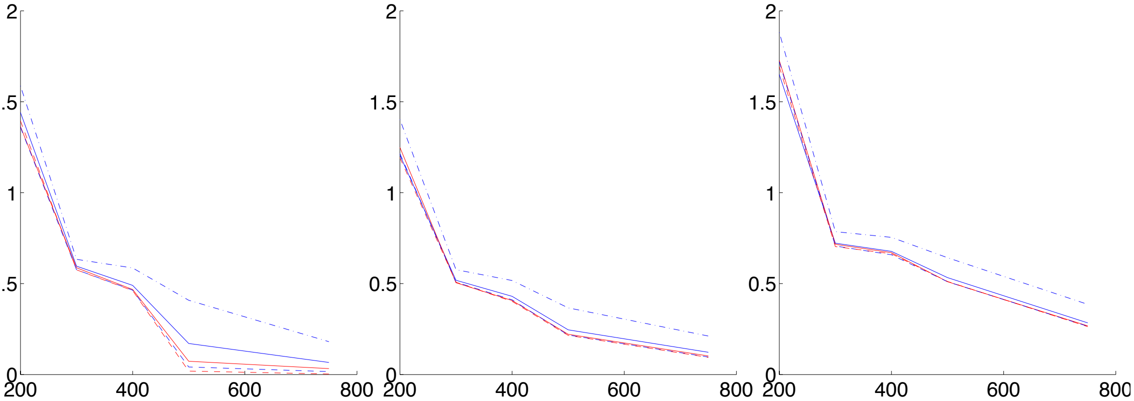


Figure S.2: Line plot of the average root mean squared error (RMSE, y-axis) for estimation of $\boldsymbol{B}$ with varying sample size (x-axis). Five lines denote the average RMSE when the noise level is 50% (blue dash-dot), 20% (blue solid), 10% (red solid), 5% (blue dash) and 1% (red dash) of the signal.

limited sample size. We reproduce the regularized tensor estimates of the "disk", "triangle", and "butterfly" signals in Figure S.3 with a rank 3 model. In addition, we also display the regular lasso estimates (i.e., lasso penalty applied to the vectorized matrix covariates) at the same sample size $n = 500$. The tuning parameter is chosen according to BIC. The results are shown in Figure S.4. It is clearly seen that the vector version of lasso estimates are far off from the truth whereas our tensor version estimates are much better.

A4. Algorithm Stability and Computing Time

We have carried out a numerical experiment to study the algorithm stability and the computing time. We report the results in Figure S.5. We adopt the setting of the illustrative example, using a "triangle" signal. Only one data instance was simulated with a fixed sample. Then the algorithm was initialized from 100 random starting points for tensor regression models at rank $r = 1, 2, 3, 4$. Box-plots of the final model deviances and wall clock run times are displayed in Figure S.5. All run times were recorded on a standard laptop computer with a 2.6 GHz Intel i7 CPU. As expected, higher rank models fit the data better, yielding smaller deviance, since the true signal is of a high rank. On the other hand, higher rank models are more vulnerable to local modes, as indicated

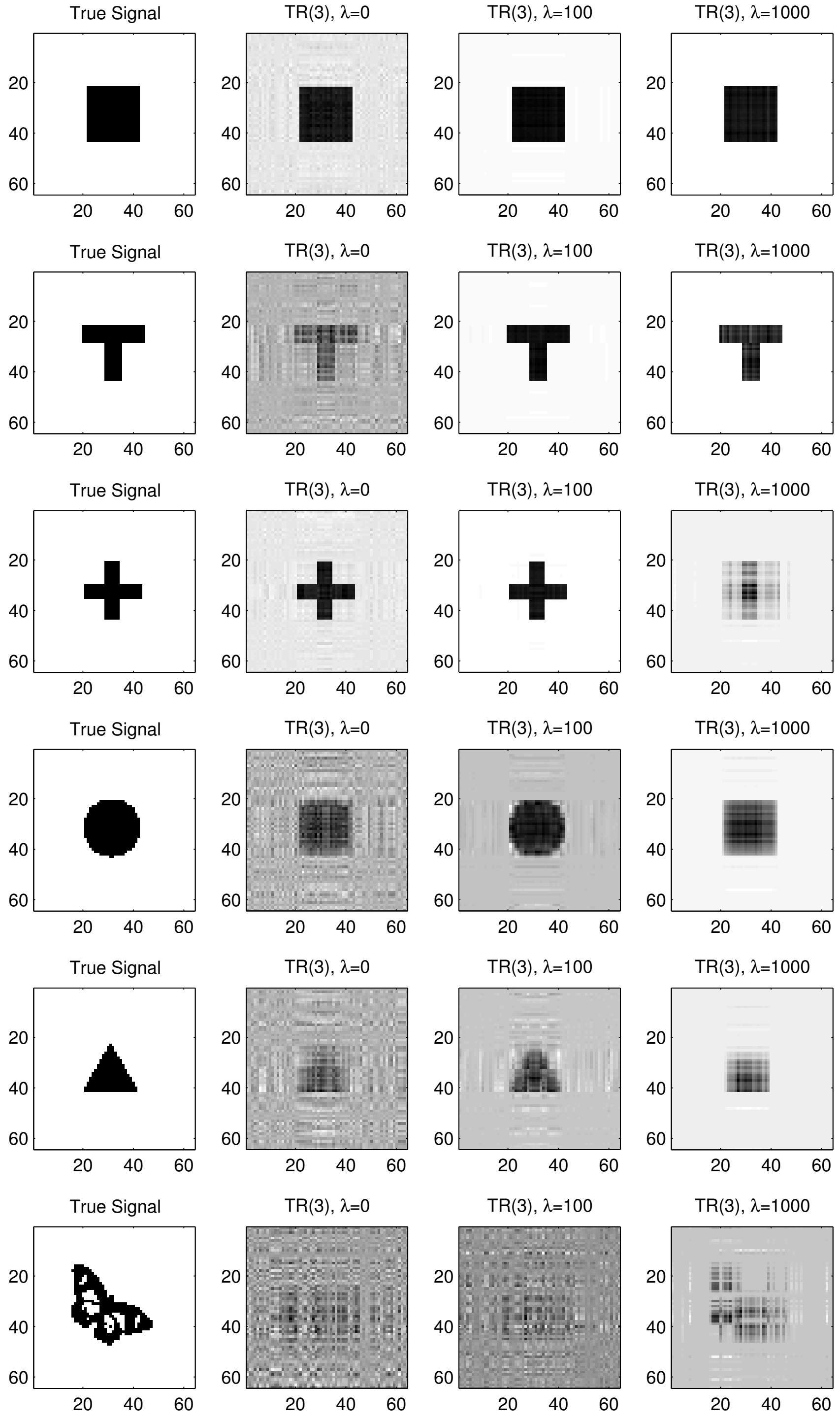


Figure S.3: Demonstration of lasso regularization. The matrix variate has size 64 by 64 with entries generated as independent standard normals. The regression coefficient for each entry is either 0 (white) or 1 (black). The sample size is 500.

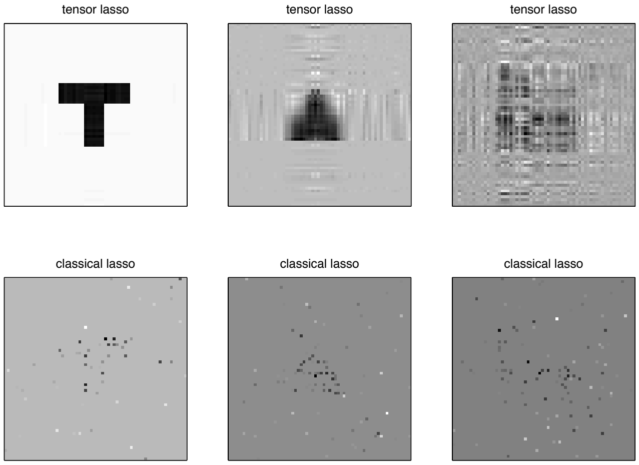


Figure S.4: Tensor lasso estimate (top) vs classical lasso estimate (bottom; applied on the vectorized matrix covariates) for T-shape, triangle and butterfly at sample size $n = 500$.

by larger variations, and takes longer to converge. The overall run time, however, is remarkably fast. For instance, the median run time of fitting a rank 3 model in this example is about 5 seconds. Fitting a rank-3 logistic model to the 3D ADHD data in Section 6.2.3 took about 285 seconds for 10 runs from 10 random starting points, averaging < 30 seconds per run.

A5. Proofs

Proof of Lemma 1

For the first identity it is enough to check that the mode-d matricization of $\boldsymbol{b}_1 \circ \cdots \circ \boldsymbol{b}_D$ is $\boldsymbol{b}_d(\boldsymbol{b}_D \otimes \cdots \otimes \boldsymbol{b}_{d+1} \otimes \boldsymbol{b}_{d-1} \otimes \cdots \otimes \boldsymbol{b}_1)^{\mathsf{T}}$, which is easily seen to hold elementwise. The scalar product $\prod_{d' \neq d} b_{d' i_{d'}}$ appears as the j-th element of the row vector $(\boldsymbol{b}_D \otimes \cdots \otimes \boldsymbol{b}_{d+1} \otimes \boldsymbol{b}_{d-1} \otimes \cdots \otimes \boldsymbol{b}_1)^{\mathsf{T}}$ where $j = 1 + \sum_{d' \neq d}(i_{d'} - 1) \prod_{d'' < d', d'' \neq d} p_{d''}$. The matricization of a sum of arrays equals sum of their matricizations. Therefore the first identity holds.

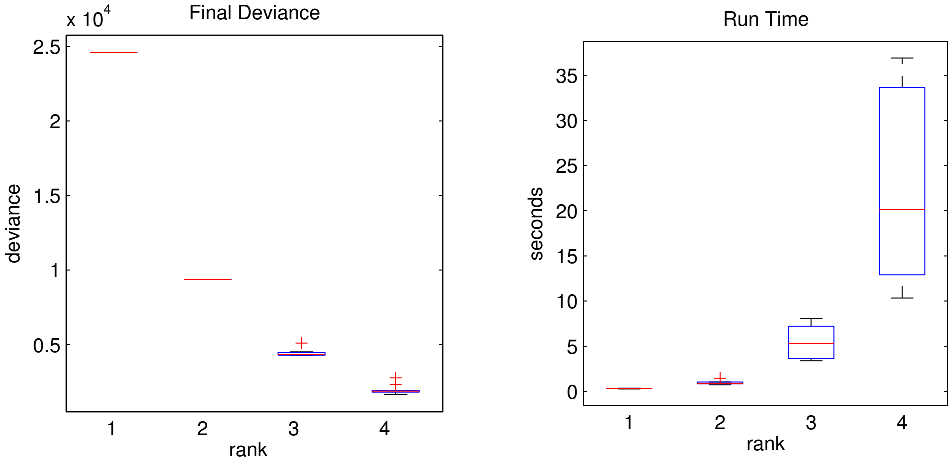


Figure S.5: Algorithm stability and run time. The same algorithm is initialized from 100 random starting points to fit tensor regression models at rank $r = 1, 2, 3, 4$ respectively. The final model deviances and wall clock timings are reported.

For the second identity,

$$\begin{aligned}\operatorname{vec} \boldsymbol{B} &= \operatorname{vec}(\sum_{r=1}^{R} \boldsymbol{b}_1^{(r)} \circ \cdots \circ \boldsymbol{b}_D^{(r)}) = \sum_{r=1}^{R} \operatorname{vec}(\boldsymbol{b}_1^{(r)} \circ \cdots \circ \boldsymbol{b}_D^{(r)}) \\ &= \sum_{r=1}^{R} \boldsymbol{b}_D^{(r)} \otimes \cdots \otimes \boldsymbol{b}_1^{(r)} = (\boldsymbol{B}_D \odot \cdots \odot \boldsymbol{B}_1)\mathbf{1}_R.\end{aligned}$$

Proof of Proposition 1

Proof of global convergence follows from standard arguments for algorithms that monotonically increase objective function (de Leeuw, 1994; Lange, 2004, 2010). Under the blockwise strict concavity condition (2), the block update is well-defined and differentiable. Then algorithmic map M is a composition of $D+1$ differentiable maps and, by implicit function theorem, continuous. Let $\boldsymbol{\theta}^{(t)}$ be the sequence generated by M and $\boldsymbol{\theta}$ be any accumulation point of $\boldsymbol{\theta}^{(t)}$. Since the algorithm always increase objective value, $\ell(M(\boldsymbol{\theta}^{(t)})) \geq \ell(\boldsymbol{\theta}^{(t)})$. Taking limit gives $\ell(M(\boldsymbol{\theta})) = \ell(\boldsymbol{\theta})$ by continuity of M and ℓ. Thus any accumulation point of algorithmic sequence is a stationary point of ℓ. The set of accumulation points is contained in $\{\boldsymbol{\theta} : \ell(\boldsymbol{\theta}) \geq \ell(\boldsymbol{\theta}^{(0)})\}$ and thus compact by condition (1). Compactness implies that this set of accumulation points is also connected (Lange, 2010, Propitions 8.2.1 and 15.4.2). Discreteness of the stationary points of ℓ implies that the number of stationary points is finite. Otherwise there is a sequence of stationary points whose limit is not isolated. Finally the set of accumulation points is a connected subset of these finite number of stationary points, thus is a single point. In other words the algorithmic sequence $\boldsymbol{\theta}^{(t)}$ converges to a stationary point of ℓ.

Proof of local convergence relies on the Ostrowski's theorem (Ostrowski, 1960), which states that the sequence $\boldsymbol{\theta}^{t+1} = M(\boldsymbol{\theta}^t)$ is locally attracted to $\boldsymbol{\theta}^\infty$ if the spectral radius of the differential of the algorithmic map $\rho[dM(\boldsymbol{\theta}^\infty)]$ is strictly less than 1. We partition the Hessian of the objective function ℓ at $\boldsymbol{\theta}^\infty$ as

$$d^2\ell(\alpha, \boldsymbol{\gamma}, \boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) = \begin{pmatrix} d_{00}^2\ell & & \mathbf{0} & \\ & d_{11}^2\ell & \cdots & d_{1D}^2\ell \\ \mathbf{0} & \vdots & \ddots & \vdots \\ & d_{D1}^2\ell & \cdots & d_{DD}^2\ell \end{pmatrix} = \boldsymbol{L} + \boldsymbol{D} + \boldsymbol{L}^\intercal,$$

where $\boldsymbol{L}$ is the strictly block lower triangular part and $\boldsymbol{D}$ is the block diagonal part. Then it can be shown that the differential of map M is

$$dM(\boldsymbol{\theta}^\infty) = -(\boldsymbol{L} + \boldsymbol{D})^{-1}\boldsymbol{L}^\intercal.$$

Note $\boldsymbol{\theta}^\infty$ being a strict local maximum implies that $d^2\ell(\boldsymbol{\theta}^\infty)$ is strictly negative definite and thus the diagonal blocks $d_{dd}^2\ell$, $d = 0, \ldots, D$ are strictly negative definite too. Therefore the block lower triangular matrix $(\boldsymbol{L} + \boldsymbol{D})$ is invertible as it shares the same eigenvalues as its diagonal blocks. The spectral radius of $-(\boldsymbol{L} + \boldsymbol{D})^{-1}\boldsymbol{L}^\intercal$ is strictly less than one. Therefore the iterates are locally attracted to $\boldsymbol{\theta}^\infty$.

Proof of Lemma 2

By Lemma 1,

$$\boldsymbol{B}_{(d)} = \boldsymbol{B}_d(\boldsymbol{B}_D \odot \cdots \odot \boldsymbol{B}_{d+1} \odot \boldsymbol{B}_{d-1} \odot \cdots \odot \boldsymbol{B}_1)^\intercal.$$

Using the well-known fact that $\text{vec}(\boldsymbol{XYZ}) = (\boldsymbol{Z}^\intercal \otimes \boldsymbol{X})\text{vec}(\boldsymbol{Y})$,

$$\text{vec}\boldsymbol{B}_{(d)} = [(\boldsymbol{B}_D \odot \cdots \odot \boldsymbol{B}_{d+1} \odot \boldsymbol{B}_{d-1} \odot \cdots \odot \boldsymbol{B}_1) \otimes \boldsymbol{I}_{p_d}]\text{vec}(\boldsymbol{B}_d).$$

Thus we have

$$\begin{aligned}\boldsymbol{J}_d &= D\boldsymbol{B}(\boldsymbol{B}_d)\\ &= D\boldsymbol{B}(\boldsymbol{B}_{(d)}) \cdot D\boldsymbol{B}_{(d)}(\boldsymbol{B}_d)\\ &= \boldsymbol{\Pi}_d \frac{\partial \text{vec} \boldsymbol{B}_{(d)}}{\partial (\text{vec} \boldsymbol{B}_d)^{\mathsf{T}}}\\ &= \boldsymbol{\Pi}_d[(\boldsymbol{B}_D \odot \cdots \odot \boldsymbol{B}_{d+1} \odot \boldsymbol{B}_{d-1} \odot \cdots \odot \boldsymbol{B}_1) \otimes \boldsymbol{I}_{p_n}].\end{aligned}$$

Combining gives

$$\begin{aligned}&D\eta(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)\\ &= D\eta(\boldsymbol{B}) \cdot D\boldsymbol{B}(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)\\ &= (\text{vec}\boldsymbol{X})^{\mathsf{T}}[\boldsymbol{J}_1 \ \boldsymbol{J}_2 \ \cdots \ \boldsymbol{J}_D].\end{aligned}$$

For the Hessian,

$$h_{(i_d,r),(i_{d'},r')} = \sum_{j_1,\ldots,j_D} x_{j_1,\ldots,j_D} \frac{\partial^2 b_{j_1,\ldots,j_D}}{\partial \beta_{i_d}^{(r)} \partial \beta_{i_{d'}}^{(r')}}.$$

The second derivative in the summand is nonzero only if $j_d = i_d$, $j_{d'} = i_{d'}$, $r = r'$, and $d \neq d'$. Therefore

$$h_{(i_d,r),(i_{d'},r')} = 1_{\{r=r',d\neq d'\}} \sum_{j_d=i_d, j_{d'}=i_{d'}} x_{j_1,\ldots,j_D} \prod_{d''\neq d,d'} \beta_{j_{d''}}^{(r)},$$

where the sum is over $\prod_{d''\neq d,d'} p_{d''}$ terms. It is easy to see that $h_{(i_d,r),(i_{d'},r')}$ are the entries of the matrix

$$\boldsymbol{X}_{(dd')}(\boldsymbol{B}_D \odot \cdots \odot \boldsymbol{B}_{d+1} \odot \boldsymbol{B}_{d-1} \odot \cdots \odot \boldsymbol{B}_{d'+1} \odot \boldsymbol{B}_{d'-1} \odot \cdots \odot \boldsymbol{B}_1).$$

Proof of Proposition 2

Since $\mu = b'(\theta)$, $d\mu/d\theta = b''(\theta) = \sigma^2/a(\phi)$ and

$$\begin{aligned}\nabla \ell(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) &= \frac{y - b'(\theta)}{a(\phi)} \frac{d\theta}{d\mu} \frac{d\mu}{d\eta} \nabla \eta(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)\\ &= \frac{(y-\mu)\mu'(\eta)}{\sigma^2} [\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]^{\mathsf{T}} (\text{vec}\boldsymbol{X})\end{aligned}$$

by Lemma 2. Further differentiating shows

$$
\begin{aligned}
& d^2\ell(\boldsymbol{B}_1,\ldots,\boldsymbol{B}_D) \\
= & -\frac{1}{\sigma^2}\nabla\mu(\boldsymbol{B}_1,\ldots,\boldsymbol{B}_D)d\mu_i(\boldsymbol{B}_1,\ldots,\boldsymbol{B}_D)+\frac{y-\mu}{\sigma^2}d^2\mu(\boldsymbol{B}_1,\ldots,\boldsymbol{B}_D) \\
= & -\frac{[\mu'(\eta)]^2}{\sigma^2}([\boldsymbol{J}_1\ldots\boldsymbol{J}_D]^{\mathsf{T}}\mathrm{vec}\boldsymbol{X})([\boldsymbol{J}_1\ldots\boldsymbol{J}_D]^{\mathsf{T}}\mathrm{vec}\boldsymbol{X})^{\mathsf{T}} \\
& +\frac{(y-\mu)\theta''(\eta)}{\sigma^2}([\boldsymbol{J}_1\ldots\boldsymbol{J}_D]^{\mathsf{T}}\mathrm{vec}\boldsymbol{X})([\boldsymbol{J}_1\ldots\boldsymbol{J}_D]^{\mathsf{T}}\mathrm{vec}\boldsymbol{X})^{\mathsf{T}} \\
& +\frac{(y-\mu)\theta'(\eta)}{\sigma^2}d^2\eta(\boldsymbol{B}).
\end{aligned}
$$

It is easy to see that $\mathbf{E}[\nabla\ell(\boldsymbol{B}_1,\ldots,\boldsymbol{B}_D)]=\mathbf{0}$ and $\mathbf{E}[-d^2\ell(\boldsymbol{B}_1,\ldots,\boldsymbol{B}_D)]=\boldsymbol{I}(\boldsymbol{B}_1,\ldots,\boldsymbol{B}_D)$, thus (9) follows.

Proof of Proposition 4

The following useful result relates local identifiability of parametric models to their Fisher information matrix.

Lemma 4. *(Rothenberg, 1971, Theorem 1) Let θ_0 be a regular point of the information matrix $I(\theta)$. Then θ_0 is locally identifiable if and only if $I(\theta_0)$ is nonsingular.*

The regularity assumptions for Lemma 4 are satisfied by tensor model: (1) the parameter space $\mathcal{B}$ is open, (2) the density $p(y,\boldsymbol{x}|\boldsymbol{B})$ is proper for all $\boldsymbol{B}\in\mathcal{B}$, (3) the support of the density $p(y,\boldsymbol{x}|\boldsymbol{B})$ is same for all $\boldsymbol{B}\in\mathcal{B}$, (4) the log density $\ell(\boldsymbol{B}|y,\boldsymbol{x})=\ln p(y,\boldsymbol{x}|\boldsymbol{B})$ is continuously differentiable, and (5) the information matrix

$$
\boldsymbol{I}(\boldsymbol{B})=[\boldsymbol{J}_1\ldots\boldsymbol{J}_D]^{\mathsf{T}}\left[\sum_{i=1}^{n}\frac{\mu'(\eta_i)^2}{\sigma_i^2}(\mathrm{vec}\,\boldsymbol{x}_i)(\mathrm{vec}\,\boldsymbol{x}_i)^{\mathsf{T}}\right][\boldsymbol{J}_1\ldots\boldsymbol{J}_D]
$$

is continuous in $\boldsymbol{B}$ by Proposition 2. Then, by Lemma 4, $\boldsymbol{B}$ is locally identifiable if and only if $\boldsymbol{I}(\boldsymbol{B})$ is nonsingular.

Proof of Theorem 1

It suffices to show the consistency of the estimated factor matrix $\hat{\boldsymbol{B}}_{nd}$, $d=1,\ldots,D$, which implies the consistency of the tensor estimate $\hat{\boldsymbol{B}}_n=[\![\hat{\boldsymbol{B}}_{n1},\ldots,\hat{\boldsymbol{B}}_{nD}]\!]$ by continuous mapping theorem. The following well-known theorem is our major tool for establishing consistency.

Lemma 5. *(van der Vaart, 1998, Theorem 5.7) Let M_n be random functions and let M be a fixed function of θ such that*

$$\sum_{\theta:d(\theta,\theta_0)\geq\epsilon} M(\theta) < M(\theta_0)$$

for every $\epsilon > 0$ and

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \to 0 \text{ in probability.}$$

Then any sequence of estimators $\hat{\theta}_n$ with $M_n(\hat{\theta}) \geq M_n(\theta_0) - o_P(1)$ converges in probability to θ_0.

To apply Lemma 5 in our setting, we take the nonrandom function M to be $\boldsymbol{B} \mapsto \mathbb{P}_{\boldsymbol{B}_0}[\ell(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{B})]$ (or its modifications) and the sequence of random functions to be $M_n : \boldsymbol{B} \mapsto \frac{1}{n}\sum_{i=1}^n \ell(y_i, \boldsymbol{x}_i|\boldsymbol{B}) = \mathbb{P}_n M$, where $\mathbb{P}_n$ denotes the empirical measure under $\boldsymbol{B}_0$. Then M_n converges to M a.s. by strong law of large number. The first condition requires that $\boldsymbol{B}_0$ is a well-separated maximum of M. This is guaranteed by the (global) identifiability of $\boldsymbol{B}_0$ and information inequality. The second uniform convergence condition is most convenient to be verified by the Glivenko-Cantelli theory (Pollard, 1984; van der Vaart, 1998; van der Vaart and Wellner, 2000).

For binary regression, the following proof is an expansion of (van der Vaart, 1998, Example 5.40) to tensor binary regression. The density is $p_{\boldsymbol{B}}(y|\boldsymbol{x}) = y\pi(\boldsymbol{B}, \boldsymbol{x}) + (1 - y)(1 - \pi(\boldsymbol{B}, \boldsymbol{x}))$, where $\pi(\boldsymbol{B}, \boldsymbol{x}) = g^{-1}(\langle \boldsymbol{B}, \boldsymbol{x}\rangle)$, where g is the link function. For instance, $\pi(\boldsymbol{B}, \boldsymbol{x}) = 1/(1 + e^{-\langle \boldsymbol{B}, \boldsymbol{x}\rangle})$ corresponds to the logit link and $\pi(\boldsymbol{B}, \boldsymbol{x}) = \Phi(\langle \boldsymbol{B}, \boldsymbol{x}\rangle)$ the probit link. Take $m_{\boldsymbol{B}} = \ln[(p_{\boldsymbol{B}} + p_{\boldsymbol{B}_0})/2]$. First we show that $\boldsymbol{B}_0$ is a well-separated maximum of the function $M(\boldsymbol{B}) := \mathbb{P}_{\boldsymbol{B}_0} m_{\boldsymbol{B}}$. The global identifiability of $\boldsymbol{B}_0$ and information inequality guarantee that $\boldsymbol{B}_0$ is the unique maximum of M. To show that it is a well-separated maximum, we need to verify that $M(\boldsymbol{B}_k) \to M(\boldsymbol{B}_0)$ implies $\boldsymbol{B}_k \to \boldsymbol{B}_0$. Suppose $M(\boldsymbol{B}_k) \to M(\boldsymbol{B}_0)$, then $\langle \boldsymbol{B}_k, \boldsymbol{X}\rangle \to \langle \boldsymbol{B}_0, \boldsymbol{X}\rangle$ in probability. If $\boldsymbol{B}_k$ are bounded, then $\mathbf{E}[\langle \boldsymbol{B}_k - \boldsymbol{B}_0, \boldsymbol{X}\rangle^2] \to 0$ and $\boldsymbol{B}_k \to \boldsymbol{B}_0$ by nonsingularity of $\mathbf{E}[(\text{vec}\boldsymbol{X})(\text{vec}\boldsymbol{X})^{\mathsf{T}}]$. On the other hand, $\boldsymbol{B}_k$ cannot escape to infinity. If they do, then $\langle \boldsymbol{B}_k, \boldsymbol{X}\rangle/\|\boldsymbol{B}_k\| \to 0$ in probability which in turn implies that $\boldsymbol{B}_k/\|\boldsymbol{B}_k\| \to \boldsymbol{0}$. For the uniform convergence, we see that the class of functions $\{\langle \boldsymbol{B}, \boldsymbol{X}\rangle, \boldsymbol{B} \in \mathcal{B}\}$ form a Vapnik-Červonenkis (VC) class. This is true because it is a collection of finite number of polynomials of degree D and then apply the VC vector space argument (van der Vaart and Wellner, 2000, 2.6.15). This implies that $\{\pi(\langle \boldsymbol{B}, \boldsymbol{X}\rangle), \boldsymbol{B} \in \mathcal{B}\}$ is a VC class since π is a monotone function (van der Vaart and Wellner, 2000, 2.16.18). Now $m_{\boldsymbol{B}}$ is Lipschitz in π and π_0 since

$$\frac{\partial m_{\boldsymbol{B}}}{\partial \pi} = \frac{\partial m_{\boldsymbol{B}}}{\partial \pi_0} = \frac{2y-1}{y\pi + (1-y)(1-\pi) + y\pi_0 + (1-y)(1-\pi_0)} \leq \frac{1}{\pi_0} + \frac{1}{1-\pi_0}.$$

A Lipschitz composition of a Donsker class is still a Donsker class (van der Vaart, 1998, 19.20). Therefore $\{\boldsymbol{B} \mapsto m_{\boldsymbol{B}}\}$ is a bounded Donsker class with the trivial envelope function 1. A Donsker class is certainly a Glivenko-Cantelli class. Finally the Glivenko-Cantelli theorem establishes the uniform convergence condition required by Lemma 5.

When the parameter is restricted to a compact set, $\mu = g^{-1}(\langle \boldsymbol{B}, \boldsymbol{x}\rangle)$ is confined in a bounded interval and the log-likelihood ℓ is Lipschitz on the finite interval. If follows that $\{\ell(\boldsymbol{B}) = \ell \circ g^{-1} \circ \langle \boldsymbol{B}, \boldsymbol{X}\rangle, \boldsymbol{B} \in \mathcal{B}\}$ is a Donsker class as composition with a monotone or Lipschitz function preserves the Donsker class. Therefore the Glivenko-Cantelli theorem establish the uniform convergence. Compactness of parameter space implies that $\boldsymbol{B}_0$ is a well-separated maximum if it is the unique maximizer of $M(\boldsymbol{B}) = \mathbb{P}_{\boldsymbol{B}_0} m_{\boldsymbol{B}}$ (van der Vaart, 1998, Exercise 5.27). Uniqueness is guaranteed by the information inequality whenever $\boldsymbol{B}_0$ is identifiable. This verifies the consistency for normal and Poisson regressions.

Proof of Lemma 3

By a well-known result (Lehmann and Romano, 2005, Theorem 12.2.2) or (van der Vaart, 1998, Lemma 7.6), it suffices to verify that the density is continuously differentiable in parameter for μ-almost all x and that the Fisher information matrix exists and is continuous. The derivative of density is

$$\nabla p(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) = \nabla e^{\ell(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)} = p(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) \nabla \ell(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D),$$

which is well-defined and continuous by Proposition 2. The same proposition shows that the information matrix exists and is continuous. Therefore the tensor regression model is q.m.d.

Proof of Theorem 2

The following result relates asymptotic normality to the densities that satisfy q.m.d.

Lemma 6. *(van der Vaart, 1998, Theorem 5.39) Suppose that the model* $(P_\theta : \theta \in \Theta)$ *is q.m.d. at an inner point* θ_0 *of* $\Theta \subset \mathbb{R}^k$*. Furthermore, suppose that there exists a measurable function* $\dot{\ell}$ *with* $\mathbf{P}_{\theta_0}\dot{\ell}^2 < \infty$ *such that, for every* θ_1 *and* θ_2 *in a neighborhood of* θ_0*,*

$$|\ln p_{\theta_1}(x) - \ln p_{\theta_2}(x)| \leq \dot{\ell}(x)\|\theta_1 - \theta_2\|.$$

If the Fisher information matrix I_{θ_0} *is nonsingular and* $\hat{\theta}_n$ *is consistent, then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = I_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{\ell}_{\theta_0}(X_i) + o_{P_{\theta_0}}(1). \tag{10}$$

In particular, the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ *is asymptotically normal with mean zero and covariance matrix* $I_{\theta_0}^{-1}$*.*

Lemma 3 shows that tensor regression model is q.m.d. By Proposition 2 and chain rule, the score function

$$\dot{\ell}_{\boldsymbol{B}}(y, \boldsymbol{x}) = d\ell(\boldsymbol{B}) = \frac{(y-\mu)\mu'(\eta)}{\sigma^2}(\operatorname{vec}\boldsymbol{x})^{\mathsf{T}}[\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]$$

is uniformly bounded in y, $\boldsymbol{x}$, and $\boldsymbol{B}$ ranging over compacta and continuous in $\boldsymbol{B}$ for every y and $\boldsymbol{x}$. For sufficiently small neighborhood U of $\boldsymbol{B}_0$, $\sup_U \|\dot{\ell}_{\boldsymbol{B}}\|$ is square-integrable. Thus the local Lipschitz condition is satisfied and Lemma 6 applies.