Bayesian Methods for Tensor Regression

Rajarshi Guhaniyogi*,

Department of Statistics, UC Santa Cruz, CA, USA

June 25, 2020

Abstract

For many applications pertaining to neuroimaging, social science, international relations, chemometrics, genomics and molecular-omics, datasets often involve variables which are best represented in the form of a multi-dimensional array or *tensor*, which extends the familiar two-way data matrix into higher dimensions. Rather than vectorizing tensor-valued variables prior to analysis which results in loss of inference, new methods have emerged developing regression relationships between variables with either tensor-valued response(s) or predictor(s). Bayesian approaches, in particular, have shown great promise in applications pertaining to tensor regressions. A remarkable feature of fully Bayesian approaches is that they allow flexible modeling of tensor-valued parameters in the regressions involving tensor variables and naturally offer characterization of uncertainty in the parametric and predictive inferences. This paper provides a review of some relevant Bayesian models on tensor regressions developed in recent years. We divide methods according to the objective of the analysis. We begin with tensor regression approaches with a scalar response and a tensor-valued covariate, discuss both parametric and nonparametric modeling options and applications in this framework. We then address the problem of making inference with a tensor response and a vector of covariates, with applications including task related brain activation and connectivity studies. Finally, we offer discussion on Bayesian models involving a tensor response and a tensor covariate. Discussion of each model is accompanied by available results on its posterior contraction properties, laying out restrictions on key model parameters (such as the tensor dimensions) to draw accurate posterior inference.

Keywords: Bayesian Statistics; Gaussian Process; Low Rank Decomposition; Multiway Shrinkage Prior; Posterior Convergence; Tensor Object.

Introduction

Of late, scientific applications in a variety of disciplines encounter datasets where one or more variables are multidimensional arrays or tensors, which are higher dimensional analogues of two dimensional matrices. For example, tensor objects are encountered in molecular-omics profiling with high dimensional data involving multiple subjects, tissues, fluids or time points within a single study [26]. Similarly, in functional magnetic resonance imaging (fMRI) or electroencephalogram (EEG) data, multidimensional arrays are common with different dimensions signifying time points, brain regions or frequencies [12]. Again, there can be time varying tensor valued objects in international relational data with dimensions signifying countries, time points, and diplomatic actions [25]. Scientific objective in such applications often pertains to developing regression relationships with either a tensor-valued response or a predictor.

A few early approaches consider reshaping tensor variables to high dimensional vectors before employing them to regression analysis [67]. Reshaping tensors introduces massive dimensional unstructured vectors in the regression framework. Applying ordinary Bayesian variable selection (see **stat05788**) or shrinkage priors on coefficients of high dimensional predictors causes computational havoc. Additionally, this may result in loss of spatial information in the tensor cells. An alternative approach within the regression framework envisions tensor variables as high dimensional functional variables to take into account the spatial information in them. Several penalized functional regression approaches have emerged in the last decade or so, mainly with functional predictors [20, 48, 68, 30, 16, 14, 61, 49], though many of them may face computational issues with large tensors due to incorporating expensive cross validation procedures for choosing the tuning parameters. Other important approaches include two stage procedures which first conduct a dimension reduction step with a tensor variable, and then fit a model using lower dimensional summaries of the tensor variable (fit) [7]. Similar to the reshaping approach, two stage approaches may also lose inference due to somewhat ignoring the inherent spatial information in the tensor variables.

Rather than summarizing the tensor objects in one way or another, novel approaches have appeared to directly incorporate a tensor valued response or predictor in regression frameworks, mostly in the frequentist paradigm. To this end, [68] propose a generalized linear model framework involving a scalar response and a tensor predictor in which the tensor predictor coefficient is assumed to have a low-rank Cannonical Parallel Factor (CP/PARAFAC) decomposition [33], which is a higher dimensional analogue to factor modeling in two dimensions, formally introduced later. [39] extend the framework to incorporate a more general Tucker decomposition of the tensor regression parameter. Both these models allow for the incorporation of sparsity-inducing regularization for the tensor parameter. Various extensions of such tensor regression frameworks have appeared to address relevant practical issues. For example, a more efficient estimation algorithm of the tensor parameter is proposed in [63], while [60] propose regularization using the total variation penalty on the tensor parameters and argue for it as a more suitable option when the tensor predictor is a piecewise smooth image with jumps and edges. There is also some literature on regressions with a tensor response and scalar predictors. For example, [44] consider an approach with a tensor response and scalar predictors. While they assume a low-rank structure on the tensor coefficient, no sparsity is enforced on the tensor coefficient. [38] propose an alternative strategy following the literature on envelope-based regression models [11], that utilizes a generalized sparsity principle to exploit the redundant information in the tensor response, by seeking linear combinations of the response that are irrelevant to the regression. In the same vein, [55] develop an approach that assumes low-rank decomposition of the tensor coefficient and imposes element-wise sparsity on individual cells of the tensor coefficient, thus offering variable selection in the tensor response regression paradigm. In the frequentist literature of tensor on tensor regression, [46] offer theoretical results under convex regularization with the tensor nuclear norm, though computation of tensor nuclear norm is NP-hard [55, 13].

In comparison, there is a limited Bayesian literature on tensor regressions. Bayesian frameworks in tensor regressions have shown great promise in effective modeling of tensor valued parameters with careful construction of priors that naturally induces sparsity within and across tensor margins and provides model based estimation of tuning parameters. In addition, the need for valid measures of uncertainty on parameter (predictive) estimates is

crucial, especially in settings with low or moderate sample sizes, which naturally motivates a Bayesian approach. This article provides a review of the most relevant Bayesian modeling approaches to tensor regression, which have mostly appeared in the last few years. We divide the article into a few sections according to the modeling objective. We begin with the review of tensor regression approaches with a scalar response and a tensor covariate. Such settings are typically useful in neuroimaging studies which usually record resting state fMRI and brain related phenotype data over a number of subjects. The main objective of such studies pertains to understanding relationships between different brain regions with the phenotype of interest. We discuss the parametric tensor linear model under this setting, provide an overview of novel multiway structured shrinkage priors specifically developed to draw inference, and then extend our discussion to the class of Bayesian non-parametric tensor regression models. Theoretical results laying out conditions for posterior consistency have also found brief mention in this section.

Another important area of Bayesian tensor regression pertains to regression settings with a tensor response and either a vector or a tensor valued predictor. An application of a tensor response with a vector predictor is found in task related brain activation studies. In a typical task-related fMRI experiment, the brain is scanned at small intervals while a subject performs a series of tasks [32, 31, 62, 66, 64]. The objective is to identify brain regions activated by an external stimulus. The Bayesian tensor response regression framework having brain scans as the tensor response with task-related predictors directly fits into drawing inference in single- and multi-subject brain activation studies. The review discusses tensor response regression models and the development of shrinkage priors on tensor parameters to draw efficient posterior inference. We extend the discussion on Bayesian tensor response regression to Bayesian mixed effect tensor response regression models which find an important application in the joint estimation of brain activation and connectivity for multi-subject studies. Connectivity signifies the interaction between brain regions to assess how information is shared across brain regions. We focus upon functional connectivity that seeks to determine brain regions with similar neuronal activity. We will additionally offer a brief discussion on some important theoretical developments in these models laying out sufficient conditions on the tensor dimensions, sparsity, and magnitudes of cell entries to achieve optimal estimation of

tensor coefficients.

As part of the review, we briefly mention some of the recent literature on Bayesian tensor-on-tensor regression. This regression framework is directly motivated by a plethora of important application, including temporal modeling of international relational data [25], the prediction of fMRI from EEG data [12] and the prediction of gene expression across multiple tissues from other genomic variables. We begin by reviewing tensor on tensor regression models which assume that the response and predictor tensors have the same number of modes. We also discuss a more general framework that allows response and predictor tensors to have different number of modes [41]. Finally, we briefly review extensions of tensor on tensor regressions with a time varying tensor response. While the review draws motivations for various approaches mainly from neuro-scientific applications, the methods reviewed in this article lend themselves to the analysis of datasets emerging from various other applications. Figure 1 shows an outline of our review.

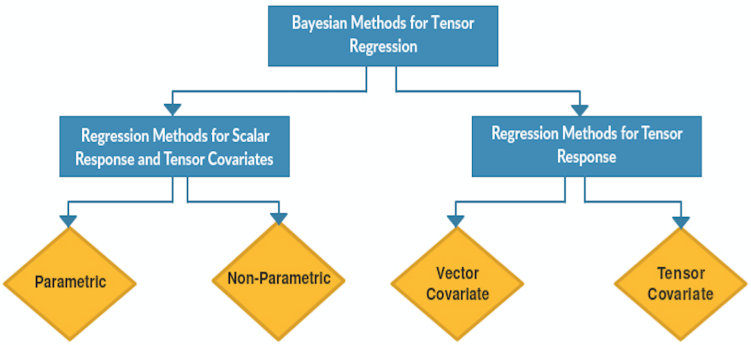
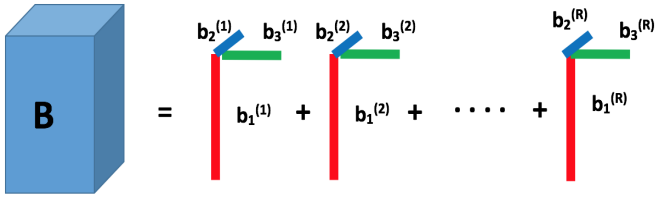


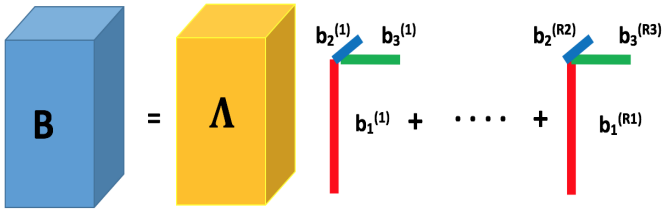
Figure 1: Outline of the Bayesian methods for tensor regressions reviewed in this article. Methods are divided according to the objectives of the analysis.

Notations

We begin by introducing essential notations related to tensors which will appear repeatedly in different sections. A tensor $\boldsymbol{B} \in \otimes_{d=1}^{D} \mathbb{R}^{p_d}$, referred to as a D-way tensor or D-mode tensor, is a multidimensional array whose $(v_1, ..., v_D)$th *cell* is denoted by $B_{(v_1,...,v_D)}$, $1 \leq v_1 \leq p_1$,...,$1 \leq v_D \leq p_D$. When $D = 2$, a tensor corresponds to a matrix. A D-way outer product between vectors $\boldsymbol{b}_d = (b_{d,1}, \ldots, b_{d,p_d})'$, $1 \leq d \leq D$, is a $p_1 \times \cdots \times p_D$



(a) CP/PARAFAC decomposition



(b) Tucker decomposition

Figure 2: Visualization of CP/PARAFAC decomposition of rank-R and Tucker decomposition of ranks R_1, R_2, R_3 for a three dimensional tensor $\boldsymbol{B}$.

tensor denoted by $\boldsymbol{B} = \boldsymbol{b}_1 \circ \boldsymbol{b}_2 \circ \cdots \circ \boldsymbol{b}_D$ with the entry in the $(v_1, .., v_D)$th *cell* given by $B_{(v_1,\ldots,v_D)} = \prod_{d=1}^{D} b_{d,v_d}$. Define a $vec(\boldsymbol{B})$ operator as one that stacks elements of this tensor into a column vector of length $\prod_{d=1}^{D} p_d$. From the definition of outer products, it is easy to see that $vec(\boldsymbol{b}_1 \circ \boldsymbol{b}_2 \circ \cdots \circ \boldsymbol{b}_D) = \boldsymbol{b}_D \otimes \cdots \otimes \boldsymbol{b}_1$. As a higher order generalization of matrix singular value decomposition, the Tucker decomposition of a D-way tensor $\boldsymbol{B} \in \otimes_{d=1}^{D} \mathbb{R}^{p_d}$ is often considered. The Tucker decomposition [57, 33] can be expressed as

$$\boldsymbol{B} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \cdots \sum_{r_D=1}^{R_D} \lambda_{r_1,\ldots,r_D} \boldsymbol{b}_1^{(r_1)} \circ \boldsymbol{b}_2^{(r_2)} \circ \cdots \circ \boldsymbol{b}_D^{(r_D)}, \tag{1}$$

where $\boldsymbol{b}_d^{(r_d)}$ is a p_d dimensional vector, $1 \leq d \leq D$, often referred to as the *tensor margins* and $\boldsymbol{\Lambda} = (\lambda_{r_1,\ldots,r_D})_{r_1,\ldots,r_D=1}^{R_1,\ldots,R_D}$, referred to as the *core tensor*. If one considers $\{\boldsymbol{b}_d^{(r_d)}; 1 \leq r_d \leq R_d, 1 \leq d \leq D\}$ as "factor loadings" and $\lambda_{r_1,\ldots,r_D}$ to be the corresponding coefficients, then the Tucker decomposition may be thought of as a multiway analogue to factor modeling. A rank-R CP/PARAFAC decomposition emerges as a special case of Tucker decomposition (1) when $R_1 = R_2 = \cdots = R_D = R$ and $\lambda_{r_1,\ldots,r_D} = \lambda_r I(r_1 = r_2 = \cdots = r_D = r)$ [22, 33]. Figure 2 provides a pictorial view of PARAFAC and Tucker decompositions for $D = 3$.

A mode-d fiber of a D-way tensor is obtained by fixing all dimensions of a tensor except the d-th one. For example, in a matrix (equivalently a 2-way tensor), a column is a mode-1

fiber and a row is a mode-2 fiber. A d-th mode vector product of a D-way tensor $\boldsymbol{B}$ and vector $\boldsymbol{a} \in \mathbb{R}^{p_d}$, denoted by $\boldsymbol{B}\bar{\times}_d\boldsymbol{a}$, is a tensor of the order of $p_1 \times \cdots \times p_{d-1} \times p_{d+1} \times \cdots \times p_D$, whose elements are the inner product of each mode-d fiber of $\boldsymbol{B}$ with $\boldsymbol{a}$. The Tucker product [57] between a D-way tensor $\boldsymbol{A}$ of dimensions $p_1 \times \cdots \times p_D$ and matrices $\boldsymbol{B}_1$, $\boldsymbol{B}_2$,..,$\boldsymbol{B}_D$ of dimensions $m_1 \times p_1$,...,$m_D \times p_D$ respectively, denoted by $\boldsymbol{A} \times \{\boldsymbol{B}_1, ..., \boldsymbol{B}_D\}$, is defined as a tensor $\boldsymbol{C}$ such that $vec(\boldsymbol{C}) = (\boldsymbol{B}_D \otimes \cdots \otimes \boldsymbol{B}_1)vec(\boldsymbol{A})$. The Tucker product essentially maps a $p_1 \times \cdots \times p_D$ tensor to a $m_1 \times \cdots \times m_D$ tensor. For $D = 2$, it follows that $\boldsymbol{C} = \boldsymbol{B}_1\boldsymbol{A}\boldsymbol{B}_2'$.

For two tensors $\boldsymbol{A}$ and $\boldsymbol{B}$ of dimensions $p_1 \times \cdots \times p_{D_1} \times m_1 \times \cdots \times m_{D_3}$ and $m_1 \times \cdots \times m_{D_3} \times q_1 \times \cdots \times q_{D_2}$ respectively, the contracted product between the two tensors is a tensor of dimension $p_1 \times \cdots \times p_{D_1} \times q_1 \times \cdots \times q_{D_2}$, whose $(v_1, ..., v_{D_1}, u_1, ..., u_{D_2})$th element is given by $\sum_{i_1=1}^{m_1} \cdots \sum_{i_{D_3}=1}^{m_{D_3}} A_{(v_1,...,v_{D_1},i_1,..,i_{D_3})}B_{(i_1,..,i_{D_3},u_1,...,u_{D_2})}$. Finally, we use $||\cdot||$ and $||\cdot||_\infty$ to denote the L_2 and L_∞ norms, respectively, for both vectors and higher order tensors.

1 Bayesian Tensor on Scalar Regression Model

1.1 Parametric Bayesian Tensor on Scalar Regression Model

Neuroscientific studies often involve 3-dimensional resting state fMRI scans, along with age, gender and other behavioral variables and brain related phenotypes for a number of subjects [40, 34]. The phenotypes of interest can be continuous, binary (e.g., presence or absence of a neuronal disease) or categorical (e.g., different levels of a specific disorder). In these applications, the inferential interest lies in predicting the brain related phenotype based on the fMRI scan (treated as a tensor object) as well as from behavioral variables. Further, it is of specific interest to identify the regions in the brain (alternatively, the cells in the tensor predictor) that are predictive of the response. Let $y \in \mathcal{Y}$ denotes a response variable, with $\boldsymbol{z} \in \mathcal{X} \subset \mathbb{R}^p$ and $\boldsymbol{X} \in \otimes_{d=1}^D \mathbb{R}^{p_d}$ being the scalar and tensor predictors, respectively. Depending on the nature of the response (continuous, binary, categorical or count), a generalized tensor regression model is proposed in these contexts as below

$$E[y|\boldsymbol{X}, \boldsymbol{z}] = g_y^{-1}(\mu + \boldsymbol{z}'\boldsymbol{\gamma} + \langle \boldsymbol{X}, \boldsymbol{B}\rangle), \quad \langle \boldsymbol{X}, \boldsymbol{B}\rangle = \text{vec}(\boldsymbol{X})'\text{vec}(\boldsymbol{B}), \tag{2}$$

where $g_y(\cdot)$ is an appropriate link function, $\boldsymbol{\gamma}$ is a $p \times 1$ coefficient for scalar predictors and $\boldsymbol{B} \in \otimes_{d=1}^{D} \mathbb{R}^{p_d}$ is the tensor coefficient corresponding to the measured tensor predictor $\boldsymbol{X}$. [19] consider the tensor linear regression model with $g_y(\cdot)$ as the identity link function, whereas [6] discuss binary tensor regression model with zero-inflated logit link.

Without imposing any additional structure on the coefficient tensor $\boldsymbol{B}$, its estimation involves $\prod_{d=1}^{D} p_d$ parameters. To reduce the number of free parameters, a rank-R PARAFAC decomposed structure is commonly adopted for $\boldsymbol{B}$ [68, 19]. Under the assumed rank-R PARAFAC decomposition for $\boldsymbol{B}$, model (2) requires estimating $R \sum_{d=1}^{D} p_d$ as opposed to $\prod_{d=1}^{D} p_d$ parameters for the unstructured $\boldsymbol{B}$. To appreciate the extent of dimension reduction offered by the PARAFAC decomposition, notice that with a tensor predictor of dimension $30 \times 30 \times 30$, the unstructured $\boldsymbol{B}$ requires estimating 27000 parameters, whereas the rank-R PARAFAC structure involves only $90R$ free parameters. In many applications, $R \sim 5 - 15$ seems sufficient, the number of free parameters ranges between a couple of hundreds to couples of thousands. As pointed out by [19], such a reduction in the number of parameters facilitates computationally efficient estimation of $\boldsymbol{B}$. However, the imposed decomposition induces a nonlinear relationship between tensor margins and cell entries in $\boldsymbol{B}$. In particular, the $\boldsymbol{v} = (v_1, ..., v_D)$th cell entry of $\boldsymbol{B}$ is given by $B_{(v_1,...,v_D)} = \sum_{r=1}^{R} \prod_{d=1}^{D} b_{d,v_d}^{(r)}$. As one is interested in identifying geometric sub-regions of the tensor in which coefficients are not close to zero, with the remaining elements being very close to zero, one wonders whether such a dramatic dimension reduction retains sufficient flexibility. We will return to this question in due course with sufficient theoretical justification.

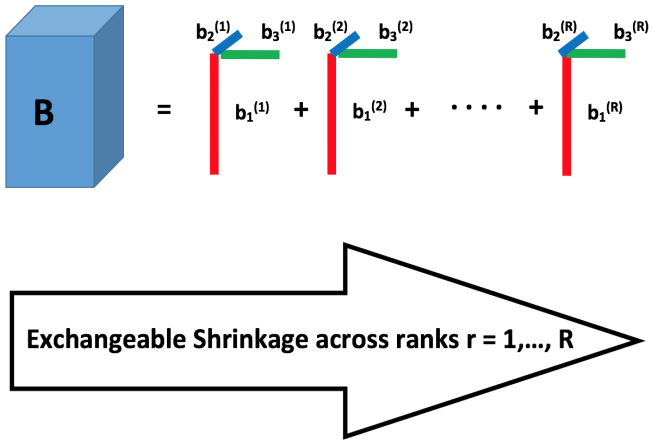
Although the tensor coefficient $\boldsymbol{B}$ is identifiable, [19] comments on non-identifiability of tensor margins $\boldsymbol{b}_d^{(r)}$, $d = 1, .., D$, $r = 1, ..., R$. Constructing a prior distribution on the tensor margins after adding necessary identifiability restrictions turns out to be inefficient. Instead, [19] propose a new class of prior distributions on tensor margins to draw efficient posterior inference on $\boldsymbol{B}$ (rather than on tensor margins) ignoring any identifiability constraints, and observe rapid convergence for the tensor parameter $\boldsymbol{B}$. To elaborate, the *multiway Dirichlet generalized double Pareto* (M-DGDP) prior introduced in [19] induces two types of shrinkage in estimating $\boldsymbol{B}$. Shrinkage across ranks (or summands in the PARAFAC decomposition) takes place in an exchangeable way, with global scale $\tau \sim \text{Ga}(a_\tau, b_\tau)$ adjusted in each rank

as $\tau_r = \phi_r \tau$ for $r = 1, \ldots, R$, where $\Phi = (\phi_1, \ldots, \phi_R) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_R)$ encourages shrinkage towards lower ranks in the assumed PARAFAC decomposition. In addition, $\boldsymbol{W}_{dr} = \text{diag}(w_{dr,1}, \ldots, w_{dr,p_d})$, $d = 1, \ldots, D$ and $r = 1, \ldots, R$ are margin-specific scale parameters for each component. The hierarchical margin-level prior is given by

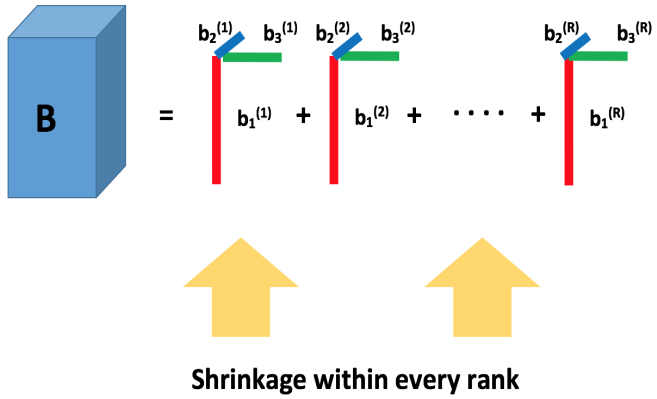
$$\boldsymbol{b}_d^{(r)} \sim \text{N}\big(\boldsymbol{0}, (\phi_r \tau)\boldsymbol{W}_{dr}\big), \quad w_{dr,v_d} \sim \text{Exp}(\lambda_{dr}^2/2), \quad \lambda_{dr} \sim \text{Ga}(a_\lambda, b_\lambda). \tag{3}$$

Collapsing over element-specific scales, notice that $b_{d,v_d}^{(r)} | \lambda_{dr}, \phi_r, \tau \overset{\text{iid}}{\sim} \text{DE}(\lambda_{dr}/\sqrt{\phi_r \tau})$, $1 \leq v_d \leq p_d$. Further, integrating over λ_{dr}, $b_{d,v_d}^{(r)} | \phi_r, \tau$ can be shown to induce a Generalized Double Pareto shrinkage prior on the individual margin coefficients, which in turn is a Bayesian analogue to an adaptive LASSO penalty [3] (see **stat07543.pub2**) and imparts shrinkage on tensor margins. Flexibility in estimating $\{\boldsymbol{b}_d^{(r)}; 1 \leq d \leq D\}$ is accommodated by modeling within-margin heterogeneity via element-specific scaling w_{dr,v_d}. Common rate parameter λ_{dr} shares information between margin elements, encouraging shrinkage at the local scale. The prior achieves shrinkage across ranks as well as within a margin, and hence is coined as a multiway shrinkage prior. Figure 3 shows a pictorial depiction of the prior. The choice of hyper-parameters $\alpha_1, ..., \alpha_R, a_\tau, b_\tau$ crucial to impose adequate tail behavior of $\boldsymbol{B}$ a apriori (see [19] for more details). Later, [6] prove that the imposed prior on $B_{(v_1,..,v_D)}$ has a thicker tail than the ordinary Bayesian LASSO [43] shrinkage prior. Advantage of the M-DGDP prior is that the posterior full conditionals of most parameters are in closed forms, leading to simple Gibbs sampling updates and rapid convergence [19].

An important question posed earlier is that if the dimension reduction step using the low-rank factorization and subsequent prior structure on tensor margins limit flexibility in estimating the tensor parameter $\boldsymbol{B}$. To this end, [19] offer theoretical results proving consistency of the proposed model in estimating the regression function. The consistency of Bayesian models and estimators are determined using the notion of posterior consistency [15, 2, 4]. Let $f(y|\boldsymbol{X}, \boldsymbol{B})$ denotes the density of y, and assume that the true data generating model also belongs to the class of tensor regression models with density given by $f(y|\boldsymbol{X}, \boldsymbol{B}^0)$.



(a) M-DGDP prior shrinkage across ranks



(b) M-DGDP prior shrinkage within a rank

Figure 3: Multiway Dirichlet Generalized Double Pareto (M-DGDP) prior imposes exchangeable shrinkage across ranks for the model to opt for a lower rank structure of $\boldsymbol{B}$. Within every rank, the elements in the tensor margin are assigned shrinkage priors for adequate estimation of $\boldsymbol{B}$.

Define a Kullback-Leibler (KL) neighborhood around the true tensor $\boldsymbol{B}^0$ as

$$\mathcal{B}_n = \left\{ \boldsymbol{B} : \frac{1}{n} \sum_{i=1}^{n} \mathrm{KL}(f(y_i|\boldsymbol{X}_i, \boldsymbol{B}^0), f(y_i|\boldsymbol{X}_i, \boldsymbol{B})) < \epsilon \right\}.$$

Further, let π_n and Π_n denote prior and posterior densities of $\boldsymbol{B}$ with n observations,

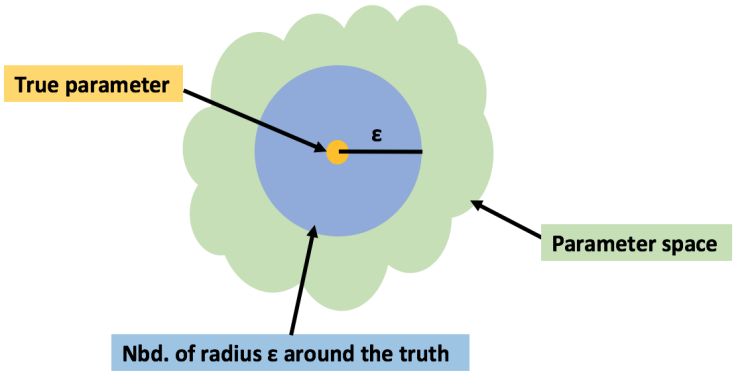


Figure 4: Visualization of posterior convergence. The yellow point presents the true parameter and blue ball presents a neighborhood of radius ϵ around the truth. Posterior consistency roughly means as n increases, most of the posterior mass concentrates inside the blue ball for any choice of ϵ. In the limit (as $n \to \infty$), the posterior probability inside the ball tends to 1, for any choice of ϵ. In the study of posterior contraction rate, ϵ is considered to be a decreasing function of n, $\epsilon_n \to 0$. The task remains to find the fastest rate of decay to 0 for ϵ_n as a function of n. The choice of the neighborhood has profound effect on the posterior convergence property of a model.

respectively. [19] have established posterior consistency by showing that the following result holds for the proposed tensor regression model with $g_y(\cdot)$ as the identity link

$$\Pi_n\left(\mathcal{B}_n^c\right) \to 0 \quad \text{under } \boldsymbol{B}^0 \quad \text{a.s. as } n \to \infty. \tag{4}$$

Figure 3 further clarifies the concept of posterior consistency. [19] have adopted a framework where the tensor dimensions $p_{d,n}$s are considered to be increasing functions of the sample size n. Such a framework allows investigating the maximum rate at which $p_{d,n}$s can grow to maintain (4). It has been shown in [19] that posterior consistency of the model can be maintained with $\sum_{d=1}^{D} p_{d,n} \log(p_{d,n}) = o(n)$ (see Theorem 2 in [19]). Notably, this condition

requires $\sum_{d=D} p_{d,n}$ to grow sub-linearly with sample size n, though the number of cells $\prod_{d=1}^{D} p_{d,n}$ in the tensor can possibly grow at a rate much faster than the sample size n. Hence, the model ensures desirable performance for tensor covariates with a massive number of cells, even in presence of moderate sample size (refer to [19] for a more detailed discussion). The theorem also assumes that $\boldsymbol{B}^0$ admits a rank-R PARAFAC decomposition and $M_n = \frac{1}{n}\sqrt{\sum_{i=1}^{n} ||\boldsymbol{X}_i||_2^2}$ grows slowly as a function of n. Later, [17] have established a much stronger "near optimal" convergence rate under similar assumptions. Specifically, [17] have shown that the risk function

$$\frac{1}{n}\sum_{i=1}^{n} E_{\boldsymbol{B}^0} \int \mathrm{KL}(f(y_i|\boldsymbol{B}^0), f(y_i|\boldsymbol{B}))\pi_n(\boldsymbol{B}|\{y_i, \boldsymbol{X}_i\}_{i=1}^n) \tag{5}$$

is bounded above by ϵ_n^2 where ϵ_n can be taken as $n^{-1/2}$ upto some $\log(n)$ factor. Importantly, [56] also showed decay of (5) to 0 at an "optimal" rate, with an alternative specification of prior distributions on the tensor margins, though the prior specified may appear to be less conducive to full scale Bayesian computation.

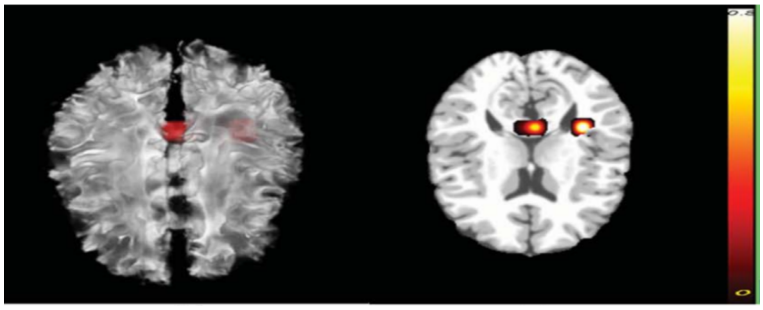


Figure 5: Application to the ADHD data. Left panel is the regularized estimate overlaid on a randomly selected subject. Right panel is a selected slice of the regularized estimate overlaid on the template. Picture courtesy [68].

To demonstrate practical application of tensor regression framework developed in this section, a brief analysis of tensor regression is presented in a neuroimaging data. The dataset records resting state fMRI and T1-weighted images of 776 subjects, either normal, or suffering from attention deficit hyperactivity disorder (ADHD). Information on demographic and behavioral variables for the subjects are also available. A binary tensor regression with $g_y(\cdot)$ as the logit link has been fitted to the data considering binary indicators for subjects as

responses and brain images as tensor covariates. The analysis reveals two regions of interest significantly associated with the ADHD: left temporal lobe white matter and the splenium that connects parietal and occipital cortices across the midline in the corpus callosum (see Figure 5); both these findings are consistent with earlier studies. In fact, prior studies have revealed prominent volume reductions in the temporal and frontal cortices in children with ADHD compared with matched controls [53]. An earlier study has also recorded a reduced size of the splenium in the corpus callosum being responsible for ADHD [58]. We refer to [56] for more details on the data and analysis.

1.2 Nonparametric Tensor on Scalar Regression

In estimating regression relationships between a scalar response and a tensor predictor, the bias-variance tradeoff is a central issue, both from theoretical and practical perspectives. In the parametric generalized tensor model discussed in Section 1.1, the function class that the model can represent is critically restricted due to its linearity in the mean function and the low-rank constraint, implying that the variance error is low but the bias error is high if the true functional relationship between y_i and $\boldsymbol{X}_i$ is either nonlinear or full rank. An alternative approach is to fit the nonparametric tensor regression model to the data [67, 27] given by

$$y_i = f(\boldsymbol{X}_i) + \epsilon_i, \epsilon_i \overset{iid}{\sim} N(0, \sigma^2). \tag{6}$$

Here $f(\cdot)$ can represent a wide range of functions and the bias error can be close to zero, though at the expense of flexibility due to the notorious high dimensionality associated with estimating (6). In order to mitigate the burden of high dimensionality, [29] propose a new class of additive-multiplicative nonparametric regression (AMNR) models. AMNR employs low-rank decomposition for both the tensor predictor $\boldsymbol{X}$ and the function $f(\cdot)$, and hence is referred to as a doubly decomposing nonparametric tensor regression framework.

To elaborate, [29] propose decomposing the tensor predictor $\boldsymbol{X} = \sum_{r=1}^{S} \xi_s \boldsymbol{x}_1^{(s)} \circ \cdots \circ \boldsymbol{x}_D^{(s)}$,

$\xi_1 \geq \xi_2 \geq \cdots$ using a rank-S PARAFAC decomposition, with each $\boldsymbol{x}_d^{(s)}$ a unit vector. Define

$$f^{AMNR}(\boldsymbol{X}) = \sum_{r=1}^{R} \sum_{s=1}^{S} \xi_s \prod_{d=1}^{D} f_d^{(r)}(\boldsymbol{x}_d^{(s)}). \tag{7}$$

$f^{AMNR}(\cdot)$ can be conceptualized as a rank-R PARAFAC decomposition of $f(\cdot)$ in an infinite space. In fact, any function in a Sobolev space of order β can be approximated to any degree by a function of this form, with each $f_d^{(r)}(\cdot)$ belonging to a Sobolev space of order β [21]. To estimate $f^{AMNR}(\cdot)$, each local function $f_d^{(r)}(\cdot)$ is assigned a Gaussian process prior distribution (see **stat04542**).

Interestingly, AMNR is interpretable as a piecewise non-parametrization of the tensor model proposed in (2) with $g_y(\cdot)$ as the identity link. Assuming $\boldsymbol{B}$ and $\boldsymbol{X}$ to have rank-R and rank-S PARAFAC decompositions respectively, $\langle \boldsymbol{X}, \boldsymbol{B} \rangle = \sum_{r=1}^{R} \sum_{s=1}^{S} \xi_s \prod_{d=1}^{D} \langle \boldsymbol{b}_d^{(r)}, \boldsymbol{x}_d^{(s)} \rangle$. Thus AMNR replaces $\langle \boldsymbol{b}_d^{(r)}, \boldsymbol{x}_d^{(s)} \rangle$ by a local function $f_d^{(r)}(\boldsymbol{x}_d^{(s)})$. [29] generalize a few earlier approaches on nonparametric tensor regression, such as [52], which use the same vector input for every $f_d^{(r)}(\cdot)$ and restrict $S = 1$. Other prominent approaches in the context of nonparametric tensor regression includes the Tensor Gaussian Process (TGP) model [67], which is essentially a GP regression model that reshapes a tensor into a high-dimensional vector and takes the high dimensional vector as a predictor. While tensor GP is found to demonstrate satisfactory practical performance, AMNR derives additional advantage by offering theoretical support for the class of nonparametric tensor regression models discussed below.

Suppose $\hat{f}_n(\cdot)$ is the Bayes estimator for estimating $f(\cdot)$ with sample size n. Let the metric $||g||_n$ for any function $g(\cdot)$ represents the quantity $\frac{1}{n} \sum_{i=1}^{n} g(\boldsymbol{X}_i)^2$. Assume that the true data generating model is given by (6) with the true regression function $f^*(\cdot)$ belonging to the space of Sobolev functions of order β. [29] mentioned two key assumptions in the theoretical development of AMNR: (a) the true data generating regression function $f^*(\cdot)$ and functions generated from the fitted Gaussian process priors have the same degree of smoothness; (b) $f^*(\cdot)$ satisfies a rank additivity condition which essentially implies that for any tensor $\boldsymbol{X}^*$ admitting a rank-R^* PARAFAC decomposition $\boldsymbol{X}^* = \sum_{r=1}^{R^*} \bar{\xi}_r \bar{\boldsymbol{x}}_1^{(s)} \circ \cdots \circ \bar{\boldsymbol{x}}_D^{(s)}$, one has $f^*(\boldsymbol{X}^*) = \sum_{r=1}^{R^*} f^*(\bar{\xi}_r \bar{\boldsymbol{x}}_1^{(s)} \circ \cdots \circ \bar{\boldsymbol{x}}_D^{(s)})$ [29]. Rank-additivity has been crucially used

in developing multivariate additive model analysis, see [23], [47]. Assume that f^* admits a rank S^* decomposition in the sense of (7). [29] observed different convergence rates of estimating f^* by $\hat{f}_n$, depending on whether S^* is finite or infinite. When S^* is finite, assuming $S = S^*$, [29] proved that $E||\hat{f}_n - f^*||_n^2$ decays at the rate of $n^{-2\beta/(2\beta+\max_{d=1}^{D} p_d)}$. On the other hand, when S^* is infinite, the convergence rate deteriorates by a factor due to a finite rank approximation of f^*, details of which can be found in [29].

Before concluding this section, we would like to make an important observation. Notice that the existing literature on Bayesian tensor regressions assumes low-rank PARAFAC decomposition either on the tensor parameter (Section 1.1) or on the regression function (Section 1.2). There is a scope of imposing a more flexible and expressive Tucker decomposition structure on them. While the difference seems to be a minor one, it can have profound effects, especially on neuro-scientific applications. For example, the freedom in the choice of unequal R_ds $d = 1, .., D$ is essential when the tensor data are skewed in dimensions, which is pretty common in EEG, with the temporal dimension often exceeding the spatial dimension. Besides, the Tucker formulation may often be handy to achieve parsimony for datasets with small to moderate sample sizes [39]. Considering these facts, we expect to see in future a more flexible Bayesian modeling with Tucker decomposition in the tensor regression contexts.

2 Bayesian Tensor Regression Models with a Tensor Response

While regression with a tensor predictor and a scalar response is able to deliver inference on many scientific problems in neuroscience and other disciplines, a variety of applications also motivate models with a tensor response and a vector or a tensor predictor. For example, detecting brain regions activated by an external stimulus or condition is probably the most common objective in fMRI studies [66]. Neuronal activation in response to a stimulus occurs in milliseconds and cannot be observed directly. However, neuronal activation is followed by the metabolic process which increases blood flow and volume in the activated areas, and can therefore be measured by fMRI. During the course of a task-related fMRI experiment, a series of brain images are acquired over multiple time points while a subject performs multiple

tasks, yielding three dimensional tensor responses over time points. The tensor response at each time point is presumed to be associated with the task related predictors and it is of scientific interest to delineate the nature and region of activation. Building a regression framework involving the tensor response and task related predictors is naturally motivated from such scientific problems. We also refer to applications in electroencephalography (EEG) studies, where voltage values are measured from numerous electrodes placed on the scalp over time. The resulting data is a two-dimensional matrix where the readings are both spatially and temporally correlated. These matrix responses are often regressed on a set of scalar predictors (e.g., if a subject is alcoholic or not) to identify their variation with the predictors. Similarly, tensor on tensor regressions have found applications including the prediction of fMRI from EEG data [12], prediction of gene expression across multiple tissues from other genomic variables [45] and temporal dynamics of international relational data [25], to name a few. In what follows, we provide a brief review of Bayesian models for both vector on tensor and tensor on tensor regressions.

2.1 Bayesian Vector on Tensor Regression Models

[18] formulate the Bayesian vector on tensor regression model, mainly from the motivation of a single-subject brain activation study, though the framework can be naturally adapted to other data application contexts. Let $\boldsymbol{Y}_i = ((Y_{i,\boldsymbol{v}}))_{v_1,\ldots,v_D=1}^{p_1,\ldots,p_D} \in \otimes_{d=1}^{D}\mathbb{R}^{p_d}$ denote a tensor valued response for the ith sample, where $\boldsymbol{v} = (v_1, ..., v_D)'$ represents the position of cell $\boldsymbol{v}$ in the D dimensional array of cells. Let $\boldsymbol{x}_i = (x_{1,i}, ..., x_{m,i})' \in \mathcal{X} \subset \mathbb{R}^m$ be the m-dimensional measured vector predictor. Assuming that both response $\boldsymbol{Y}_i$ and predictors $\boldsymbol{x}_i$ are centered around their respective means, the proposed tensor response regression model of $\boldsymbol{Y}_i$ on $\boldsymbol{x}_i$ is given by

$$\boldsymbol{Y}_i = \boldsymbol{\Gamma}_1 x_{1,i} + \cdots + \boldsymbol{\Gamma}_m x_{m,i} + \boldsymbol{E}_i, \tag{8}$$

for $i = 1, ..., n$. $\boldsymbol{\Gamma}_k \in \otimes_{d=1}^{D}\mathbb{R}^{p_d}$, $k = 1, .., m$, is the tensor coefficient corresponding to the predictor $x_{k,i}$. $\boldsymbol{E}_i \in \otimes_{d=1}^{D}\mathbb{R}^{p_d}$ represents the error tensor corresponding to the ith sample. Flexibly modeling the distribution of $\boldsymbol{E}_i$ allows developing the correlation structure between samples and among the cells of $\boldsymbol{Y}_i$. In the context of single-subject brain activation studies,

the ith tensor response represents the brain image observed at time i. In this specific context, the error tensor $\boldsymbol{E}_i$ is assumed to follow a component-wise AR(1) structure (see **stat03473**) across i, $vec(\boldsymbol{E}_i) = \kappa vec(\boldsymbol{E}_{i-1}) + vec(\boldsymbol{\eta}_i)$, where $\kappa \in (-1, 1)$ is the autocorrelation coefficient, and $\boldsymbol{\eta}_i \in \otimes_{d=1}^{D} \mathbb{R}^{p_d}$, with each cell in $\boldsymbol{\eta}_i$ following $N(0, \sigma^2/(1-\kappa^2))$. This ensures both computational simplicity and stationarity in the AR(1) structure.

To tackle the ultra-high dimensional modeling pursuit in estimating the $\boldsymbol{\Gamma}_k$s, [18] propose a rank-R PARAFAC decomposition of each $\boldsymbol{\Gamma}_k$, i.e., $\boldsymbol{\Gamma}_k = \sum_{r=1}^{R} \boldsymbol{\gamma}_{1,k}^{(r)} \circ \cdots \circ \boldsymbol{\gamma}_{D,k}^{(r)}$, where $\boldsymbol{\gamma}_{d,k}^{(r)} = (\gamma_{d,k,1}^{(r)}, ..., \gamma_{d,k,p_d}^{(r)})'$ is a p_d dimensional vector, $1 \leq r \leq R$, $1 \leq d \leq D$ and $k = 1, .., m$. Although the M-DGDP prior constructed on tensor margins in [19] becomes an obvious choice, [18] observe that a straightforward application of the M-DGDP prior on $\boldsymbol{\Gamma}_k$ leads to less accurate uncertainty estimation, perhaps due to less desirable tail behavior of the posterior distribution of the $\Gamma_{\boldsymbol{v},k}$ parameters. Instead, [18] propose a new multiway stick breaking shrinkage (M-SB) prior on the tensor coefficients $\boldsymbol{\Gamma}_k$. The proposed multiway shrinkage prior bears close connection with the M-DGDP prior, the main difference being how it achieves shrinkage across ranks. More specifically, set $\tau_{r,k} = \phi_{r,k}\tau_k$, as the scaling specific to rank $r = 1, ..., R$. Effective shrinkage across ranks is achieved in [18] by adopting a stick breaking construction for the rank-specific parameters $\phi_{r,k}$s, $\phi_{r,k} = \xi_{r,k} \prod_{l=1}^{r-1}(1 - \xi_{l,k})$, $r = 1, ..., R-1$, and $\phi_{R,k} = \prod_{l=1}^{R-1}(1 - \xi_{l,k})$, where $\xi_{r,k} \overset{iid}{\sim} Beta(1, \alpha_k)$. In contrast with the exchangeable shrinkage offered by the M-DGDP prior across ranks, the M-SB prior imposes increasing shrinkage across ranks, favoring a low-rank solution. The global scale parameter is modeled as $\tau_k \sim IG(a_\tau, b_\tau)$. The prior specification is completed by specifying priors on tensor margins with local scale parameters $\boldsymbol{W}_{dr,k} = \text{diag}(w_{dr,k,1}, ..., w_{dr,k,p_d})$ to achieve adequate shrinkage,

$$\boldsymbol{\gamma}_{d,k}^{(r)} \sim N(\boldsymbol{0}, \tau_{r,k}\boldsymbol{W}_{dr,k}), \;\; w_{dr,k,k_1} \sim Exp(\lambda_{dr,k}^2/2), \;\; \lambda_{dr,k} \sim Ga(a_\lambda, b_\lambda), \;\; k_1 = 1, ..., p_d.$$

[18] develop theoretical results providing more insights into the model and the proposed prior. Similar to Section 1.1, the theoretical framework of [18] assumes the number of predictors as well as dimensions of tensor margins as functions of the sample size n (hence using a subscript n). [18] derive restrictions on the growth of m_n and $p_{d,n}$s as functions of n

to ensure asymptotically consistent estimation of $\mathbf{\Gamma}$.

Let $\mathbf{\Gamma}$ be a $m_n \times p_{1,n} \times \cdots \times p_{D,n}$ tensor whose kth slice is given by $\mathbf{\Gamma}_k$. Let the true data generating model be given by (8), with $\mathbf{\Gamma}^0$ as the true tensor coefficient. Since the shrinkage prior on $\mathbf{\Gamma}$ assigns zero probability at point zero, the exact number of nonzero elements of $\mathbf{\Gamma}$ is always $m_n \prod_{d=1}^{D} p_{d,n}$. A meaningful comparison with the number of nonzero elements s_n of the true tensor coefficient $\mathbf{\Gamma}^0$ is made by considering $\tilde{s}_n$, the number of elements of $\mathbf{\Gamma}$ exceeding in absolute value a threshold a_n, which will be specified later. In other words, only elements with absolute values larger than a_n will be treated as significant and counted towards non-zero entries.

Define $\mathcal{B}_n = \left\{\text{At least } \tilde{s}_n \text{ absolute values of } \mathbf{\Gamma} \text{ are greater than } a_n = \frac{\epsilon}{\prod_{d=1}^{D} p_{d,n}}\right\}$, $\mathcal{C}_n = \left\{\mathbf{\Gamma} : ||\mathbf{\Gamma} - \mathbf{\Gamma}^0||_2 > \epsilon\right\}$ and $\mathcal{A}_n = \mathcal{B}_n \cup \mathcal{C}_n$. [18] show that $\Pi_n(\mathcal{A}_n) \to 0$, a.s., when $n \to \infty$ under the assumptions given below:

(a) $\mathbf{\Gamma}_k^0$ assumes a rank-R_0 PARAFAC decomposition, $\mathbf{\Gamma}_k^0 = \sum_{r=1}^{R_0} \boldsymbol{\gamma}_{1,k}^{0(r)} \circ \cdots \circ \boldsymbol{\gamma}_{D,k}^{0(r)}$, for $k = 1, .., m_n$, with $R > R_0$ and $||\boldsymbol{\gamma}_{d,k}^{0(r)}|| < \infty$;

(b) $||\mathbf{\Gamma}_k^0||_0 = s_n$, with $s_n \log(p_n) = o(n)$;

(c) $\tilde{s}_n = O(s_n)$;

(d) $m_n \sum_{d=1}^{D} p_{d,n} \log(p_{d,n}) = o(n)$;

(e) There exists $\lambda_0, \lambda_1 > 0$ s.t. $e_{min}(\boldsymbol{X}_\nabla' \boldsymbol{R}^{-1} \boldsymbol{X}_\nabla) \geq n\lambda_0^2$ and $e_{max}(\boldsymbol{X}_\nabla' \boldsymbol{R}^{-1} \boldsymbol{X}_\nabla) \leq n\lambda_1^2$, for any set $\nabla \subseteq \{1, ..., m_n\}$, where $\boldsymbol{X}_\nabla$ is a submatrix of $\boldsymbol{X} = [\boldsymbol{x}_1' : \cdots : \boldsymbol{x}_n']'$ with columns corresponding to the indices ∇. $\boldsymbol{R}$ is an $n \times n$ matrix with $var(\boldsymbol{E_v}) = \boldsymbol{R}$, $\boldsymbol{E_v} = (E_{1,\boldsymbol{v}}, ..., E_{n,\boldsymbol{v}})'$.

Importantly, the result proves accurate estimation of $\mathbf{\Gamma}$, also ensuring that the true number of nonzero elements in $\mathbf{\Gamma}^0$ and the number of elements identified as nonzero in $\mathbf{\Gamma}$ (i.e., above the threshold a_n) are of the same order. In fact the L_2 metric between $\mathbf{\Gamma}$ and $\mathbf{\Gamma}^0$ is stronger than the KL-divergence metric used in Section 1.1. Similar to Section 1.1, assumptions on $p_{d,n}$ and m_n allow the number of tensor cells to grow much faster than the sample size n without disturbing the desirable posterior consistency of the model.

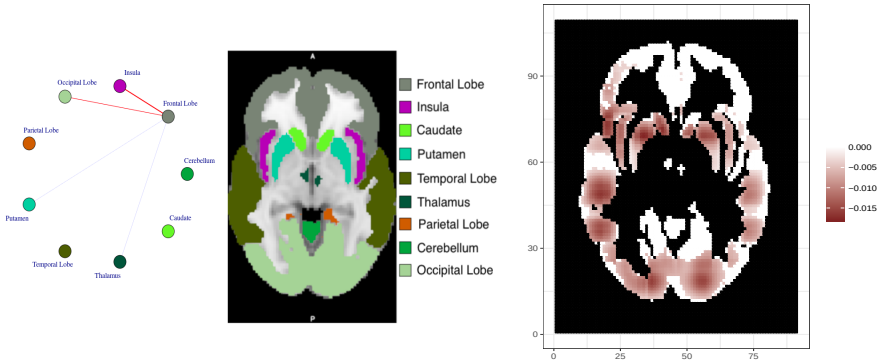
Building on (8), [54] develop the mixed effect tensor response regression model with an application to modeling the joint estimation of brain activation at the voxel level and brain connectivity at the lobe level for multi-subject fMRI studies. Let $\boldsymbol{Y}_{i,g,t}$ be the tensor of observed fMRI data in brain region g for the ith subject at the tth time point. $\boldsymbol{Y}_{i,g,t}$ is observed in the form of a tensor with dimensions $p_{1,g} \times \cdots \times p_{D,g}$. To simultaneously measure activation due to stimulus at voxels in the gth brain region and connectivity among G brain regions, [54] employ an additive mixed effect model with a tensor-valued fMRI response and activation-related predictors $x_{1,i,t}, ..., x_{m,i,t} \in \mathbb{R}$,

$$\mathbf{Y}_{i,g,t} = \mathbf{\Gamma}_{1,g} x_{1,i,t} + \cdots + \mathbf{\Gamma}_{m,g} x_{m,i,t} + d_{i,g} + \mathbf{E}_{i,g,t}, \tag{9}$$

for subject $i = 1, \ldots, n$, in region $g = 1, \ldots, G$, and time $t = 1, \ldots, T$. $\mathbf{\Gamma}_{k,g}$ $(k = 1, ..., m)$ represents activation due to the kth stimulus at the gth brain region, and hence multiway stick breaking priors are independently used for each $\mathbf{\Gamma}_{k,g}$ to determine the nature of activation. Additionally, the connectivity between different regions is ascertained by jointly modeling $d_{i,g}$s with a Gaussian graphical LASSO prior [59] given by,

$$\begin{aligned} &\mathbf{d}_i = (d_{i,1}, .., d_{i,G})' \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma}^{-1}), \;\; i = 1, ..., n, \\ &p(\boldsymbol{\sigma}|\zeta) = C^{-1} \prod_{g<g_1} [DE(\sigma_{gg_1}|\zeta)] \prod_{g=1}^{G} \left[\mathrm{Exp}(\sigma_{gg}|\frac{\zeta}{2}) \right] \mathbf{1}_{\mathbf{\Sigma} \in \mathcal{P}^+}, \end{aligned} \tag{10}$$

where $\mathcal{P}^+$ is the class of all symmetric positive definite matrices and C is a normalization constant. The covariance $\boldsymbol{\sigma} = (\sigma_{gg_1} : g \leq g_1)$ is a vector of upper triangular and diagonal entries of the precision matrix $\mathbf{\Sigma}$. Using properties of the multivariate Gaussian distribution (see **stat05651,stat05654**), a small value of σ_{gg_1} stands for weak connectivity between regions g and g_1, given the other regions. In fact, $\sigma_{gg_1} = 0$ $(g < g_1)$ implies that there is no connectivity between regions g and g_1, given the other regions. In practice, a double exponential prior on the off-diagonal entries will a priori favor shrinkage. An efficient Markov chain Monte Carlo algorithm has been developed for estimation of model parameters, even in presence of high resolution fMRI images, and post burn-in MCMC samples are processed to determine activated voxels in a brain region and connectivity between different brain regions.



(a) Brain connectivity map among lobes (b) Brain activation map in voxels

Figure 6: An example of brain activation and connectivity estimated jointly from the Balloon Analog Risk-Taking Task (BART) Experiment data.

All details related to implementation and inference on (9) can be found in [54].

[54] analyze data collected in a study examining the fMRI scans of individuals undergoing a test which introduces risk-taking scenarios. This study is known as the Balloon Analog Risk-Taking Task (BART) Experiment [50], which requires participants to make active decisions, and whose design has been found to correlate with real-world risk behavior such as alcohol use, cigarette and drug use, gambling, stealing, unsafe sex [36, 35, 37] and trait measures of risk-taking propensity like sensation seeking, trait impulsivity [36] and trait psychopathy [28]. [54], in particular, employed the model (9) for the joint inference on voxel level brain activation and connectivity between lobes related to the risk taking task. Figures 6b shows the estimated map of activated voxels which shows localized activation mainly in the frontal lobe. This localized pattern of activation in the frontal lobe while performing the higher-order processing task of risk assessment is consistent with various previous studies [5]. Figure 6a shows estimates for the significantly nonzero partial correlations between lobes. This figure also indicates that the frontal lobe plays an important role in this task showing nonzero partial correlations with the insula, caudate, putamen, and occipital lobes, placing the frontal lobe at the center of a connective network. This agrees with earlier experiments suggesting that the frontal lobe plays a role in the assessment of risk [42]. We emphasize that the Bayesian model shows rapid convergence of the MCMC chain and efficiently analyzes fMRI data with more than $11,000$ voxels.

2.2 Bayesian Tensor on Tensor Regression Models

In the realm of Bayesian modeling with a tensor response and a tensor predictor, one of the significant earlier contributions comes from [25]. Let $\boldsymbol{Y}_i$ be a tensor of dimension $p_1 \times \cdots \times p_D$ for sample $i = 1, ..., n$, with the corresponding predictor $\boldsymbol{X}_i$ of dimension $q_1 \times \cdots \times q_D$. [25] proposes the multilinear regression model given by

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \times \{\boldsymbol{C}_1, ..., \boldsymbol{C}_D\} + \boldsymbol{E}_i, \tag{11}$$

where $\boldsymbol{C}_1, .., \boldsymbol{C}_D$ are matrices of dimensions $q_1 \times p_1$,...,$q_D \times p_D$ respectively, and $\times$ refers to the Tucker product. The error tensors $\boldsymbol{E}_i$ are i.i.d with dimensions $q_1 \times \cdots \times q_D$, and are assumed to follow a tensor normal distribution (see **stat02360**), implying $vec(\boldsymbol{E}_i) \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a $(\prod_{d=1}^{D} q_d) \times (\prod_{d=1}^{D} q_d)$ dimensional covariance matrix. Estimating so many parameters from an unstructured $\boldsymbol{\Sigma}$ without adding some restrictions on its form appears to be infeasible. As an alternative, a flexible, reduced-parameter covariance model that retains the tensor structure of the data is the tensor normal model [1, 24], which assumes a separable (Kronecker structured) covariance matrix. Specifically, the tensor normal distribution is denoted by $N_{p_1,...,p_D}(\boldsymbol{0}, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_D)$ which essentially implies $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 \otimes \cdots \otimes \boldsymbol{\Sigma}_D$, where $\boldsymbol{\Sigma}_d \in \mathbb{R}^{q_d \times q_d}$. [25] employs matrix normal priors for $\boldsymbol{C}_d | \boldsymbol{\Sigma}_d$ for $d = 1, ..., D$, and inverse wishart priors for $\boldsymbol{\Sigma}_d$ to deliver efficient posterior computation. Although the model is the first of its kind in developing a Bayesian regression framework between two tensors, it comes with a restrictive assumption that $\boldsymbol{Y}_i$ and $\boldsymbol{X}_i$ both have D modes. This assumption is often violated in pertinent neuroscience applications, e.g., in the problem of developing a regression relationship between fMRI and EEG tensors for subjects. Later on, [41] develop another framework that relaxes this assumption.

Let $\boldsymbol{Y}_i$ be the tensor response of dimension $p_1 \times \cdots \times p_{D_1}$ for sample $i = 1, .., n$, with the corresponding tensor predictor $\boldsymbol{X}_i$ of dimension $q_1 \times \cdots \times q_{D_2}$. Let $\boldsymbol{Y}$ be a tensor of dimension $n \times p_1 \times \cdots \times p_{D_1}$, obtained by stacking the $\boldsymbol{Y}_i$'s over $i = 1, .., n$. A tensor $\boldsymbol{X}$ of dimension $n \times q_1 \times \cdots \times q_{D_2}$ is created in a similar fashion by stacking the $\boldsymbol{X}_i$s. [41] propose

a regression framework of $\boldsymbol{X}$ on $\boldsymbol{Y}$ as follows

$$\boldsymbol{Y} = \langle \boldsymbol{X}, \boldsymbol{B} \rangle_L + \boldsymbol{E}, \;\; E_{(v_1,\ldots,v_{D_1+1})} \sim N(0, \sigma^2), \tag{12}$$

$1 \leq v_1 \leq n$, $1 \leq v_2 \leq p_1$,...,$1 \leq v_{D_1+1} \leq p_{D_1}$. $\boldsymbol{B}$ is the tensor coefficient of dimension $p_1 \times \cdots \times p_{D_1} \times q_1 \times \cdots \times q_{D_2}$ and $\langle \cdot, \cdot \rangle_L$ represents the contracted tensor product. When $D_2 = 1$, i.e., $\boldsymbol{X}$ becomes a vector, (12) reduces to (8). Reshaping each $\boldsymbol{Y}_i$ and $\boldsymbol{X}_i$ to vectors of dimensions $\prod_{d_1=1}^{D_1} p_{d_1}$ and $\prod_{d_2=1}^{D_2} q_{d_2}$, (12) can be rewritten in the standard multivariate linear regression form given below,

$$\boldsymbol{Y}^{(1)} = \boldsymbol{X}^{(1)} \boldsymbol{B}^{(1)} + \boldsymbol{E}^{(1)},$$

where $\boldsymbol{Y}^{(1)}$ and $\boldsymbol{X}^{(1)}$ are $n \times \prod_{d_1=1}^{D_1} p_{d_1}$ and $n \times \prod_{d_2=1}^{D_2} q_{d_2}$ matrices obtained by stacking reshaped vectors over all samples. $\boldsymbol{B}^{(1)}$ is a $\prod_{d_1=1}^{D_1} p_{d_1} \times \prod_{d_2=1}^{D_2} q_{d_2}$ matrix whose columns are obtained by vectorizing the first D_1 modes of $\boldsymbol{B}$, and rows are obtained by vectorizing the last D_2 modes of $\boldsymbol{B}$.

Estimating $\boldsymbol{B}$ involves $\prod_{d_1=1}^{D_1} p_{d_1} \times \prod_{d_2=1}^{D_2} q_{d_2}$ parameters. Hence [41] have adopted the low-rank PARAFAC decomposition for $\boldsymbol{B}$. To estimate $\boldsymbol{B}$ in a Bayesian framework, [41] have proposed

$$\pi(\boldsymbol{B}) \propto \begin{cases} \exp(-\frac{\tilde{\lambda}}{2\sigma^2}||\boldsymbol{B}||^2) & \text{if } rank(\boldsymbol{B}) \leq R \\ 0 & o.w. \end{cases}$$

This specification leads to $-\log(\pi(\boldsymbol{B}|\boldsymbol{Y}, \boldsymbol{X})) = \frac{||\boldsymbol{Y} - \langle \boldsymbol{X}, \boldsymbol{B} \rangle_L||^2}{2\sigma^2} + \frac{\tilde{\lambda}}{2\sigma^2}||\boldsymbol{B}||^2$, which resembles a penalized likelihood framework (see **stat05934**). Here $\tilde{\lambda}$ acts as a tuning parameter, which is optimized within the Bayesian estimation of the model. Notice that the inferential framework in [41] is semi-Bayesian since the tuning parameter $\tilde{\lambda}$ is chosen using a cross-validation procedure. Further, in contrast with the multiway shrinkage priors discussed earlier, the prior in [41] imposes a similar degree of shrinkage on every tensor cell. Later, [6] extend the modeling framework in [41] to develop dynamic tensor on tensor regression models and use M-DGDP prior as a more effective tool to impose shrinkage on the cells of

tensor coefficients.

If $\boldsymbol{Y}_t$ represents the response tensor of dimension $p_1 \times \cdots \times p_{D_1}$ at time t, and $\boldsymbol{X}_t$ is the tensor predictor at time t of dimension $q_1 \times \cdots \times q_{D_2}$, the dynamic tensor on tensor regression model proposed in [6] is given by

$$\boldsymbol{Y}_t = \sum_{j=1}^{q} \langle \tilde{\boldsymbol{B}}_j, \boldsymbol{Y}_{t-j} \rangle_L + \langle \tilde{\boldsymbol{A}}, \boldsymbol{X}_t \rangle_L + \boldsymbol{E}_t, \;\; \boldsymbol{E}_t \sim N_{p_1,..,p_{D_1}}(\boldsymbol{0}, \boldsymbol{\Sigma}_1, .., \boldsymbol{\Sigma}_{D_1}), \tag{13}$$

where $N_{p_1,..,p_{D_1}}(\boldsymbol{0}, \boldsymbol{\Sigma}_1, .., \boldsymbol{\Sigma}_{D_1})$ represents a D_1 dimensional tensor normal distribution with $\boldsymbol{\Sigma}_j$ of dimension $p_j \times p_j$. Here $\tilde{\boldsymbol{B}}_j$, $j = 1, .., q$, and $\tilde{\boldsymbol{A}}$ are tensors of dimensions $p_1 \times \cdots \times p_{D_1} \times p_1 \times \cdots \times p_{D_1}$ and $p_1 \times \cdots \times p_{D_1} \times q_1 \times \cdots \times q_{D_2}$, respectively. Using tensor calculus (ref), equation (13) can be rewritten as

$$\boldsymbol{Y}_t = \sum_{j=1}^{q} \boldsymbol{B}_j \bar{\times}_{D_1+1} vec(\boldsymbol{Y}_{t-j}) + \boldsymbol{A} \bar{\times}_{D_2+1} vec(\boldsymbol{X}) + \boldsymbol{E}_t, \;\; \boldsymbol{E}_t \sim N_{p_1,..,p_{D_1}}(\boldsymbol{0}, \boldsymbol{\Sigma}_1, .., \boldsymbol{\Sigma}_{D_1}), \tag{14}$$

where $\boldsymbol{B}_j$ is of dimension $p_1 \times \cdots \times p_{D_1} \times \prod_{d_1=1}^{D_1} p_{d_1}$ and $\boldsymbol{A}$ is of dimension $q_1 \times \cdots \times a_{D_2} \times \prod_{d_2=1}^{D_2} q_{d_2}$. The model presented in (14) emerges as a generalization of several econometric models, such as the seemingly unrelated regression (SUR) model [65], the VARX and panel VAR model [8, 9] and the vector error correlation model (VECM) [10, 51], to name a few. In order to achieve parsimony in estimating tensor coefficients, [6] consider a rank-R PARAFAC decomposition of the tensor coefficients, as discussed in earlier sections. The choice of the prior distribution on the PARAFAC margins is crucial for recovering the sparsity pattern of the coefficient tensor and for the efficiency of the inference. To this end, [6] follow the proposal of the multiway shrinkage prior in [19] (reviewed in Section 1.1) on tensor coefficients. As discussed before, the global-local structure of the M-DGDP prior imposes a careful shrinkage on cell coefficients and offers better scalability properties in high-dimensional settings.

3 Conclusion

Motivated by applications in neuroscience, genomics, social science, chemometrics and other physical and biological sciences, there has been an increasing interest in the use of tensor-valued objects within a regression setting. As opposed to summarizing tensors, us-

ing tensor valued objects within a regression framework reaps important modeling benefits. These advantages include efficient computation, parsimony and importantly, accurate inference due to accounting for the neighborhood structure in a tensor. As a consequence, this decade has witnessed rapid developments in statistical methods for tensor regression. In this review article, we have focused on Bayesian methods related to tensor regression. Although the literature on tensor regression is predominantly frequentist, Bayesian methods have made inroads into the literature in the last few years, with some additional benefits. Unlike many of the classical frequentist techniques, Bayesian models allow for flexibility, mainly via their choices of carefully structured prior distributions on tensor valued coefficients that provide data dependent shrinkage of tensor coefficients and model-based learning of tuning parameters. Additionally, they offer uncertainty in parametric and predictive inferences, and can be easily implemented via full MCMC techniques. In this review article, we have divided methods according to the objective of the analysis. First, we have described tensor regression techniques with a tensor predictor and a scalar response. We have first reviewed parametric tensor regression models, and then discussed nonparametric tensor regression models as well. The article then reviewed tensor regression models with a tensor response and a vector- or tensor-valued predictor. We mention applications in each case, mainly from neuroscience, relevant to the methodology reviewed. We also offer a brief discussion on available theoretical results regarding posterior convergence of these models.

Even though the Bayesian tensor regression methods generally exhibit rapid convergence in model fitting using MCMC, high dimensionality of the tensor data may limit the practical usage of Bayesian methodology. For example, fMRI experiments produce massive amount of correlated data over millions of brain voxels, which is difficult to be dealt with using Bayesian tensor regression. Such a computation issue is not unique to Bayesian methods for tensor regression, since frequentist tensor regression techniques also employ cross validation steps to estimate tuning parameters which are computationally costly. As noted before, tensor regression methods implicitly incorporate spatial information in the data, though combining tensor regression directly with spatial modeling techniques may draw additional inferential advantages. EEG or fMRI experiments are examples which produce high resolution spatially correlated data ready to be mined with Bayesian methodologies. The latter topic is quite

recent and is certainly an important area of research. In fact, Bayesian methods for big spatial and structured data still constitute a very active area of research and many more important contributions are expected. Finally, it needs to be mentioned that although there are open source codes available on Bayesian tensor regression methods, to the best of our knowledge, an open source user-friendly software package is still unavailable. More effort in disseminating codes and software in the public domain for tensor regression methods in the near future would be greatly beneficial.

4 Acknowledgement

The author is partially supported by the Office of Naval Research, award no. N00014-18-2741, and the National Science Foundation, grant DMS-1854662.

References

[1] D. Akdemir and A. K. Gupta. Array variate random variables with multiway kronecker delta covariance matrix structure. *J. Algebr. Stat*, 2(1):98–113, 2011.

[2] M. Amewou-Atisso, S. Ghosal, J. K. Ghosh, and R. Ramamoorthi. Posterior consistency for semi-parametric regression problems. *Bernoulli*, 9(2):291–312, 2003.

[3] A. Armagan, D. B. Dunson, and J. Lee. Generalized double Pareto shrinkage. *Statistica Sinica*, 23(1):119–143, 2013.

[4] A. Armagan, D. B. Dunson, J. Lee, W. U. Bajwa, and N. Strawn. Posterior consistency in linear models under shrinkage priors. *Biometrika*, 100(4):1011–1018, 2013.

[5] M. G. Berman, J. Jonides, and D. E. Nee. Studying mind and brain with fmri. *Social cognitive and affective neuroscience*, 1(2):158–161, 2006.

[6] M. Billio, R. Casarin, S. Kaufmann, and M. Iacopini. Bayesian dynamic tensor regression. *University Ca'Foscari of Venice, Dept. of Economics Research Paper Series No*, 13, 2018.

[7] B. S. Caffo, C. M. Crainiceanu, G. Verduzco, S. Joel, S. H. Mostofsky, S. S. Bassett, and

J. J. Pekar. Two-stage decompositions for the analysis of functional connectivity for fmri with application to alzheimer's disease risk. *NeuroImage*, 51(3):1140–1149, 2010.

[8] F. Canova and M. Ciccarelli. Forecasting and turning point predictions in a Bayesian panel var model. *Journal of Econometrics*, 120(2):327–359, 2004.

[9] F. Canova and M. Ciccarelli. Estimating multicountry var models. *International economic review*, 50(3):929–959, 2009.

[10] F. Canova, M. Ciccarelli, and E. Ortega. Similarities and convergence in G-7 cycles. *Journal of Monetary economics*, 54(3):850–878, 2007.

[11] R. D. Cook, B. Li, and F. Chiaromonte. Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, pages 927–960, 2010.

[12] F. De Martino, A. W. De Borst, G. Valente, R. Goebel, and E. Formisano. Predicting eeg single trial responses with simultaneous fMRI and relevance vector machine regression. *Neuroimage*, 56(2):826–836, 2011.

[13] S. Friedland and L.-H. Lim. Computational complexity of tensor nuclear norm. *arXiv preprint arXiv:1410.6072*, 2014.

[14] J. Gertheiss, A. Maity, and A.-M. Staicu. Variable selection in generalized functional linear models. *Stat*, 2(1):86–101, 2013.

[15] S. Ghosal, J. K. Ghosh, and R. Ramamoorthi. Posterior consistency of dirichlet mixtures in density estimation. *Ann. Statist*, 27(1):143–158, 1999.

[16] J. Goldsmith, J. Bobb, C. M. Crainiceanu, B. Caffo, and D. Reich. Penalized functional regression. *Journal of Computational and Graphical Statistics*, 20(4):830–851, 2011.

[17] R. Guhaniyogi. Convergence rate of Bayesian supervised tensor modeling with multiway shrinkage priors. *Journal of Multivariate Analysis*, 160:157–168, 2017.

[18] R. Guhaniyogi and D. Spencer. Bayesian tensor response regression with an application to brain activation studies. Technical report, Technical report, UCSC. 2, 13, 2018.

[19] R. Guhaniyogi, S. Qamar, and D. B. Dunson. Bayesian tensor regression. *The Journal of Machine Learning Research*, 18(1):2733–2763, 2017.

[20] S. Guillas and M.-J. Lai. Bivariate splines for spatial functional regression models. *Journal of Nonparametric Statistics*, 22(4):477–497, 2010.

[21] W. Hackbusch. *Tensor spaces and numerical tensor calculus*, volume 42. Springer Science & Business Media, 2012.

[22] R. A. Harshman and M. E. Lundy. PARAFAC: Parallel factor analysis. *Computational Statistics & Data Analysis*, 18(1):39–72, 1994.

[23] J. Hastie Trevor and J. Tibshirani Robert. Generalized additive models. vol. 43, 1990.

[24] P. D. Hoff. Hierarchical multilinear models for multiway data. *Computational Statistics & Data Analysis*, 55(1):530–543, 2011.

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

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

[27] M. Hou, Y. Wang, and B. Chaib-draa. Online local Gaussian process for tensor-variate regression: Application to fast reconstruction of limb movements from brain signal. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5490–5494. IEEE, 2015.

[28] M. K. Hunt, D. R. Hopko, R. Bare, C. Lejuez, and E. Robinson. Construct validity of the balloon analog risk task (BART) associations with psychopathy and impulsivity. *Assessment*, 12(4):416–428, 2005.

[29] M. Imaizumi and K. Hayashi. Doubly decomposing nonparametric tensor regression. In *International Conference on Machine Learning*, pages 727–736, 2016.

[30] G. M. James, J. Wang, J. Zhu, et al. Functional linear regression that?s interpretable. *The Annals of Statistics*, 37(5A):2083–2108, 2009.

[31] S. Kalus, P. G. Sämann, and L. Fahrmeir. Classification of brain activation via spatial Bayesian variable selection in fMRI regression. *Advances in Data Analysis and Classification*, 8(1):63–83, 2014.

[32] S. Kim, P. Smyth, and H. Stern. A nonparametric Bayesian approach to detecting spatial activation patterns in fMRI data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 217–224. Springer, 2006.

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

[34] N. Lazar. *The statistical analysis of functional MRI data*. Springer Science & Business Media, 2008.

[35] C. Lejuez, W. M. Aklin, H. A. Jones, J. B. Richards, D. R. Strong, C. W. Kahler, and J. P. Read. The balloon analogue risk task (BART) differentiates smokers and nonsmokers. *Experimental and clinical psychopharmacology*, 11(1):26, 2003.

[36] C. W. Lejuez, J. P. Read, C. W. Kahler, J. B. Richards, S. E. Ramsey, G. L. Stuart, D. R. Strong, and R. A. Brown. Evaluation of a behavioral measure of risk taking: the balloon analogue risk task (bart). *Journal of Experimental Psychology: Applied*, 8(2): 75, 2002.

[37] C. W. Lejuez, W. M. Aklin, M. J. Zvolensky, and C. M. Pedulla. Evaluation of the balloon analogue risk task (BART) as a predictor of adolescent real-world risk-taking behaviours. *Journal of adolescence*, 26(4):475–479, 2003.

[38] L. Li and X. Zhang. Parsimonious tensor response regression. *Journal of the American Statistical Association*, 112(519):1131–1146, 2017.

[39] X. Li, D. Xu, H. Zhou, and L. Li. Tucker tensor regression and neuroimaging analysis. *Statistics in Biosciences*, 10(3):520–545, 2018.

[40] M. A. Lindquist. The statistical analysis of fmri data. *Statistical science*, 23(4):439–464, 2008.

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

[42] L. Miller and B. Milner. Cognitive risk-taking after frontal or temporal lobectomy-II. The synthesis of phonemic and semantic information. *Neuropsychologia*, 23(3):371–379, 1985.

[43] T. Park and G. Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[44] G. Rabusseau and H. Kadri. Low-rank regression with tensor responses. In *Advances in Neural Information Processing Systems*, pages 1867–1875, 2016.

[45] A. Ramasamy, D. Trabzuni, S. Guelfi, V. Varghese, C. Smith, R. Walker, T. De, J. Hardy, M. Ryten, M. E. Weale, et al. Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature neuroscience*, 17(10):1418, 2014.

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

[47] P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

[48] P. T. Reiss and R. T. Ogden. Functional generalized linear models with images as predictors. *Biometrics*, 66(1):61–69, 2010.

[49] P. T. Reiss, L. Huo, Y. Zhao, C. Kelly, and R. T. Ogden. Wavelet-domain regression and predictive inference in psychiatric neuroimaging. *The annals of applied statistics*, 9 (2):1076, 2015.

[50] T. Schonberg, C. R. Fox, J. A. Mumford, E. Congdon, C. Trepel, and R. A. Poldrack. Decreasing ventromedial prefrontal cortex activity during sequential risk-taking: an fmri investigation of the balloon analog risk task. *Frontiers in neuroscience*, 6:80, 2012.

[51] P. Schotman and H. K. Van Dijk. A Bayesian analysis of the unit root in real exchange rates. *Journal of Econometrics*, 49(1-2):195–238, 1991.

[52] M. Signoretto, L. De Lathauwer, and J. A. Suykens. Learning tensors in reproducing kernel Hilbert spaces with multilinear spectral penalties. *arXiv preprint arXiv:1310.4977*, 2013.

[53] E. R. Sowell, P. M. Thompson, S. E. Welcome, A. L. Henkenius, A. W. Toga, and B. S. Peterson. Cortical abnormalities in children and adolescents with attention-deficit hyperactivity disorder. *The Lancet*, 362(9397):1699–1707, 2003.

[54] D. Spencer, R. Guhaniyogi, and R. Prado. Bayesian mixed effect sparse tensor response regression model with joint estimation of activation and connectivity. *arXiv preprint arXiv:1904.00148*, 2019.

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

[56] T. Suzuki. Convergence rate of bayesian tensor estimator and its minimax optimality. In *International Conference on Machine Learning*, pages 1273–1282, 2015.

[57] L. R. Tucker. The extension of factor analysis to three-dimensional matrices. *Contributions to mathematical psychology*, 110119, 1964.

[58] E. M. Valera, S. V. Faraone, K. E. Murray, and L. J. Seidman. Meta-analysis of structural imaging findings in attention-deficit/hyperactivity disorder. *Biological psychiatry*, 61(12):1361–1369, 2007.

[59] H. Wang. Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, 7(4):867–886, 2012.

[60] X. Wang, H. Zhu, and A. D. N. Initiative. Generalized scalar-on-image regression models via total variation. *Journal of the American Statistical Association*, 112(519):1156–1168, 2017.

[61] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar. Cdnet 2014: an expanded change detection benchmark dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 387–394, 2014.

[62] L. Xu, T. D. Johnson, T. E. Nichols, and D. E. Nee. Modeling inter-subject variability in fmri activation location: a Bayesian hierarchical spatial model. *Biometrics*, 65(4): 1041–1051, 2009.

[63] R. Yu and Y. Liu. Learning from multiway data: Simple and efficient tensor regression. In *International Conference on Machine Learning*, pages 373–381, 2016.

[64] Z. Yu, R. Prado, E. B. Quinlan, S. C. Cramer, and H. Ombao. Understanding the impact of stroke on brain motor function: a hierarchical Bayesian approach. *Journal of the American Statistical Association*, 111(514):549–563, 2016.

[65] A. Zellner. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association*, 57(298):348–368, 1962.

[66] L. Zhang, M. Guindani, and M. Vannucci. Bayesian models for functional magnetic resonance imaging data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(1):21–41, 2015.

[67] Q. Zhao, G. Zhou, L. Zhang, and A. Cichocki. Tensor-variate Gaussian processes regression and its application to video surveillance. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1265–1269. IEEE, 2014.

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