Tensor Regression with Applications in Neuroimaging Data Analysis

Hua ZHOU, Lexin LI, and Hongtu ZHU

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. Supplementary materials for this article are available online.

KEY WORDS: Brain imaging; Dimension reduction; Generalized linear model; Magnetic Resonance Imaging; Multidimensional array; Tensor regression.

1. INTRODUCTION

Understanding the inner workings of 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, and positron emission tomography (PET), among others. The sheer size and complexity of medical imaging data, however, pose unprecedented challenge to classical statistical methods and have received increasing interest in recent years (Lindquist 2008; Lazar 2008; Martino et al. 2008; Friston 2009; Hinrichs et al. 2009; Ryali et al. 2010; 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 (Worsley et al. 2004; Lazar 2008). A major drawback is that all voxels are treated as independent units and important spatially correlation is ignored (Polzehl, Voss, and Tabelow 2010; Yue, Loh, and Lindquist 2010; Li et al. 2011). 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 three-dimensional and higher dimensional images, however, are far from trivial and require 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 *multidimensional array*, as covariates. Most classical regression methods take vectors as covariates. Naively turning an image array into a vector is evidently unsatisfactory. 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 severely compromised by this ultrahigh dimensionality. More seriously, vectorizing an array destroys the inherent spatial structure of the image that possesses wealth of information.

Our new regression method effectively exploits the array structure in imaging data, and substantially reduces the dimensionality, which in turn leads 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, for example, 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 (MLE), as well as statistical inferential tools. Regularized tensor regression is also investigated to identify regions of interest in brains that are relevant to a clinical 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

trends 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 multimodality imaging analysis, multiphenotype analysis, and imaging genetics (Friston 2009; Casey et al. 2010), all of 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, for example, 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 MLE. 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 supplementary material.

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 article (Kolda and Bader 2009). In this article, we use the terms multidimensional array and tensor interchangeably.

Given two matrices $A = [a_1 \ldots a_n] \in \mathbb{R}^{m \times n}$ and $B = [b_1 \ldots b_q] \in \mathbb{R}^{p \times q}$, the *Kronecker product* is the mp-by-nq matrix $A \otimes B = [a_1 \otimes B \; a_1 \otimes B \; \ldots \; a_n \otimes B]$. If A and 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 $A \odot B = [a_1 \otimes b_1 \; a_2 \otimes b_2 \; \ldots \; a_n \otimes b_n]$. If $n = q = 1$, then $A \odot B = A \otimes B$. Some useful operations transform a tensor into a matrix/vector. The vec(B) *operator* stacks the entries of a D-dimensional tensor $B \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ into a column vector. Specifically, an entry $b_{i_1 \ldots i_D}$ maps to the jth entry of vec 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*, $B_{(d)}$, maps a tensor 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 B maps to the (i_d, j) element of the matrix $B_{(d)}$, 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 B is the same as vectorizing the mode-1 matricization $B_{(1)}$. The *mode-(d, d') matricization* $B_{(dd')} \in \mathbb{R}^{p_d p_{d'} \times \prod_{d'' \neq d, d'} p_{d''}}$ is defined in a similar fashion (Kolda 2006). We also introduce an operator that turns vectors into an array. Specifically,

an *outer product*, $b_1 \circ b_2 \circ \cdots \circ b_D$, of D vectors $b_d \in \mathbb{R}^{p_d}$ is a $p_1 \times \cdots \times p_D$ array with entries $(b_1 \circ b_2 \circ \cdots \circ b_D)_{i_1 \cdots i_D} = \prod_{d=1}^{D} b_{d i_d}$.

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

$$B = \sum_{r=1}^{R} \beta_1^{(r)} \circ \cdots \circ \beta_D^{(r)}, \tag{1}$$

where $\beta_d^{(r)} \in \mathbb{R}^{p_d}, d = 1, \ldots, D, r = 1, \ldots, R$ are all column vectors, and B cannot be written as a sum of less than R outer products. For convenience, the decomposition is often represented by a shorthand, $B = [\![B_1, \ldots, B_D]\!]$, where $B_d = [\beta_d^{(1)}, \ldots, \beta_d^{(R)}] \in \mathbb{R}^{p_d \times R}$ (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 $B \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ admits a rank-R decomposition (1), then

$$B_{(d)} = B_d (B_D \odot \cdots \odot B_{d+1} \odot B_{d-1} \odot \cdots \odot B_1)^\mathsf{T} \text{ and}$$
vec $B = (B_D \odot \cdots \odot B_1) \mathbf{1}_R$.

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

2.2 Motivation and Basic Model

To motivate our model, we first start with a vector-valued X and absorb Z into 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 $X \in \mathbb{R}^p$ to the mean $\mu = E(Y|X)$ via $g(\mu) = \eta = \alpha + \beta^\mathsf{T} X$, where $g(\cdot)$ is a strictly increasing link function, and η denotes the linear systematic part with intercept α and the coefficient vector $\beta \in \mathbb{R}^p$.

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

$$g(\mu) = \alpha + \beta_1^\mathsf{T} X \beta_2,$$

where $\beta_1 \in \mathbb{R}^{p_1}$ and $\beta_2 \in \mathbb{R}^{p_2}$, respectively. The bilinear form $\beta_1^\mathsf{T} X \beta_2$ is a natural extension of the linear term $\beta^\mathsf{T} X$ in the classical GLM with a vector covariate X. It is interesting to note that, this bilinear form was first proposed by Li, Kim, and Altman (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 $\beta_1^\mathsf{T} X \beta_2 = (\beta_2 \otimes \beta_1)^\mathsf{T} \text{vec}(X)$.

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

$$g(\mu) = \alpha + \gamma^\top Z + (\beta_D \otimes \cdots \otimes \beta_1)^\top \text{vec}(X), \qquad (3)$$

where $\gamma \in \mathbb{R}^{p_0}$ and $\beta_d \in \mathbb{R}^{p_d}$ 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 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 $B \in \mathbb{R}^{64 \times 64}$ (shown in the first column), along with the estimated images by model (3) (in the second column labeled by TR(1)). Specifically, we simulated 1000 univariate responses y_i according to a normal model with mean $\mu_i = \gamma^\top z_i + \langle B, x_i \rangle$, where $\gamma = \mathbf{1}_5$. The inner product between two arrays is defined as $\langle B, X \rangle = \langle \text{vec} B, \text{vec} X \rangle = \sum_{i_1, \ldots, i_D} \beta_{i_1 \ldots i_D} x_{i_1 \ldots i_D}$. The coefficient array B is binary, with the true signal region equal to one and the rest zero. The regular covariate z_i and image covariate 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 B using data (y_i, z_i, 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 X follow the same distribution. The signal region is defined through the coefficient image B and needs to be inferred from the association between Y and X after adjusting for Z. Second, the classical GLM is difficult to apply in this example if we simply treat $\text{vec}(X)$ as a covariate vector, since the sample size $n = 1000$ 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 type region (parallel to the image edges). On the other hand, since the parameter vector β_d in a rank-1 model is only able to capture the accumulative signal along the dth dimension of the array variate X, it is unsurprising that it does not perform well for signals that are far away from "square," 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 $X \in \mathbb{R}^{p_1 \times \cdots \times p_D}$, and a full coefficient array B of same size that captures the effects of each array element. Then the most flexible GLM suggests a linear systematic part

$$g(\mu) = \alpha + \gamma^\top Z + \langle B, X \rangle.$$

The issue with this model is that B has the same number of parameters, $\prod_{d=1}^D p_d$, as X, which is ultrahigh dimensional and far exceeds the usual sample size. Then a natural idea is to approximate B with less parameters. If B admits a rank-1 decomposition (1), that is, $B = \beta_1 \circ \beta_2 \circ \cdots \circ \beta_D$, where $\beta_d \in \mathbb{R}^{p_d}$, then by Lemma 1, we have

$$\begin{aligned} \text{vec } B &= \text{vec}(\beta_1 \circ \beta_2 \circ \cdots \circ \beta_D) \\ &= \beta_D \odot \cdots \odot \beta_1 = \beta_D \otimes \cdots \otimes \beta_1. \end{aligned}$$

In other words, model (3) is indeed a *data-driven* model with a rank-1 approximation to the general signal array 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

$$\begin{aligned} g(\mu) &= \alpha + \gamma^\top Z + \left\langle \sum_{r=1}^R \beta_1^{(r)} \circ \beta_2^{(r)} \circ \cdots \circ \beta_D^{(r)}, X \right\rangle \\ &= \alpha + \gamma^\top Z + \langle (B_D \odot \cdots \odot B_1) \mathbf{1}_R, \text{vec} X \rangle, \qquad (4) \end{aligned}$$

where $B_d = [\beta_d^{(1)}, \ldots, \beta_d^{(R)}] \in \mathbb{R}^{p_d \times R}$, $B_D \odot \cdots \odot B_1 \in \mathbb{R}^{\prod_d p_d \times R}$ 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 $B = [\![B_1, \ldots, B_D]\!]$. 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 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 multidimensional 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
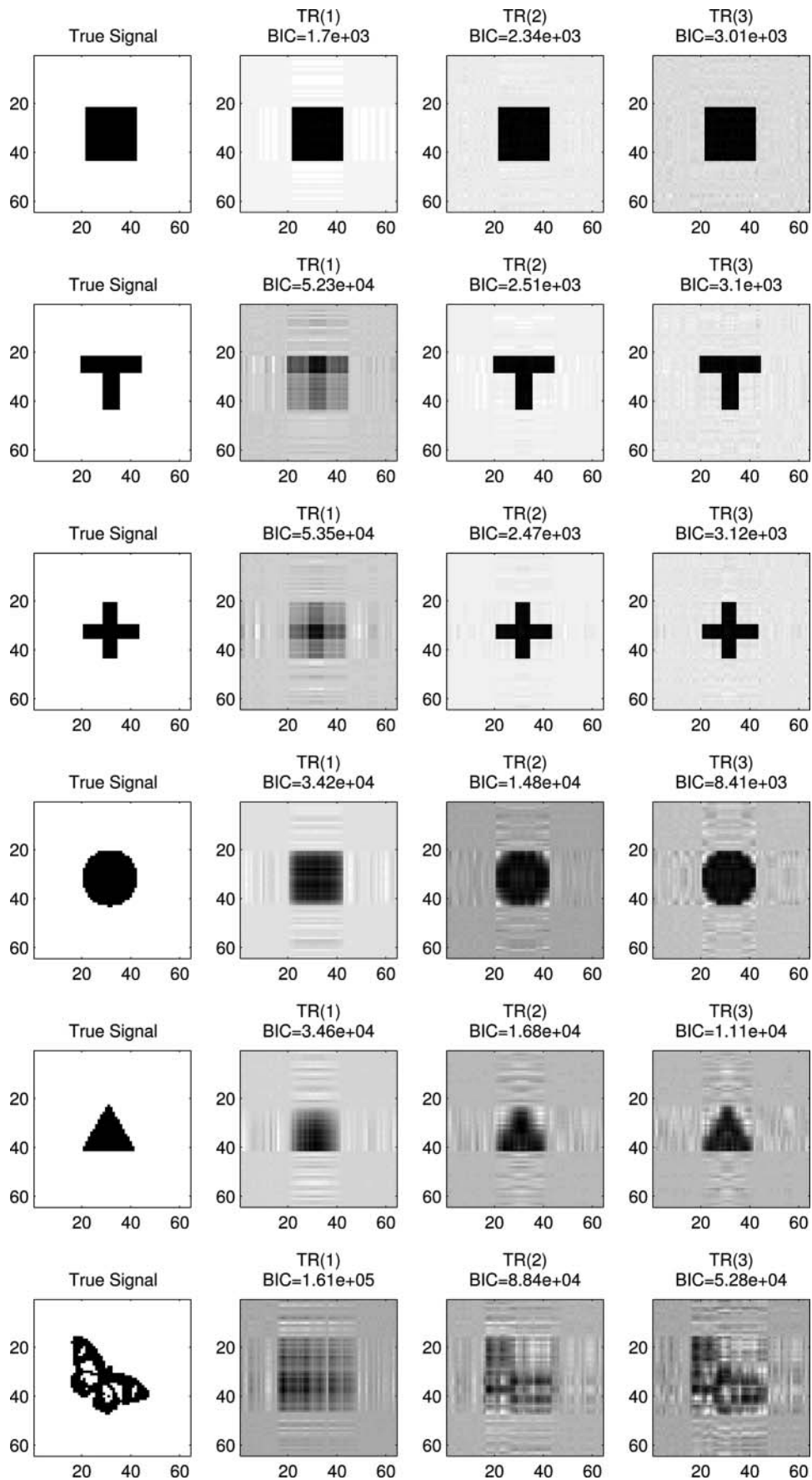
Figure 1. True and recovered image signals by tensor regression. The matrix variate has size 64×64 with entries generated as independent standard normals. The regression coefficient for each entry is either 0 (white) or 1 (black). The sample size is 1000. TR(R) means estimate from the rank-R tensor regression.

general tensor regression model (4) is able to capture significantly more tensor signals than the basic model (3).

3. ESTIMATION

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

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

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 $\boldsymbol{B}_d$,

---

Algorithm 1 Block relaxation algorithm for maximizing (5).

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

$d = 1, \ldots, D$, 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

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

Consequently, the problem turns into a traditional GLM regression with $R p_d$ 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-square procedure (de Leeuw, Young, and Takane 1976).

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, that is, $M(\boldsymbol{\theta}^{(t)}) = \boldsymbol{\theta}^{(t+1)}$, 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, that is, 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 indeterminacy) 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, for example, 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 MLE for the tensor regression model defined by (2) and (4). For simplicity, we omit the intercept α and the covariate part $\boldsymbol{\gamma}^\mathsf{T} \boldsymbol{Z}$, though the conclusions generalize to an arbitrary combination of covariates. In this article, 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.

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

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

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 $\text{vec}\boldsymbol{B}$, that is, $\text{vec}\boldsymbol{B} = \boldsymbol{\Pi}_d \, \text{vec}\boldsymbol{B}_{(d)}$.

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 that 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

$$H(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) = -\frac{[\mu'(\eta)]^2}{\sigma^2} ([\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]^{\mathsf{T}} \text{vec}\boldsymbol{X}) \\ \times ([\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]^{\mathsf{T}} \text{vec}\boldsymbol{X})^{\mathsf{T}} \\ + \frac{(y - \mu)\theta''(\eta)}{\sigma^2} ([\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]^{\mathsf{T}} \text{vec}\boldsymbol{X}) \\ \times ([\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]^{\mathsf{T}} \text{vec}\boldsymbol{X})^{\mathsf{T}} \\ + \frac{(y - \mu)\theta'(\eta)}{\sigma^2} d^2\eta(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D), \tag{8}$$

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

3. The Fisher information matrix is

$$I(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) = E[-H(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)] \\ = \text{var}[\nabla \ell(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) d\ell(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)] \\ = \frac{[\mu'(\eta)]^2}{\sigma^2} [\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]^{\mathsf{T}} (\text{vec}\boldsymbol{X})(\text{vec}\boldsymbol{X})^{\mathsf{T}} \\ \times [\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]. \tag{9}$$

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 nonstochastic, 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:

1. Scaling: $\boldsymbol{B} = [\![\boldsymbol{B}_1 \boldsymbol{\Lambda}_1, \ldots, \boldsymbol{B}_D \boldsymbol{\Lambda}_D]\!]$ for any diagonal matrices $\boldsymbol{\Lambda}_d = \text{diag}(\lambda_{d1}, \ldots, \lambda_{dR})$, $d = 1, \ldots, D$, such that $\prod_d \lambda_{dr} = 1$ for $r = 1, \ldots, R$.
2. 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 non-singular transformation indeterminacy: $\boldsymbol{B}_1 \boldsymbol{B}_2^{\mathsf{T}} = \boldsymbol{B}_1 \boldsymbol{O} \boldsymbol{O}^{-1} \boldsymbol{B}_2^{\mathsf{T}}$ for any R-by-R nonsingular matrix $\boldsymbol{O}$. Note the scaling and per-mutation indeterminacy is subsumed in the nonsingular trans-formation indeterminacy. The singular value decomposition 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 article, we adopt the following convention. $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_{D-1}$ are scaled such that $\beta_{d1}^{(r)} = 1$, that is, 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 indeterminacy, 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

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

which is open and convex. The formulas for score, Hessian and information in Proposition 2 require changes accordingly, that is, 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 $\boldsymbol{B}$ is arbitrary and exclude many arrays that might be of interest,

for example, 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 nonuniqueness 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 $\boldsymbol{A}$, that is, 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 $\min_{d=1,\ldots,D} \mathrm{rank}(\boldsymbol{B}_1 \odot \cdots \odot \boldsymbol{B}_{d-1} \odot \boldsymbol{B}_{d+1} \odot \cdots \odot \boldsymbol{B}_D) = R$, which in turn implies that $\min_{d=1,\ldots,D}(\prod_{d' \neq d} \mathrm{rank}(\boldsymbol{B}_{d'})) \geq R$.

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 $R(R - 1) \leq p_1 p_2 p_3 (3 p_1 p_2 p_3 - p_1 p_2 - p_1 p_3 - p_2 p_3 - p_1 - p_2 - p_3 + 3)/4$, 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 $\boldsymbol{B}_0 \in \boldsymbol{B}$ 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

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

is nonsingular.

Remark 3. Proposition 4 explains the merit of tensor regression from another angle. For identifiability, the classical linear regression requires $\mathrm{vec}\, \boldsymbol{x}_i \in \mathbb{R}^{\prod_d p_d}$, $i = 1, \ldots, n$, to be linearly independent 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

$[\boldsymbol{J}_1 \ldots \boldsymbol{J}_D]^{\mathsf{T}} \mathrm{vec}\, \boldsymbol{x}_i \in \mathbb{R}^{R(\sum_d p_d - D + 1)}$, $i = 1, \ldots, n$. The requirement on sample size is greatly lessened by imposing structure on the arrays.

Remark 4. Although global identifiability is hard to check for a finite sample, a parameter point $\boldsymbol{B} \in \boldsymbol{B}$ 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 $\{\langle \boldsymbol{B}, \boldsymbol{X} \rangle, \boldsymbol{B} \in \boldsymbol{B}\}$ forms a Vapnik-Červonenkis class. Then standard uniform convergence theory for *M*-estimation (van der Vaart 1998) applies.

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

Remark 5. (Misspecified rank) In practice it is rare that the true regression coefficient $\boldsymbol{B}_{\mathrm{true}} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ 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 q.m.d.

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

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

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

5. REGULARIZED ESTIMATION

Sample size in typical neuroimaging studies is often limited. Even for a low-rank tensor model, it is likely that the number of parameters exceeds the sample size. As such the $p > n$ challenge is often the rule rather than the exception in neuroimaging analysis. Regularization is essential to handle $p > n$, and is also useful for stabilizing the estimates and improving their risk property when $p < n$. We emphasize that there are a large number of regularization techniques for different purposes. Here, we illustrate with using *sparsity* regularization for identifying subregions that are associated with the response traits. This problem can be viewed as an analog of variable selection in the traditional vector-valued covariates. Toward that end, we maximize a regularized log-likelihood function

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

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 $\partial/\partial|\beta| P_\lambda(|\beta|, \rho) = \rho\{1_{\{|\beta|\le\rho\}} + (\lambda\rho - |\beta|)_+/(\lambda - 1)\rho 1_{\{|\beta|>\rho\}}\}$, $\lambda > 2$, 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,

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

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 $\lambda \ge 1$, 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 Akaike information criterion, BIC, and cross-validation. For instance, the recent work (Zhou, Armagan, and Dunson 2011) derives BIC-type criterion for GLM with possibly nonconcave 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 co-

efficient array $\boldsymbol{B} \in \mathbb{R}^{\prod_d p_d}$ without considering any structured decomposition as in our models. That is, one simply treats $\operatorname{vec}\boldsymbol{X}$ 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 Two-Dimensional Shape Examples

We first elaborate on the illustrative example given in Section 2.2 with selected two-dimensional 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 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 two-dimensional matrix, $\boldsymbol{Z}$ is a five-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 supplementary Section A.1. 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.
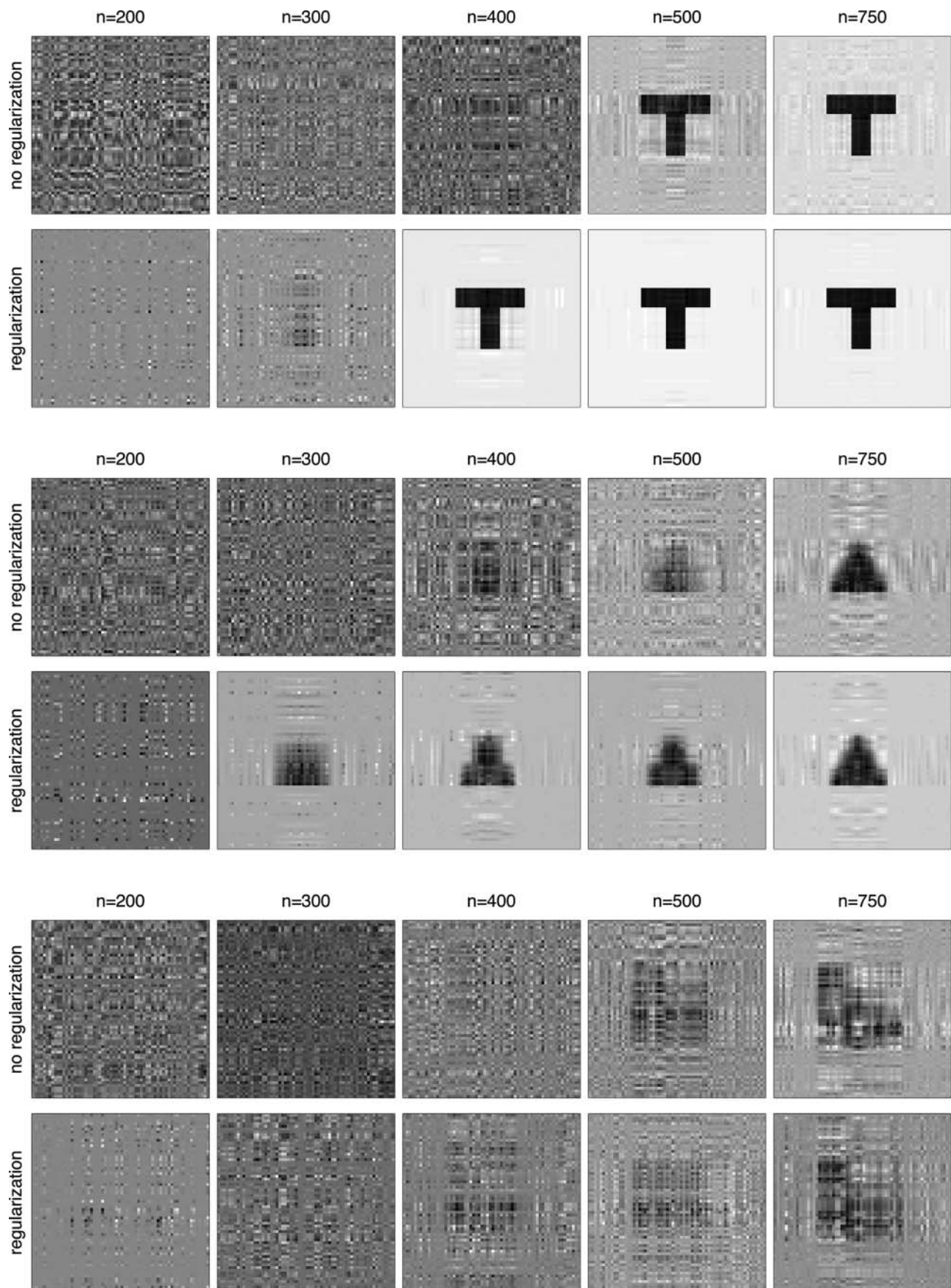
Figure 2. Snapshots of tensor estimation with varying sample size. The matrix variate has size 64×64 with entries generated as independent standard normals. The regression coefficient for each entry is either 0 (white) or 1 (black).

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 supplementary Section 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 overactivity. The dataset that we used is part of the ADHD-200 Global Competition datasets. 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: gray 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 image 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 three-dimensional MRI image with dimension $256 \times 256 \times 198$, 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-1 decomposition, with $\boldsymbol{B}_1 \in \mathbb{R}^{256 \times 1}$, $\boldsymbol{B}_2 \in \mathbb{R}^{256 \times 1}$, and $\boldsymbol{B}_3 \in \mathbb{R}^{198 \times 1}$,

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	γ	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	γ	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)

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 three-dimensional space. The second admits a rank-2 decomposition, with $\boldsymbol{B}_1 \in \mathbb{R}^{256 \times 2}$, $\boldsymbol{B}_2 \in \mathbb{R}^{256 \times 2}$, and $\boldsymbol{B}_3 \in \mathbb{R}^{198 \times 2}$. 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 three-dimensional 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, $Y \sim \text{Poisson}(\mu)$, 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.

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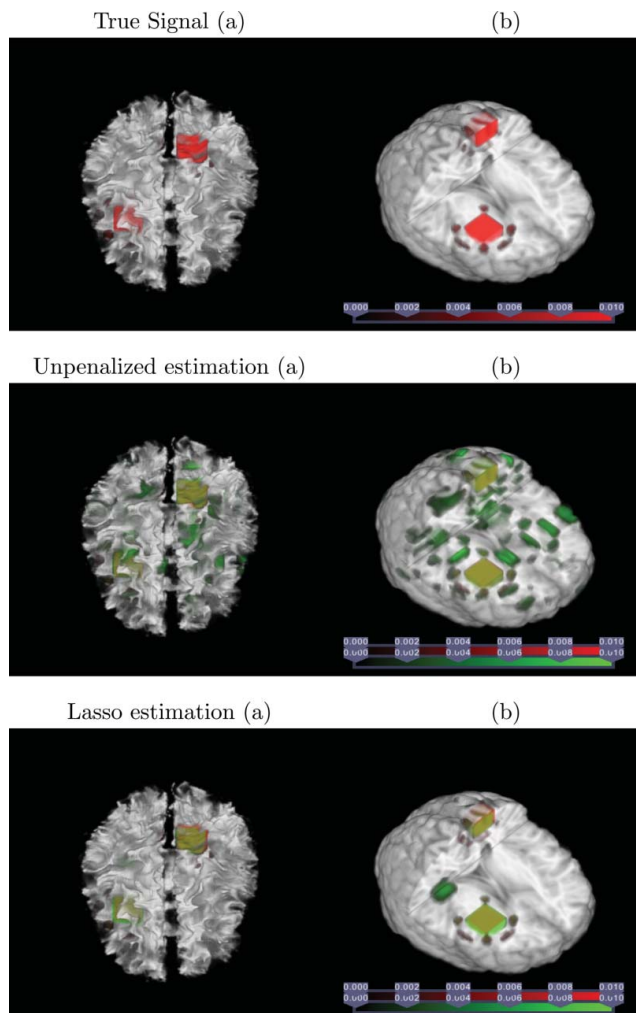
Figure 3. Region selection. The true signal regions are colored in red, the estimated signal regions are in green, and the overlapped regions are in yellow. The left panel is the true or estimated signal overlaid on a randomly selected subject, and the right panel is a three-dimensional rendering of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

temporal and frontal cortices in children with ADHD compared with matched controls (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 have been previous imaging studies using tensor structure (Li, Du, and Lin 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*) framework.

A key motivation of our work is the recent emergence of large-scale neuroimaging data. Traditional 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. The Alzheimer's Disease Neuroimaging Initiative (ADNI 2012) database, which includes over 3000 participants with MRI, fMRI, PET, and genomics data. Our proposed tensor regression model was motivated by and aims to address the computational and modeling challenges of such large-scale imaging data. Meanwhile, our approach is equally applicable to smaller scale imaging dataset, for example, images acquired from a single lab with a moderate number of subjects. In that scenario, the regularization strategy outlined in Section 5 is expected to play a central role in 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–Leibler distance. Moreover, the regular asymptotics is useful for testing significance of a low-rank sparse model in a replication study. Classical hypothesis tests such as likelihood ratio test can be formulated based on the asymptotic normality
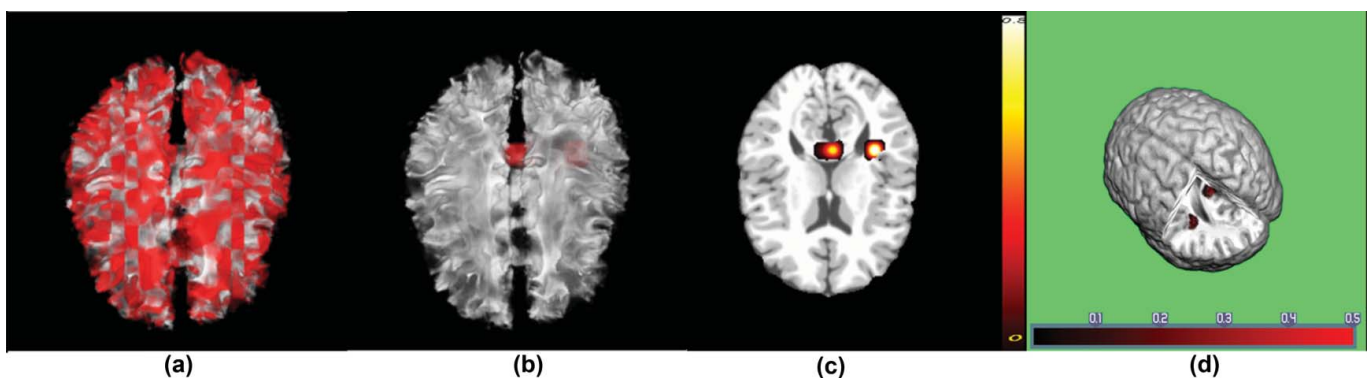

Figure 4. Application to the ADHD data. Panel (a) is the unpenalized estimate overlaid on a randomly selected subject; (b) is the regularized estimate overlaid on a randomly selected subject; (c) is a selected slice of the regularized estimate overlaid on the template; and (d) is a three-dimensional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

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.

Regularization plays a crucial role in practical applications, as the challenge of $p > n$ is the rule rather than the exception in neuroimaging analysis. We consider the tensor regression presented in this article an analog of the *hard thresholding* in classical regression, where the rank of the model is fixed. 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.

In real imaging analysis, the signal is hardly of an exact low-rank structure. However, given the limited sample size, 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 numerical experiments, where a rank-3 model yields an informative recovery of a butterfly shape in Figure 1, and a two-ball structure in Figure 3. We also note that, our proposed regularized tensor model often yields a significantly better recovery than the classical regularized regression model. This is partly verified in an experiment comparing our solution to a classical lasso with vectorized image covariates. The results are given in supplementary Section A.3.

The scale and complexity of neuroimaging data require the estimation algorithm to be highly scalable, efficient, and stable. The methods in this article are implemented in an efficient Matlab toolbox, and the run time is remarkably fast. For instance, the median run time of fitting a rank-3 model to the two-dimensional triangle shape in Figure 1 was about 5 sec. Fitting a rank-3 logistic model to the three-dimensional ADHD data in Section 6.2.3 took about 285 sec for 10 runs from 10 random starting points, averaging < 30 sec per run. Supplementary Section 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 multimodality 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

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

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.

SUPPLEMENTARY MATERIALS

Supplementary materials contains additional numerical results as indicated in Section 6, and all technical proofs.

REFERENCES

ADHD (2012), "The ADHD-200 Sample," available at *http://fcon_1000.projects.nitrc.org/indi/adhd200/*. [550]

ADNI (2012), "Alzheimer's Disease Neuroimaging Initiative," available at *http://adni.loni.ucla.edu*. [550]

Caffo, B., Crainiceanu, C., Verduzco, G., Joel, S., Mostofsky, S. H., 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. [540]

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. [541]

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. [549]

de Lathauwer, L. (2006), "A Link Between the Canonical Decomposition in Multilinear Algebra and Simultaneous Matrix Diagonalization," *SIAM Journal Matrix Analysis Applications*, 28, 642–666. [546]

de Leeuw, J. (1994), "Block-Relaxation Algorithms in Statistics," in *Information Systems and Data Analysis*, eds. H. H. Bock, W. Lenski, and M. M. Richter, Berlin: Springer, pp. 308–325. [544]

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. [544]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [547]

Frank, I. E., and Friedman, J. H. (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35, 109–135. [547]

Friston, K. J. (2009), "Modalities, Modes, and Models in Functional Neuroimaging," *Science*, 326, 399–403. [540]

Hinrichs, C., Singh, V., Mukherjee, L., Xu, G., Chung, M. K., Johnson, S. C.; and ADNI (2009), "Spatially Augmented LP Boosting for AD Classification With Evaluations on the ADNI Dataset," *NeuroImage*, 48, 138–149. [540]

Hung, H., and Wang, C.-C. (2013), "Matrix Variate Logistic Regression Model with Application to EEG Data," *Biostatistics*, 14, 189–202. [541]

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*, 107, 568–577. [540]

Kolda, T. G. (2006), "Multilinear Operators for Higher-Order Decompositions," Technical Report SAND2006-2081, Sandia National Laboratories. [541]

Kolda, T. G., and Bader, B. W. (2009), "Tensor Decompositions and Applications," *SIAM Review*, 51, 455–500. [541,542]

Lange, K. (2010), *Numerical Analysis for Statisticians (Statistics and Computing)* (2nd ed.), New York: Springer. [544]

Lazar, N. A. (2008), *The Statistical Analysis of Functional MRI Data*, New York: Springer. [540]

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. [541]

Li, Y., Du, Y., and Lin, X. (2005), "Kernel-Based Multifactor Analysis for Image Synthesis and Recognition," in *10th IEEE International Conference on Computer Vision (ICCV)*, pp. 114–119. [550]

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. [540]

Lindquist, M. (2008), "The Statistical Analysis of fMRI Data," *Statistical Science*, 23, 439–464. [540]

Liu, X., and Sidiropoulos, N. D. (2001), "Cramér-Rao Lwer Bounds for Low-Rank Decomposition of Multidimensional Arrays," *IEEE Transactions on Signal Processing*, 49, 2074–2086. [546]

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. [540]

McCullagh, P., and Nelder, J. A. (1983), *Generalized Linear Models (Monographs on Statistics and Applied Probability)*, London: Chapman & Hall. [541,542]

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. [550]

Polzehl, J., Voss, H. U., and Tabelow, K. (2010), "Structural Adaptive Segmentation for Statistical Parametric Mapping," *NeuroImage*, 52, 515–523. [540]

Qiu, P. (2005), *Image Processing and Jump Regression Analysis (Wiley Series in Probability and Statistics)*, New York: Wiley. [542]

——— (2007), "Jump Surface Estimation, Edge Detection, and Image Restoration," *Journal of the American Statistical Association*, 102, 745–756. [542]

Rao, C. R., and Mitra, S. K. (1971), *Generalized Inverse of Matrices and its Applications*, New York-London-Sydney: Wiley. [541]

Reiss, P., and Ogden, R. (2010), "Functional Generalized Linear Models With Images as Predictors," *Biometrics*, 66, 61–69. [540]

Rothenberg, T. J. (1971), "Identification in Parametric Models," *Econometrica*, 39, 577–591. [546]

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. [540]

Sidiropoulos, N. D., and Bro, R. (2000), "On the Uniqueness of Multilinear Decomposition of N-Way Arrays," *Journal of Chemometrics*, 14, 229–239. [546]

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. [550]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society,* Series B, 58, 267–288. [547]

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," *Biological Psychiatry*, 61, 1361–1369. [550]

van der Vaart, A. W. (1998), *Asymptotic Statistics of Cambridge Series in Statistical and Probabilistic Mathematics* (Vol. 3), Cambridge: Cambridge University Press. [546]

Worsley, K. J., Taylor, J. E., Tomaiuolo, F., and Lerch, J. (2004), "Unified Univariate and Multivariate Random Field Theory," *NeuroImage*, 23, 189–195. [540]

Yue, Y., Loh, J. M., and Lindquist, M. A. (2010), "Adaptive Spatial Smoothing of fMRI Images," *Statistics and its Interface*, 3, 3–13. [540]

Zhou, H., Armagan, A., and Dunson, D. (2011), "Path Following and Empirical Bayes Model Selection for Sparse Regressions," *arXiv:1201.3528*. [547]

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society,* Series B, 67, 301–320. [547]