Vol. 00 (0000) 1
DOI: 0000

Bayesian Tensor Response Regression With an Application to Brain Activation Studies

Rajarshi Guhaniyogi[1] **and Daniel Spencer**[1]

[1]*Department of Statistics, Baskin School of Engineering, 1156 High Street Santa Cruz, CA 95060*
e-mail: rguhaniy@ucsc.edu daspence@ucsc.edu

Abstract: This article proposes a novel Bayesian implementation of regression with multi-dimensional array (tensor) response on scalar covariates. The recent emergence of complex datasets in various disciplines presents a pressing need to devise regression models with a tensor valued response. This article considers one such application of detecting neuronal activation in fMRI experiments in presence of tensor valued brain images and scalar predictors. The overarching goal in this application is to identify spatial regions (voxels) of a brain activated by an external stimulus. In such and related applications, we propose to regress responses from all cells (or voxels in brain activation studies) together as a tensor response on scalar predictors, accounting for the structural information inherent in the tensor response. To estimate model parameters with proper cell specific shrinkage, we propose a novel *multiway stick breaking shrinkage prior* distribution on tensor structured regression coefficients, enabling identification of cells which are related to the predictors. The major novelty of this article lies in the theoretical study of the contraction properties for the proposed shrinkage prior in the tensor response regression when the number of cells grows faster than the sample size. Specifically, estimates of tensor regression coefficients are shown to be asymptotically concentrated around the true sparse tensor in L_2-sense under mild assumptions. Various simulation studies and analysis of a brain activation data empirically verify desirable performance of the proposed model in terms of estimation and inference on cell-level parameters.

Keywords and phrases: brain activation, BOLD predictor, fMRI studies, multiway stick breaking shrinkage prior, posterior consistency, tensor response.

1. Introduction

Of late, neuroscience or related applications routinely encounter regression scenarios involving a multidimensional array or tensor structured response and scalar predictors. An important motivating example occurs in single-subject Functional MRI (fMRI) studies to detect localized regions where neuronal activation takes place in presence of external stimuli (e.g., during a task). During the course of an fMRI experiment, the three dimensional brain image space is divided into a large number of rectangular cells, also referred to as "voxels". A series of brain images are acquired over multiple time points across all voxels

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 using a regression framework involving the tensor response and task related predictors. Similarly, in electroencephalography (EEG) studies voltage values are measured from numerous electrodes placed on 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. All these applications involve a response tensor $\boldsymbol{Y}_t \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ and a vector of predictors $\boldsymbol{x}_t \in \mathbb{R}^m$ at time t respectively, with an objective to understand the cells in $\boldsymbol{Y}_t$ influenced by the changes in $\boldsymbol{x}_t$. Although the tensor response regression framework is motivated by aforementioned neuroimaging studies, the proposed methodology equally applies to a variety of scientific applications, including chemometrics (Bro, 2006), psychometrics (Kiers and Mechelen, 2001) and relational data (Gerard and Hoff, 2015), among others, where tensor valued responses are collected routinely.

Rather than analyzing cells in a tensor response together, the most popular mass univariate analysis (MUA) fits a regression model at each cell in the tensor response independently of the others and calculates the test statistic corresponding to each cell to identify if the response is significantly associated with a predictor in that cell, accounting for multiple testing corrections (Penny et al., 2011; Friston et al., 1995; Genovese et al., 2002). MUA is conceptually simple and computationally efficient, though it fails to accommodate spatial associations across cells in the tensor response. Additionally, neuroimaging data are usually pre-processed using a kernel convolution based spatial smoothing approach. Performing MUA on pre-smoothed data may result in inaccurate estimation and testing of the covariate effects (Chumbley and Friston, 2009; Li et al., 2011). More principled approaches vectorize the tensor response to construct a multivariate vector response regression. Some notable structures employed to estimate parameters in the multivariate vector response regression include sparse regressions with various penalties incorporating correlated response variables (Similä and Tikka, 2007; Peng et al., 2010), reduced-rank regressions (Yuan et al., 2007; Chen et al., 2013) and sparse reduced-rank regressions (Chen and Huang, 2012). While these methods view tensor response as a high dimensional vector without any spatial association among its cells, our goal is to incorporate spatial information in the multidimensional tensor into the proposed model.

To this end, sophisticated approaches include adaptive multiscale smoothing methods and spatially varying coefficient (SVC) models. The former estimates parameters by building iteratively increasing neighbors around each cell and combining observations within the neighbors with weights (Li et al., 2011). The SVC models add spatial components in the cell by cell regression that account for the spatial correlations between cell (Zhang et al., 2015, 2014; Descombes et al., 1998; Zhu et al., 2014). There is a parallel literature to model spatial de-

pendence among regression coefficients induced by Markov random fields (MRF) (Smith and Fahrmeir, 2007). These approaches introduce distinct parameters for different cell specific regressions and propose to model them jointly. For a tensor response of dimensions $p_1 \times \cdots \times p_D$, where p_1,.. p_D are moderately large, such strategies lead to the joint modeling of *at least* $\prod_{i=1}^{D} p_i$ parameters, which may turn out to be computationally challenging.

Recently, Li and Zhang (2015) propose a novel approach of regressing the tensor variate response on scalar predictors, where recently developed *envelope* technique by Cook et al. (2010) is employed to yield point estimates of the parameters. Subsequently, Sun and Li (2017) provide convergence rates of the frequentist penalized regression approaches with a tensor response and vector predictors. This approach proposes low rank decomposition of the tensor coefficient and introduces multiple constraints on the parameter space. While such constraints can be easily accommodated by frequentist optimization algorithms, they offer a steep challenge for Bayesian implementation. Additionally, frequentist optimization frameworks are dependent on tuning parameters (e.g., the envelope dimensions in Li and Zhang (2015)), with choices for these parameters being sensitive to the tensor dimensions and the signal-to-noise ratio (degree of sparsity).

In the same vein as Li and Zhang (2015), we propose a regression scenario with tensor response $\boldsymbol{Y}_t$ and predictors $\boldsymbol{x}_t$, referred to as the *tensor response regression* (TRR). The coefficient corresponding to each predictor in the vector $\boldsymbol{x}_t$ is a tensor, and is assumed to possess a "low rank" PARAFAC/CP (defined in Section 2.1) decomposition. For the Bayesian implementation, we employ a novel *multiway stick breaking shrinkage prior* distribution to shrink the cells of the tensor coefficient corresponding to unimportant voxels close to zero while maintaining accurate estimation and uncertainty of cell coefficients related to important voxels. Our framework is, to the best of our knowledge, the first Bayesian framework for regressing a tensor response on scalar predictors. Additionally, TRR retains the tensor structure of the response to implicitly preserve correlations between cells and yet substantially reduces the number of parameters using the CP decomposition to accrue computational benefits. The TRR framework with the multiway stick breaking prior gives rise to model-based shrinkage towards a "low rank" solution for the tensor coefficient, with a carefully constructed shrinkage prior that naturally induces sparsity within and across ranks for the tensor coefficient and results in identification of important cells in the tensor related to a predictor. In addition, there is a strong need for uncertainty quantification for parametric estimates, especially when the tensor dimension far exceeds the sample size, or the signal to noise ratio is low, motivating the Bayesian TRR (BTRR) approach.

There is a recent literature on regressing a tensor covariate on a scalar response (Guhaniyogi et al., 2017; Zhou et al., 2013; Zhou and Li, 2014) that focuses on identifying voxels in the tensor which are related to the response. In contrast, we flip the role and regress a tensor response on scalar predictors. Our approach differs from the existing frequentist and Bayesian tensor modeling approaches (Gerard and Hoff, 2015; Dunson and Xing, 2009) as we offer a

supervised tensor regression framework that accommodates scalar predictors.

One important contribution of this article remains proving posterior consistency for the proposed BTRR model with the multiway stick breaking shrinkage prior. Theory of posterior contraction for high dimensional regression models has gained traction lately, though the literature is less developed in shrinkage priors compared to point-mass priors. For example, Castillo et al. (2012) and Belitser and Nurushev (2015) have established posterior concentration and variable selection properties for certain point-mass priors in the many normal-means model. The latter article also establishes coverage of Bayesian credible sets. Results on posterior concentration and variable selection in high dimensional linear models are also established by Castillo et al. (2015a) and Martin et al. (2017) for certain point-mass priors. In contrast, Armagan et al. (2013b) show posterior consistency in the linear regression model with shrinkage priors for low-dimensional settings where the number of covariates *does not* exceed the number of observations. Using direct calculations, Van Der Pas et al. (2014) show that the posterior based on the horseshoe prior concentrates at the optimal rate for the many normal-mean problem. Song and Liang (2017) and Wei and Ghosal (2017) consider a general class of continuous shrinkage priors and obtain posterior contraction rates in ordinary high dimensional linear regression models and logistic regression models respectively, depending on the concentration and tail properties of the density of the continuous shrinkage prior. In contrast, the study of posterior contraction properties for tensor regression models in the Bayesian paradigm has been given far too less attention. A recent article by Guhaniyogi (2017) is of interest in this regard. Developing theory for tensor response regression models is faced with two major challenges. While high dimensional regression models directly impose a well investigated shrinkage prior on the predictor coefficients, BTRR imposes shrinkage priors on margins of the CP decomposition of tensor coefficients. As a result, the prior distribution on voxel level elements of the tensor coefficient is difficult to deal with. Additionally, in typical applications, the dimensions of tensor coefficients are much larger than the sample size. Both of these present obstacles which we overcome in this work. We also emphasize that the posterior contraction of tensor regression in Guhaniyogi (2017) is shown for the Kullback-Leibler neighborhood. In contrast, Bayesian tensor response regression develops a much stronger result with L_2-neighborhood around the true tensor coefficient.

The remainder of the article flows as following. Section 2 introduces the model and describes prior distributions on the parameters. Section 3 describes results on posterior consistency of the proposed model. Section 4 details out the posterior computations. Section 5 and 6 show performance of the proposed model through simulation studies and brain activation data analysis. Section 7 concludes the paper.

2. Framework & Model

2.1. Basic Notation

Let $\boldsymbol{\gamma}_1 = (\gamma_{11}, \ldots, \gamma_{1p_1})'$ and $\boldsymbol{\gamma}_2 = (\gamma_{21}, \ldots, \gamma_{2p_2})'$ be $p_1 \times 1$ and $p_2 \times 1$ vectors, respectively. The vector outer product $\boldsymbol{\gamma}_1 \circ \boldsymbol{\gamma}_2$ is a $p_1 \times p_2$ array with (i,j)-th entry $\gamma_{1i}\,\gamma_{2j}$. A D-way outer product between vectors $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jp_j})$, $1 \leq j \leq D$, is a $p_1 \times \cdots \times p_D$ dimensional array denoted by $\boldsymbol{\Gamma} = \boldsymbol{\gamma}_1 \circ \boldsymbol{\gamma}_2 \circ \cdots \circ \boldsymbol{\gamma}_D$ with entries $\boldsymbol{\Gamma}_{i_1,\ldots,i_D} = \prod_{j=1}^{D} \gamma_{ji_j}$. Define a $vec(\boldsymbol{\Gamma})$ operator as one that stacks elements of this tensor into a column vector of length $\prod_{j=1}^{D} p_j$. From the definition of outer products, it is easy to see that $vec(\boldsymbol{\gamma}_1 \circ \boldsymbol{\gamma}_2 \circ \cdots \circ \boldsymbol{\gamma}_D) = \boldsymbol{\gamma}_D \otimes \cdots \otimes \boldsymbol{\gamma}_1$. A tensor $\boldsymbol{\Gamma} \in \otimes_{j=1}^{D} \Re^{p_j}$ is known as a D-way tensor. A mode-k fiber of a D-way tensor is obtained by fixing all dimensions of a tensor except the k-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 k-th mode vector product of a D-way tensor $\boldsymbol{\Gamma}$ and vector $\boldsymbol{a} \in \mathcal{R}^{p_k}$, denoted by $\boldsymbol{\Gamma} \bar{\times}_k \boldsymbol{a}$, is a tensor of the order of $p_1 \times \cdots \times p_{k-1} \times p_{k+1} \times \cdots \times p_D$, whose elements are the inner product of each mode-k fiber of $\boldsymbol{\Gamma}$ with $\boldsymbol{a}$.

A D-way tensor $\boldsymbol{\Gamma} \in \otimes_{j=1}^{D} \Re^{p_j}$ assumes a rank-R PARAFAC decomposition (Kiers, 2000) if $\boldsymbol{\Gamma}$ can be expressed as

$$\boldsymbol{\Gamma} = \sum_{r=1}^{R} \boldsymbol{\gamma}_1^{(r)} \circ \cdots \circ \boldsymbol{\gamma}_D^{(r)} \tag{2.1}$$

where $\boldsymbol{\gamma}_j^{(r)}$ is a p_j dimensional column vector as before, for $1 \leq j \leq D$ and $1 \leq r \leq R$. Terminology refers to these vectors as 'margins' of a particular rank. The PARAFAC decomposition is generally preferred in most modeling applications involving tensors, both in terms of interpretability (i.e., invariance to the order of summation) and from a computational tractability point of view (Kolda and Bader, 2009).

2.2. Model framework

Let $\boldsymbol{Y}_t = ((Y_{t,\boldsymbol{v}}))_{v_1,\ldots,v_D=1}^{p_1,\ldots,p_D} \in \otimes_{j=1}^{D} \Re^{p_j}$ denote a tensor valued response at time t, where $\boldsymbol{v} = (v_1, \ldots, v_D)'$ represents the position of voxel $\boldsymbol{v}$ in the D dimensional array of voxels. Let $\boldsymbol{x}_t = (x_{1,t}, \ldots, x_{m,t})' \in \mathcal{X} \subset \Re^m$ be the m-dimensional measured vector predictor. Assuming that both response $\boldsymbol{Y}_t$ and predictors $\boldsymbol{x}_t$ are centered around their respective means, the proposed tensor response regression model of $\boldsymbol{Y}_t$ on $\boldsymbol{x}_t$ is given by

$$\boldsymbol{Y}_t = \boldsymbol{\Gamma}_1 x_{1,t} + \cdots + \boldsymbol{\Gamma}_m x_{m,t} + \boldsymbol{E}_t, \tag{2.2}$$

for $t = 1, \ldots, T$. $\boldsymbol{\Gamma}_k \in \otimes_{j=1}^{D} \Re^{p_j}$, $k = 1, .., m$ is the tensor coefficient corresponding to the predictor $x_{k,t}$. To account for the temporal correlation of the response tensor, the error tensor $\boldsymbol{E}_t \in \otimes_{j=1}^{D} \Re^{p_j}$ is assumed to follow a componentwise

AR(1) structure, $vec(\boldsymbol{E}_t) = \kappa\, vec(\boldsymbol{E}_{t-1}) + vec(\boldsymbol{\eta}_t)$, where $\kappa \in (-1, 1)$ is the autocorrelation coefficient and $\boldsymbol{\eta}_t \in \otimes_{j=1}^{D} \Re^{p_j}$ with each cell in $\boldsymbol{\eta}_t$ following $N(0, \sigma^2/(1-\kappa^2))$. This ensures both computational simplicity and stationarity in the AR(1) structure.

Naive voxel by voxel regression of $Y_{t,v}$ on $\boldsymbol{x}_t$ requires introducing m regression parameters per voxel, hence a total of $m\prod_{j=1}^{D} p_j$ parameters, resulting in an ultra-high dimensional modeling pursuit, and fails to incorporate tensor structural information into the estimation procedure. This necessitates imposing a sufficiently expressive structure on $\boldsymbol{\Gamma}_k$ which simultaneously achieves a large dimensionality reduction. We propose flexible 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 $\boldsymbol{\gamma}_{j,k}^{(r)} = (\gamma_{j,k,1}^{(r)}, ..., \boldsymbol{\gamma}_{j,k,p_j}^{(r)})'$ is a p_j dimensional vector, $1 \leq r \leq R$, $1 \leq j \leq D$ and $k = 1, .., m$.

A few remarks on (2.2) are in order. First, since we deal with modeling the linear predictor part of the model, our framework can easily be extended to a GLM set up. Second, the formulation also assumes easy extensions to settings with a more complicated spatio-temporal correlation structure in $\boldsymbol{E}_t$. Additionally, PARAFAC decomposition reveals that the cell level parameters are nonlinear functions of the tensor margins $\boldsymbol{\gamma}_{k,j}^{(r)}$. Careful choice of prior distributions on the tensor margins implicitly imposes correlations among voxels and facilitates identifying significantly nonzero cells in $\boldsymbol{\Gamma}_k$.

Imposing this additional rank-R PARAFAC structure on $\boldsymbol{\Gamma}_k$ remarkably reduces the total number of parameters in the model from $m\prod_{j=1}^{D} p_j$ to $Rm\sum_{j=1}^{D} p_j$. A critical question remains whether such a dimension reduced structure can identify geometric sub-regions in the tensor response which are related to the predictors. Additionally, we also intend to accurately estimate coefficients corresponding to these sub-regions of the tensor coefficient. The next section proposes a careful elicitation of the prior distribution on the tensor parameters to achieve our goal.

2.3. Multiway stick breaking shrinkage prior on tensor coefficients

Although the spike-slab prior for selective predictor inclusion (George and McCulloch, 1993; Clyde et al., 1996) possesses attractive theoretical properties, intractability of exploring an exponentially large space of predictor inclusion along with the belief that many regression coefficients may be small rather than exactly zero has led to considerable growth in the appeal for continuous shrinkage priors. An impressive variety of Bayesian shrinkage priors for ordinary high dimensional regression with a scalar/vector response on high dimensional vector predictors has been proposed in recent times, see for example Hans (2009); Carvalho et al. (2010); Armagan et al. (2013a) and references therein. Shrinkage priors are based on the principle of artfully shrinking predictor coefficients of unimportant predictors to zero, while maintaining proper estimation and uncertainty of the important predictor coefficients. Polson and Scott (2010) further show that most of the existing shrinkage priors can be expressed as the scale mixture of normal distributions with a global parameter common to all pre-

dictors and predictor specific local parameters. The global parameter imposes shrinkage globally while local parameters carefully balance shrinkage for large and small coefficients.

Literature on the vector shrinkage priors provides an excellent starting point for studying multiway shrinkage priors on tensor coefficient $\boldsymbol{\Gamma}_k$, though the latter presents a lot more challenges. Assuming that $\boldsymbol{\Gamma}_k$ admits a rank-R PARAFAC decomposition, proposing a prior on $\boldsymbol{\Gamma}_k$ is equivalent to specifying priors over tensor margins $\boldsymbol{\gamma}_{j,k}^{(r)}$. Given that every cell coefficient in $\boldsymbol{\Gamma}_k$ is a nonlinear function of the tensor margins, care should be taken while imposing prior shrinkage on them. To this end, Guhaniyogi et al. (2017) have characterized multiple restrictions on putting prior distributions on $\boldsymbol{\Gamma}_k$'s and have proposed the multiway dirichlet generalized double pareto (M-DGDP) shrinkage prior satisfying all the restrictions. However, in the context of BTRR, a straightforward application of M-DGDP prior on $\boldsymbol{\Gamma}_k$ leads to inaccurate estimation due to less desirable tail behavior of the distribution of $\Gamma_{\boldsymbol{v},k}$ parameters.

This article proposes a multiway stick breaking shrinkage prior on $\boldsymbol{\Gamma}_k$ to ensure desirable tail behavior for the tensor coefficient. More specifically, set $\tau_{r,k} = \phi_{r,k}\tau_k$, as the scaling specific to rank $r = 1, ..., R$. To achieve effective shrinkage across ranks we adopt a stick breaking construction for the rank-specific scale 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 $\xi_{r,k} \overset{iid}{\sim} Beta(1, \alpha_k)$. The global scale parameter is modeled as $\tau_k \sim IG(a_\tau, b_\tau)$. Additionally, the local scale parameters $\boldsymbol{W}_{jr,k} = \text{diag}(w_{jr,k,1}, ..., w_{jr,k,p_j})$ are employed to achieve margin level shrinkage in the following way

$$\boldsymbol{\gamma}_{j,k}^{(r)} \sim N(\mathbf{0}, \tau_{r,k}\boldsymbol{W}_{jr,k}),\ \ w_{jr,k,i} \sim Exp(\lambda_{jr,k}^2/2),\ \ \lambda_{jr,k} \sim Ga(a_\lambda, b_\lambda),\ \ i = 1, ..., p_j.$$

The construction tacitly exploits the finite stick breaking construction for the local parameters $\phi_{r,k}$'s. As $\alpha_k \to 0$, most of $\phi_{r,k}$'s will be close to being sparse. Therefore, careful learning of α_k leads to a sparse and parsimonious representation of the tensor. α_k is assigned a discrete uniform prior on a grid and learnt using a greedy Gibbs algorithm. Additionally, flexibility in estimating tensor margins $\{\boldsymbol{\gamma}_{j,k}^{(r)} : 1 \le j \le D,\ 1 \le r \le R\}$ is accommodated by modeling heterogeneity within margins via element-specific scaling $\boldsymbol{W}_{jr,k}$. A common rate parameter $\lambda_{jr,k}$ encourages sharing of information between the margin elements. In fact, it is easy to see that $\gamma_{j,k,i}^{(r)}|\phi_{r,k}, \tau_k$ follows the well known generalized double pareto (GDP) (Armagan et al., 2013a) shrinkage prior distribution. Exploiting more efficient computational techniques, TRR with the multiway stick breaking shrinkage prior accurately estimates the posterior distribution of $\boldsymbol{\Gamma}_k$ for a relatively large number of cells compared to the ordinary spike and slab prior on cell coefficients.

3. Posterior consistency in tensor response regression

3.1. Notations

In what follows, we add a subscript (T) to the dimensions of tensor margins $p_{1,(T)}, ..., p_{D,(T)}$ and the number of predictors $m_{(T)}$ to indicate that the size of both the response tensor $\boldsymbol{Y}_t$ and covariates $\boldsymbol{x}_t$ can increase with the sample size T. This asymptotic paradigm is also meant to capture the fact that the number of cells $\prod_{j=1}^{D} p_{j,(T)}$ is typically larger than the sample size T for each tensor coefficients $\boldsymbol{\Gamma}_{1,(T)}, .., \boldsymbol{\Gamma}_{m_{(T)},(T)}$. Define $\boldsymbol{\Gamma}$ as a $\Re^m \otimes_{j=1}^{D} \Re^{p_j}$ tensor with the $(v_1, .., v_D, k)$th cell being given by the $(v_1, ..., v_D)$th cell of $\boldsymbol{\Gamma}_{k,(T)}$. Naturally, the tensor coefficient $\boldsymbol{\Gamma}$ and tensor margins $\boldsymbol{\gamma}_{j,k}^{(r)}$s are also functions of the sample size T and we denote them by $\boldsymbol{\Gamma}_{(T)}$ and $\boldsymbol{\gamma}_{j,k,(T)}^{(r)}$s respectively. We use superscript (0) to indicate true parameters, e.g. the true tensor regression parameter and the true error variance are denoted by $\boldsymbol{\Gamma}_{(T)}^{(0)}$ and $\sigma^{(0)2}$ respectively. For simplicity, we assume that $\sigma^2 = \sigma^{(0)2}$ is known and fixed at 1. We also assume that κ is fixed an known, so that $var(\boldsymbol{E}_{\boldsymbol{v}}) = \boldsymbol{R}$ is fixed, where $\boldsymbol{E}_{\boldsymbol{v}} = (E_{1,\boldsymbol{v}}, ..., E_{T,\boldsymbol{v}})'$. While κ and σ^2 are unknown in practice and are assigned prior distributions, our setup assumes them to be fixed and known. This is a common assumption in the asymptotic study (Van der Vaart and Van Zanten, 2011). Furthermore, it is known that the theoretical results obtained by assuming these parameters as known constants are equivalent to those obtained by assigning priors with bounded supports on these parameters (Van der Vaart and Van Zanten, 2009). For vectors, we let $||\cdot||_2$ denote the L_2-norm, $||\cdot||_1$ denote the L_1-norm and $||\cdot||_\infty$ denote the L_∞ norm. With a slight abuse of notations, for a D-dimensional tensor object $\boldsymbol{A}$, the L_1, L_2 and L_∞ norms are defined as $||\boldsymbol{A}||_1 = \sum_{v_1,...,v_D} |A_{v_1,...,v_D}|$, $||\boldsymbol{A}||_2 = \sqrt{\sum_{v_1,...,v_D} A^2_{v_1,...,v_D}}$ and $||\boldsymbol{A}||_\infty = \max_{v_1,..,v_D} |A_{v_1,...,v_D}|$. $||\cdot||_0$ denotes the L_0-norm, i.e. the number of non-zero entries, for both vectors and tensors. Further, assume $\mathcal{F}_1 = \{\boldsymbol{h}_1 = (v_1, ..., v_D) : 1 \leq v_1 \leq p_{1,(T)}, ..., 1 \leq v_D \leq p_{D,(T)}\}$, $\mathcal{F}_2 = \{h_2 = v_{D+1} : 1 \leq v_{D+1} \leq m_{(T)}\}$. Denote $\zeta^{(0)} = \{(\boldsymbol{h}_1, h_2) : \Gamma^{(0)}_{\boldsymbol{h}_1,h_2,(T)} \neq 0, \boldsymbol{h}_1 \in \mathcal{F}_1, h_2 \in \mathcal{F}_2\}$ as a set of indices corresponding to the nonzero cells of the true tensor coefficient, and also denote $\zeta_1^{(0)} = \{\boldsymbol{h}_1 \in \mathcal{F}_1 : \Gamma^{(0)}_{\boldsymbol{h}_1,h_2,(T)} \neq 0$, for some $h_2 \in \mathcal{F}_2\}$. Similarly, for any set $\zeta \subseteq \mathcal{F}_1 \times \mathcal{F}_2$, define $\zeta_1 = \{\boldsymbol{h}_1 \in \mathcal{F}_1 : (\boldsymbol{h}_1, h_2) \in \zeta\}$ and $\zeta_{2,\boldsymbol{h}_1} = \{h_2 \in \mathcal{F}_2 : (\boldsymbol{h}_1, h_2) \in \zeta\}$. $|\zeta|$ denotes the cardinality of the set ζ. We let $s_{(T)}$ (dependent on T) denote the number of nonzero entries in the true tensor coefficient, i.e., $s_{(T)} = ||\boldsymbol{\Gamma}^{(0)}_{(T)}||_0$. Let $e_{max}(\cdot)$ and $e_{min}(\cdot)$ denote the largest and smallest eigenvalues of a square matrix, respectively.

Since the shrinkage prior on $\boldsymbol{\Gamma}_{(T)}$ assigns zero probability at the point zero, the exact number of nonzero elements of $\boldsymbol{\Gamma}_{(T)}$ is always $m_{(T)} \prod_{j=1}^{D} p_{j,(T)}$. A meaningful comparison with the value $s_{(T)}$ is made by considering $\tilde{s}_{(T)}$, the number of elements of $\boldsymbol{\Gamma}_{(T)}$ exceeding in absolute value a threshold a_T, which will be specified later. In other words, only elements with absolute value larger than a_T will be treated as significant and counted towards non-zero entries.

Define $\mathcal{B}_T = \{\text{At least } \tilde{s}_{(T)} \text{ absolute values of } \boldsymbol{\Gamma}_{(T)} \text{ are greater than } a_T\}$, $\mathcal{C}_T = \left\{\boldsymbol{\Gamma}_{(T)} : ||\boldsymbol{\Gamma}_{(T)} - \boldsymbol{\Gamma}^{(0)}_{(T)}||_2 > \epsilon\right\}$ and $\mathcal{A}_T = \mathcal{B}_T \cup \mathcal{C}_T$. Further suppose $\pi_T(\cdot)$ and $\Pi_T(\cdot)$ are the prior and posterior densities of $\boldsymbol{\Gamma}_{(T)}$ with T observations, so that

$$\Pi_T(\mathcal{A}_T) = \frac{\int_{\mathcal{A}_T} f(\boldsymbol{Y}_1, ..., \boldsymbol{Y}_T|\boldsymbol{\Gamma}_T)\pi_T(\boldsymbol{\Gamma}_T)}{\int f(\boldsymbol{Y}_1, ..., \boldsymbol{Y}_T|\boldsymbol{\Gamma}_T)\pi_T(\boldsymbol{\Gamma}_T)},$$

where $f(\boldsymbol{Y}_1, ..., \boldsymbol{Y}_T|\boldsymbol{\Gamma}_{(T)})$ is the joint density of $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_T$ under model (2.2). This article intends to show

$$\Pi_T(\mathcal{A}_T) \to 0, \text{ a.s., when } T \to \infty. \tag{3.1}$$

3.2. Main results

The following theorem shows that (3.1) holds under mild sufficient conditions on $s_{(T)}$, $\tilde{s}_{(T)}$ and $p_{j,(T)}$s. The proof of the theorem is given in the appendix.

Theorem 3.1. *Denote* $p_{(T)} = m_{(T)} \prod_{j=1}^{D} p_{j,(T)}$. *Let*

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

(b) $||\boldsymbol{\Gamma}^{(0)}_{k,(T)}||_0 = s_{(T)}$, *with* $s_{(T)} \log(p_{(T)}) = o(T)$;

(c) $\tilde{s}_{(T)} \log(p_{(T)}) = O(T)$;

(d) $m_{(T)} \sum_{j=1}^{D} p_{j,(T)} \log(p_{j,(T)}) = o(T)$;

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

Under conditions (a)-(e), (3.1) holds with $a_T = \frac{\epsilon}{2p_{(T)}}$.

Remark: Condition (a) in Theorem 3.1 assumes a low-rank decomposition for the true tensor coefficient. This is a mild condition as most applications allow low-rank structure for the true tensor coefficients. Regarding condition (b), note that $s_{(T)}$ is the sparsity of the true tensor and $p_{(T)}$ is the total number of cells in the tensor. When the tensor is just a scalar ($D = 0$), i.e., the tensor regression reduces to an ordinary high dimensional regression with $m_{(T)}$ predictors, the condition reduces to $s_{(T)} \log(m_{(T)}) = o(T)$, which is a typical assumption in ordinary high dimensional regression, see Song and Liang (2017). Condition (c) also assumes the same condition for the "near sparsity" in the estimated $\boldsymbol{\Gamma}_{(T)}$ in the sense of $\mathcal{B}_T$. Condition (d) in Theorem 3.1 requires that $m_{(T)} \sum_{j=1}^{D} p_{j,(T)}$ grows sub-linearly with sample size T. However, the number of cells $m_{(T)} \prod_{j=1}^{D} p_{j,(T)}$ in the tensor $\boldsymbol{\Gamma}_{(T)}$ can grow at a rate much faster than the sample size T; hence, the modeling framework allows large tensor responses even for moderate sample sizes. Condition (e) is equivalent to a lower bounded compatibility number condition assumed in the theoretical study of ordinary high

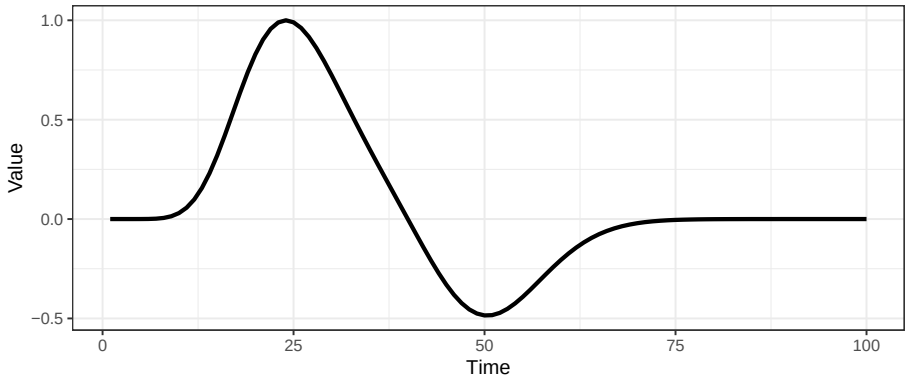

FIG 1. *Values taken by the simulated covariate through time when the number of total time steps was set to* $T = 100$.

dimensional regression, (see Song and Liang (2017); Castillo et al. (2015b)). Finally, condition (e) also ensures $e_{max}(\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X})$ grows sub-linearly with T.

4. Posterior Computation

Under a Bayesian framework, parameter estimation can be achieved via Markov chain Monte Carlo (MCMC) algorithms, in which posterior distributions for the unknown quantities are approximated with empirical distributions of samples from a Markov chain. The full conditional distributions for developing Gibbs within Metropolis algorithms are provided in the supplementary material.

5. Simulated Data Results

This section showcases parametric inference from Bayesian tensor response regression (BTRR) with various simulation studies. Since the major motivation of model development is drawn from the fMRI based brain activation study, the simulation study is performed on simulated datasets reminiscent of real world fMRI data. Scalar predictors are simulated with the block experimental design. A single stimulus block was convolved with the canonical double-gamma haemodynamic response function. An example of the values taken by the covariate can be seen in figure 1.

The block design consisted of a single discrete epoch of activity and rest, with the "activity" representing a period of stimulus presentation, and the "rest" referring to a state of rest or baseline. The stimulus was assumed to take place at time $t = 0$ for a duration of one time step, with a stimulus value of 1.

The response tensor is simulated from (2.2) with $D = 2$ and $p_{D+1} = 1$ and $\sigma^{(0)2} = 1$. Thus, the true coefficient tensor $\boldsymbol{\Gamma}^{(0)}$ is assumed to be sparse and two-dimensional (i.e. $D = 2$). The `specifyregion` function within the `neuRosim`

package in R (Welvaert et al., 2011) is employed to simulate the nonzero regions of the true coefficient tensor $\mathbf{\Gamma}^{(0)}$. Lengths of each dimension (p_1, p_2) of the tensor coefficient are drawn from a Poisson distribution with shared parameter μ and the nonzero elements assuming value η. The scenarios were created by constructing a grid over different values for $T \in \{20, 50, 100, 200\}$, $\mu \in \{5, 10, 20, 30\}$, and $\eta = \{0.1, 0.25, 0.5, 0.75, 1, 1.5\}$.

The model is fitted in each simulation scenario along with the naive maximum likelihood estimator for the coefficient tensor $\mathbf{\Gamma}$ to highlight the advantages of joint Bayesian modeling with tensor coefficients. In all cases, the log-likelihood was examined in order to verify that the Markov chain converged. The model witnesses rapid convergence, so that in each model fitting 1,100 draws are taken from the joint posterior distribution, out of which the first 100 draws are discarded as burn-in. Average effective sample sizes shown in Figure 2 for the $1,000$ post burn samples calculated using the `coda` package in `R` confirm sufficiently uncorrelated post burn-in samples.

Point estimation of $\mathbf{\Gamma}$. A comparison of the posterior mean of the elements of $\mathbf{\Gamma}$ for different values of R and η when $\mu = 30$ and $T = 20$ can be seen in Figure 3. We especially show figures in this case since this case represents higher tensor dimensions and smaller sample size. The posterior mean estimates show the effects of the regularization in the prior, which pulls the posterior mean values corresponding to unimportant cells closer to zero. The true and the estimated activation maps demonstrate excellent performance of BTRR in capturing the true activation pattern under moderate contrast to noise ratio η. When contrast to noise ratio drops below 1, identifying signal from noise remains a challenging task which causes less accurate identification of activated regions. It should be mentioned that this simulation scenario is well outside the umbrella of theoretical guarantee observed in Theorem 3.1 since $s_T \log(p_T)$ is much larger than T, and yet the model is able to identify the truly activated regions.

Figure 4 shows the root mean squared error of the estimates of $\mathbf{\Gamma}$ under different scenarios. In each scenario, BTRR model with ranks (R) 1, 2, and 3 are tested, and further testing suggests that additional ranks are not required. In a real data application, the final rank used to fit a model can be selected using the deviance information criterion (Gelman et al., 2014). The model at ranks 1, 2, and 3 is compared to a naive maximum likelihood estimate, which is found by regressing each $\mathbf{Y}_{t,v}$ on $\mathbf{x}_t$ separately for each cell in the experiment.

Based on the results, the model performs well, both for low and high contrast to noise ratio. Although the root mean squared error (RMSE) metric seems to be lower for $\eta = 0.5$ compared to $\eta = 1.5$, this does not contradict Figure 3. It is worth noting that the shrinkage mechanism pulls every coefficient towards zero, with significant cell coefficients observing less shrinkage than unimportant coefficients. Since for $\eta = 0.5$, even important coefficients are close to zero, all estimated coefficients are close to the truth. On the contrary, for $\eta = 1.5$, shrinkage of important coefficients leads to an increase in RMSE. When RMSE figures are normalized with the true signal strength, the model shows much improved performance for $\eta = 1.5$ than $\eta = 0.5$. Note that the naive MLE does

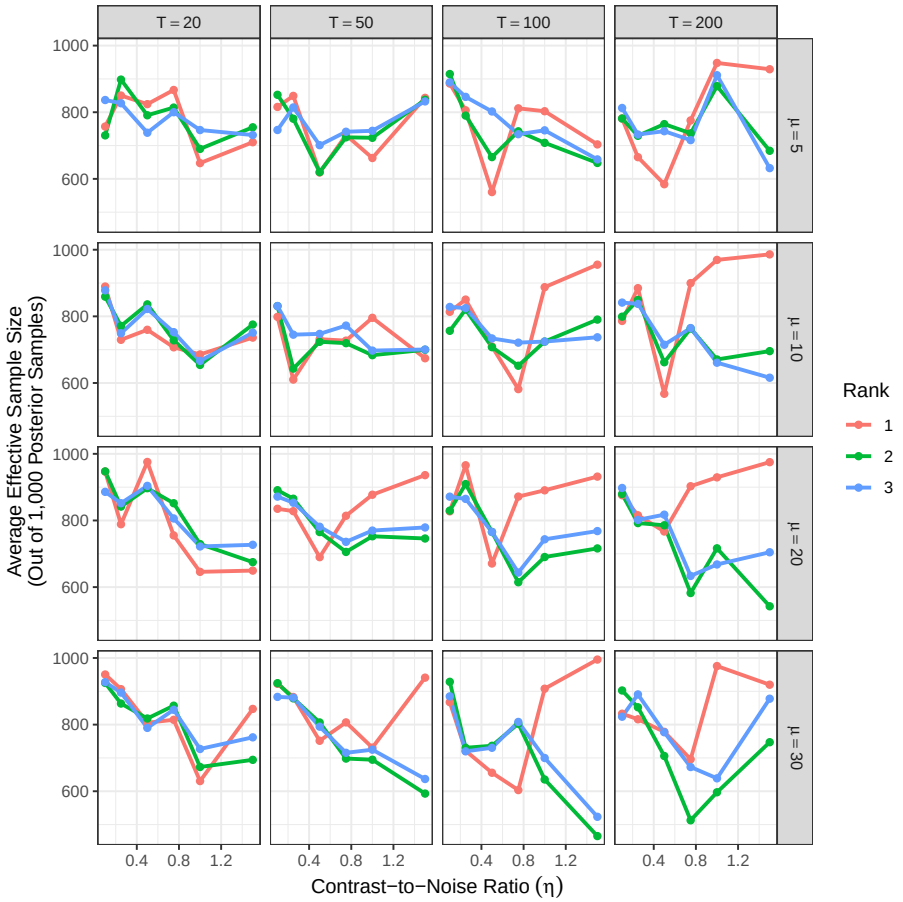

FIG 2. *The average effective sample size for elements of* $\mathbf{\Gamma}$ *under each of 288 scenarios.*

not assume sparsity and uses more parameters in almost every case, which is a disadvantage in low-signal sparse regressions. It is used as a comparison to show that the proposed model provides a reasonable point estimate.

Parametric uncertainty of $\mathbf{\Gamma}$. To assess uncertainty quantification of $\mathbf{\Gamma}$ from BTRR, we focus on coverage and length of 95% credible intervals (CI) of cells of $\mathbf{\Gamma}$, shown in Figures 5 and 6 respectively. Given that in almost all the scenarios, the coverage of the 95% credible intervals is close to nominal, attention turns to the length of the 95% credible intervals. Two visible patterns emerge from the figures. First, the 95% credible intervals shrink as T increases, since the posterior variance lowers with increased observed data. Secondly, the credible intervals are wider for higher contrast to noise ratio, which can be attributed to the fact that estimating a few high signals with lots of zero coefficients involves more uncertainties.

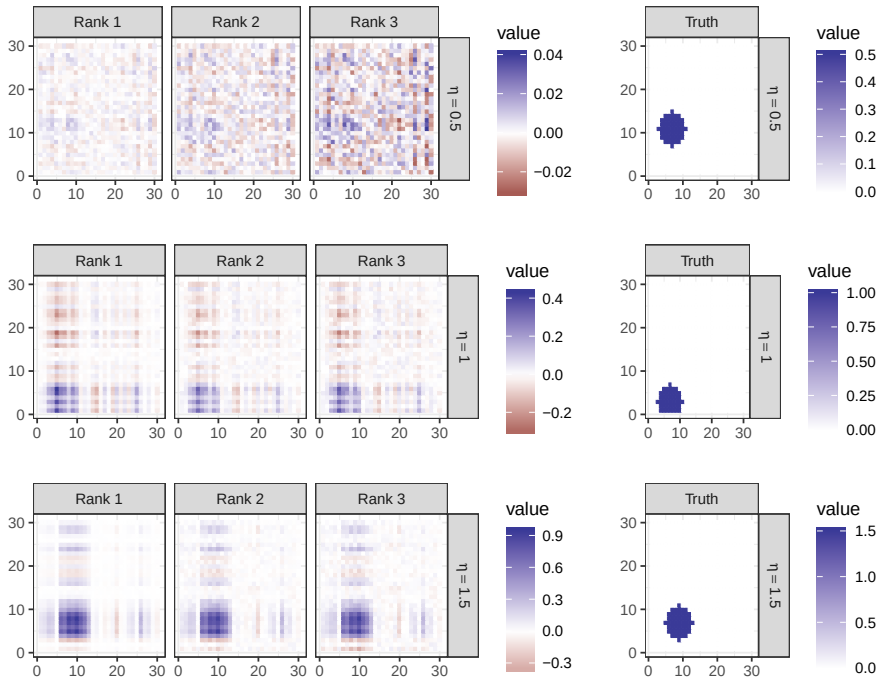

FIG 3. *Estimated and true values for* $\mathbf{\Gamma}^0$ *when* $\mu = 30$ *and* $T = 20$ *under different values for* R *and* η.

All scenarios conclusively establish the strength of BTRR as a principled Bayesian approach that accurately detects brain activation with proper characterization of uncertainties. It is particularly appealing to observe BTRR outperforming MLE estimates in smaller contrast to noise ratios reminiscent of real fMRI data. We have also replaced naive MLE by a few penalized optimizers (not shown here) and found BTRR continuing to be the clear winner in sparse and low signal scenarios. Application of the model to real data is explored in the following section.

6. Application to Balloon Analog Risk Taking Data

Neuroscientists at the University of California Los Angeles conducted an experiment intended to infer about the regions of the brain that are involved in the process of evaluating risk (Schonberg et al., 2012). Sixteen young adults (average age of 23.56 years) were subjects in an experiment with the following design. Each subject entered an fMRI machine with a computer display and a controller with two buttons. On the screen, the image of a balloon would be shown, along with a payout amount, starting with a value of \$0.25. The buttons

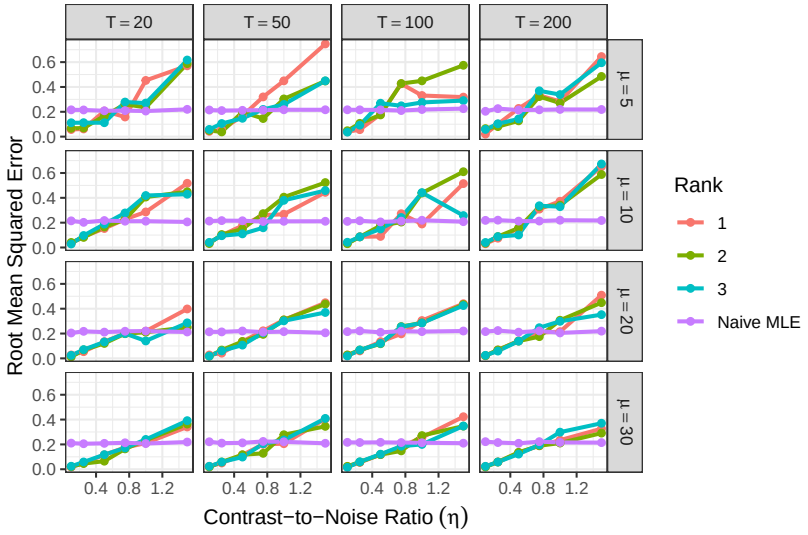

FIG 4. *Root mean squared error from analyses on simulated data.*

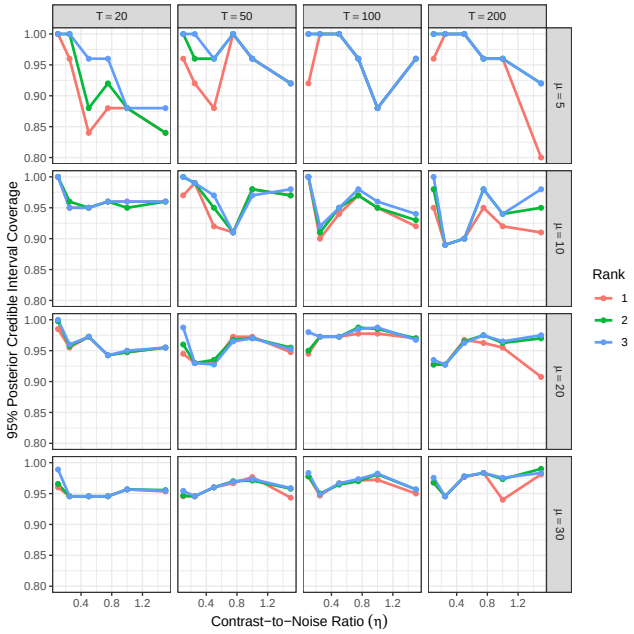

FIG 5. *The average coverage of the 95% posterior credible intervals for the posterior draws for the elements of* $\mathbf{\Gamma}$ *under varying conditions.*

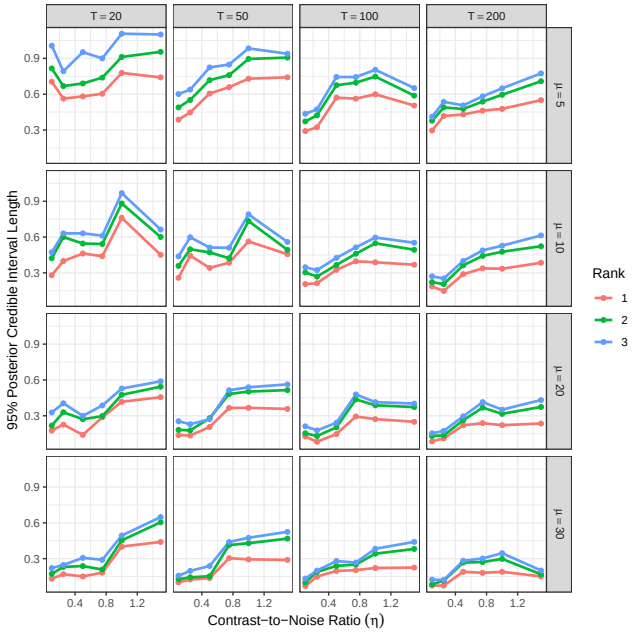

FIG 6. *The average length of the 95% posterior credible intervals for the posterior draws for the elements of* $\mathbf{\Gamma}$ *under varying conditions.*

on the controller allowed the subject to either inflate the balloon or take the payout. If the subject inflated the balloon, and the balloon did not explode, the payout amount increased by \$0.25. If the subject inflated the balloon, and the balloon exploded, no payout was received, the payout value was reset to \$0.25, and a new balloon was displayed. Balloons were assigned a number of pumps at which the balloon would explode from a discrete uniform distribution with a lower bound of 1, and an upper bound of 8, 12, or 16, depending on whether the balloon was red, green, or blue, respectively. A grey "control" balloon, offering no payout and an upper bound of 12 pumps before exploding, was also part of the trial to record a riskless scenario. Each subject participated in three runs. Each run consisted of either 10 minutes, or 48 balloons exploding, whichever came first. In order to be able to analyze this data on a laptop with 8GB memory, only the first run was used for each subject. Before analysis, the data was preprocessed to correct for motion and physiological differences between subjects using FSL (Smith et al., 2004), sliced into a two-dimensional cross-section, and then separated into 9 different regions of interest based on an atlas from the Montreal Neurological Institute. Table 1 records these Regions of Interest (ROI) with the number of voxels in them. Our preliminary investigation confirms that a single tensor response regression model is inadequate for all the ROIs, so that ROI specific tensor response regression model is fitted to the data. This separation additionally serves to make the MCMC easily parallelizable, group physical

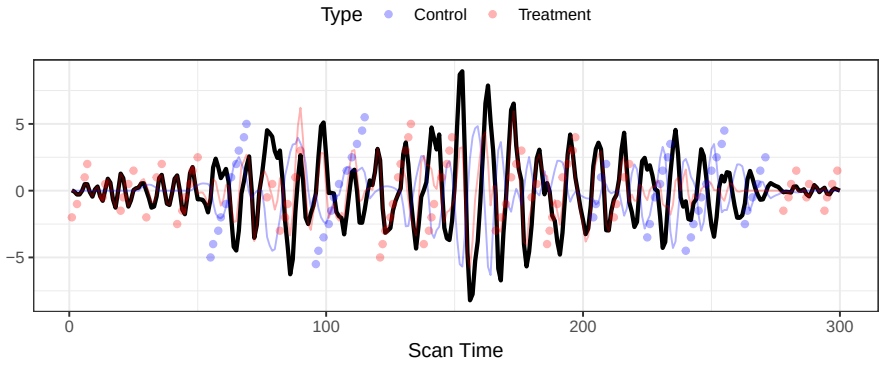

FIG 7. *The raw values of the demeaned number of pumps (points), their convolution with the double-gamma haemodynamic response function (light lines), and the final covariate resulting from their difference (heavy black line) for a single run from a single subject.*

components of the brain together, and allow for the selection of different values of R for fitting the model on different regions of interest in the fMRI scan.
To measure the level of risk to a subject at a given time, we measure the centered number of pumps that an individual gives a "treatment" balloon before they "cash-out" or the balloon explodes. It is assumed that the higher the number of pumps becomes, the more the risk to the individual. This value is then convolved with the double-gamma haemodynamic response function, which takes into account the physiological lag between stimulus and response, and smooths the stepwise function for the centered number of pumps. An illustration of this calculation can be seen in Figure 7.

Finally, the centered, convolved number of pumps on the control balloon is subtracted from the treatment series to provide a basis for comparison. The multiway stick-breaking shrinkage model is applied to this data under ranks 1, 2, and 3 in each region; and each Markov chain is run for 1,100 iterations. The first 100 iterations are discarded as burn-in after checking the stationarity of the log-likelihood. The mean effective sample sizes of the 1,000 post burn-in samples are 893.6343, 860.1507, and 791.4332 for ranks 1, 2, and 3, respectively. The point estimates for activation coefficient for different ranks can be seen in Figure 8. As the brain images are split into nine regions of interest and BTRR fitted separately for these regions, the final estimate for activation coefficient is obtained by estimating the posterior mean of $\mathbf{\Gamma}$ from each ROI, with rank R in each region estimated using the Deviance Information Criterion (DIC). The final estimate obtained in this fashion is presented in Figure 9.

The same naive ordinary least squares (OLS) estimator described in section 5 is also added for comparison. The scale of the results for the OLS estimator is about four times that of the multiway stick-breaking shrinkage model. This is likely due in part to the shrinkage prior favoring smaller values, and the multiplicative nature of the PARAFAC decomposition of the coefficient tensor.

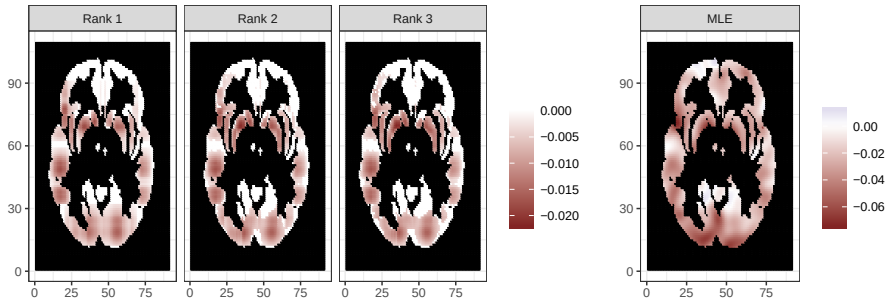

FIG 8. *A comparison of estimates of* $\mathbf{\Gamma}$ *under the naive maximum likelihood estimate, and ranks 1, 2, and 3.*

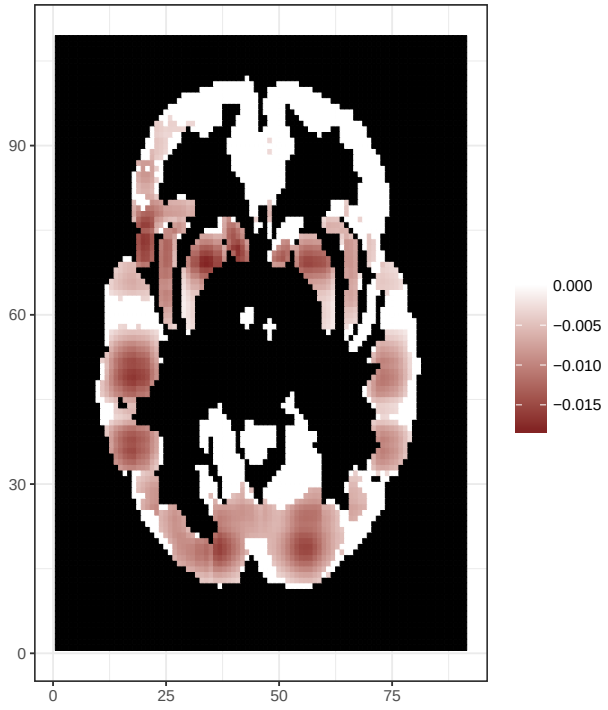

FIG 9. *The final estimate of the effect of increased perceived risk on the relative levels of oxygen in different regions of the brain after selecting R for each region using the DIC.*

These features likely improve the OLS estimator, due to the sparse nature of signal (with the magnitude of nonzero voxel coefficients likely to be small) in fMRI applications (Daubechies et al., 2009).

Region	p_1	p_2	Chosen Rank
Cerebellum	8	8	Rank 1
Putamen	34	19	Rank 1
Parietal Lobe	22	7	Rank 2
Caudate	19	9	Rank 3
Frontal Lobe	57	35	Rank 3
Insula	45	23	Rank 3
Occipital Lobe	56	31	Rank 3
Temporal Lobe	72	40	Rank 3
Thalamus	8	6	Rank 3

TABLE 1

The different values for R selected by the deviance information criterion (DIC), along with the dimensions associated with the response tensors in each region.

7. Conclusion

This article proposes a Bayesian framework to regress a tensor valued response on scalar covariates. Adopting the rank-R PARAFAC decomposition for the tensor coefficient, the proposed model is able to reduce the number of free parameters. We employ a novel multiway stick breaking shrinkage prior distribution on the tensor coefficient to be able to identify significantly nonzero cell coefficients. New results on posterior consistency have been developed to show convergence in L_2 sense of the tensor coefficient to the true tensor as data size increases.

As an illustrative example, the present article focuses on analysis of fMRI data to detect voxels of the brain which exhibit neuronal activity in response to stimuli, while simultaneously inferring on the association of spatially remote groups of voxels with similar characteristics. Analysis of simulated fMRI time series and real fMRI data demonstrates excellent performance of BTRR in identifying the regions of activation with required uncertainties. Additionally, BTRR is able to achieve remarkable parsimony, even after being a Bayesian model. This facilitates its usage in presence of images with a fine resolution.

The core idea of the proposal is to recognize the importance of retaining the tensor structure of the image response during the entire statistical analysis for studies including brain activation. An immediate extension to the proposed model would be meant to investigate both voxel level activation and ROI level connectivity from multi-subject fMRI data.

Supplementary Material

Supplementary Material: Bayesian Tensor Response Regression With an Application to Brain Activation Studies:
(). Supplementary material consists of full posterior conditionals for all the parameters to implement the MCMC algorithm.

8. Acknowledgement

The research of Rajarshi Guhaniyogi is partially supported by grants from the Office of Naval Research (ONR-BAA N000141812741) and the National Science Foundation (DMS-1854662).

Appendix

The proof of Theorem 3.1 relies in part on the existence of exponentially consistent sequence of tests.

Definition An exponentially consistent sequence of test functions Φ_T for testing $H_0 : \mathbf{\Gamma}_{(T)} = \mathbf{\Gamma}^{(0)}_{(T)}$ vs. $H_1 : \mathbf{\Gamma}_{(T)} \in \mathcal{A}_T$ satisfies

$$E_{\mathbf{\Gamma}^{(0)}_{(T)}}(\Phi_T) \leq c_1 \exp(-b_1 T), \qquad \sup_{\mathbf{\Gamma}_{(T)} \in \mathcal{A}_T} E_{\mathbf{\Gamma}_{(T)}}(1 - \Phi_T) \leq c_2 \exp(-b_2 T)$$

for some $c_1, c_2, b_1, b_2 > 0$.

Theorem 8.1. *There exist an exponentially consistent sequence of tests Φ_T for testing $H_0 : \mathbf{\Gamma}_{(T)} = \mathbf{\Gamma}^{(0)}_{(T)}$ vs. $H_1 : \mathbf{\Gamma}_{(T)} \in \mathcal{A}_T$.*

Proof. Let $\zeta \in \mathcal{F}_1 \times \mathcal{F}_2$. For any $\boldsymbol{h}_1 \in \zeta_1$, let $\hat{\mathbf{\Gamma}}_{(T),\boldsymbol{h}_1,\zeta_{2,\boldsymbol{h}_1}} = (\boldsymbol{X}'_{\zeta_{2,\boldsymbol{h}_1}} \boldsymbol{R}^{-1} \boldsymbol{X}_{\zeta_{2,\boldsymbol{h}_1}})^{-1} \boldsymbol{X}'_{\zeta_{2,\boldsymbol{h}_1}} \boldsymbol{R}^{-1} \boldsymbol{y}_{\boldsymbol{h}_1}$, where $\boldsymbol{y}_{\boldsymbol{h}_1} = (Y_{1,\boldsymbol{h}_1}, ..., Y_{T,\boldsymbol{h}_1})'$ and $\boldsymbol{X}_{\zeta_{2,\boldsymbol{h}_1}}$ is a $T \times |\zeta_{2,\boldsymbol{h}_1}|$ dimensional matrix whose tth row is given by $(\boldsymbol{x}_{j,t} : j \in \zeta_{2,\boldsymbol{h}_1})$. Define a test function $\Phi_T = \max_{|\zeta| \leq \tilde{s}_T + s_T, \zeta \supseteq \zeta^{(0)}} 1\left\{||\hat{\mathbf{\Gamma}}_{(T),\zeta} - \mathbf{\Gamma}^{(0)}_{(T),\zeta}||_2 > \epsilon/4\right\}$. In what follows, we will show that Φ_T is an exponentially consistent sequence of tests.

$$\begin{aligned}
E_{\mathbf{\Gamma}^{(0)}_{(T)}}(\Phi_T) &\leq \sum_{|\zeta| \leq \tilde{s}_{(T)} + s_{(T)}, \zeta \supseteq \zeta^{(0)}} P\left(||\hat{\mathbf{\Gamma}}_{(T),\zeta} - \mathbf{\Gamma}^{(0)}_{(T),\zeta}||_2 > \epsilon/4\right) \\
&\leq \sum_{|\zeta| \leq \tilde{s}_{(T)} + s_{(T)}, \zeta \supseteq \zeta^{(0)}} P\left(\sum_{\boldsymbol{h} \in \zeta_1} (\hat{\mathbf{\Gamma}}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}} - \mathbf{\Gamma}^{(0)}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}})'(\hat{\mathbf{\Gamma}}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}} - \mathbf{\Gamma}^{(0)}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}}) > \epsilon^2/16\right) \\
&\leq \sum_{|\zeta| \leq \tilde{s}_{(T)} + s_{(T)}, \zeta \supseteq \zeta^{(0)}} P\Big(\sum_{\boldsymbol{h} \in \zeta_1} (\hat{\mathbf{\Gamma}}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}} - \mathbf{\Gamma}^{(0)}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}})'(\boldsymbol{X}'_{\zeta_{2,\boldsymbol{h}}} \boldsymbol{R}^{-1} \boldsymbol{X}_{\zeta_{2,\boldsymbol{h}}})(\hat{\mathbf{\Gamma}}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}} - \mathbf{\Gamma}^{(0)}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}}) > \\
&\qquad T\lambda_0^2 \epsilon^2/16\Big) \\
&= \sum_{|\zeta| \leq \tilde{s}_{(T)} + s_{(T)}, \zeta \supseteq \zeta^{(0)}} P\left(\sum_{\boldsymbol{h} \in \zeta_1} \chi^2_{|\zeta_{2,\boldsymbol{h}}|} > T\lambda_0^2 \epsilon^2/16\right) \\
&= \sum_{|\zeta| \leq \tilde{s}_{(T)} + s_{(T)}, \zeta \supseteq \zeta^{(0)}} P\left(\chi^2_{|\zeta|} > T\lambda_0^2 \epsilon^2/16\right) \leq \binom{p_{(T)}}{\tilde{s}_{(T)} + s_{(T)}} \exp(-T\lambda_0^2 \epsilon^2/16),
\end{aligned}$$

where the last inequality follows from Lemma A.1 and A.2 in Song and Liang (2017). Note that $\binom{p_{(T)}}{\tilde{s}_{(T)}+s_{(T)}} \leq p_{(T)}^{\tilde{s}_{(T)}+s_{(T)}} \leq \exp((\tilde{s}_{(T)}+s_{(T)})\log(p_{(T)})) \leq \exp(T\lambda_0^2\epsilon^2/32)$, by assumptions (b) and (c). Thus $E_{\mathbf{\Gamma}_{(T)}^{(0)}}(\Phi_T) \leq \exp(-T\lambda_0^2\epsilon^2/32)$.

Let $\tilde{\zeta} = \zeta^{(0)} \cup \{(\boldsymbol{h}_1, h_2) : |\Gamma_{(T),\boldsymbol{h}_1,h_2}| \geq a_T\}$

$$\sup_{\mathbf{\Gamma}_{(T)}\in\mathcal{A}_T} E_{\mathbf{\Gamma}_{(T)}}(1-\Phi_T) \leq \sup_{\mathbf{\Gamma}_{(T)}\in\mathcal{A}_T} E_{\mathbf{\Gamma}_{(T)}}(1-1\left\{||\hat{\mathbf{\Gamma}}_{(T),\tilde{\zeta}} - \mathbf{\Gamma}^{(0)}_{(T),\tilde{\zeta}}||_2 > \epsilon/4\right\})$$
$$= \sup_{\mathbf{\Gamma}_{(T)}\in\mathcal{A}_T} P_{\mathbf{\Gamma}_{(T)}}\left(||\hat{\mathbf{\Gamma}}_{(T),\tilde{\zeta}} - \mathbf{\Gamma}^{(0)}_{(T),\tilde{\zeta}}||_2 \leq \epsilon/4\right).$$

Under $\mathcal{A}_T$, $||\mathbf{\Gamma}_{(T),\tilde{\zeta}} - \mathbf{\Gamma}^{(0)}_{(T),\tilde{\zeta}}||_2 \geq ||\mathbf{\Gamma}_{(T)} - \mathbf{\Gamma}^{(0)}_{(T)}||_2 - ||\mathbf{\Gamma}_{(T),\tilde{\zeta}^c} - \mathbf{\Gamma}^{(0)}_{(T),\tilde{\zeta}^c}||_2 \geq \epsilon - a_T p_T \geq \epsilon/2$. Where the last inequality follows due to the fact $\mathbf{\Gamma}^{(0)}_{(T),\tilde{\zeta}^c} = \mathbf{0}$ and $|\Gamma_{(T),\boldsymbol{h}_1,h_2}| \leq a_T$ for $(\boldsymbol{h}_1, h_2) \in \tilde{\zeta}^c$.

Using the above fact

$$\sup_{\mathbf{\Gamma}_{(T)}\in\mathcal{A}_T} E_{\mathbf{\Gamma}_{(T)}}(1-\Phi_T) \leq \sup_{\mathbf{\Gamma}_{(T)}\in\mathcal{A}_T} P_{\mathbf{\Gamma}_{(T)}}\left(||\hat{\mathbf{\Gamma}}_{(T),\tilde{\zeta}} - \mathbf{\Gamma}^{(0)}_{(T),\tilde{\zeta}}||_2 \leq \epsilon/4\right)$$
$$\leq \sup_{\mathbf{\Gamma}_{(T)}\in\mathcal{A}_T} P_{\mathbf{\Gamma}_{(T)}}\left(||\hat{\mathbf{\Gamma}}_{(T),\tilde{\zeta}} - \mathbf{\Gamma}_{(T),\tilde{\zeta}}||_2 \geq -||\hat{\mathbf{\Gamma}}_{(T),\tilde{\zeta}} - \mathbf{\Gamma}^{(0)}_{(T),\tilde{\zeta}}||_2 + ||\mathbf{\Gamma}_{(T),\tilde{\zeta}} - \mathbf{\Gamma}^{(0)}_{(T),\tilde{\zeta}}||_2\right)$$
$$\leq \sup_{\mathbf{\Gamma}_{(T)}\in\mathcal{A}_T} P_{\mathbf{\Gamma}_{(T)}}\left(||\hat{\mathbf{\Gamma}}_{(T),\tilde{\zeta}} - \mathbf{\Gamma}_{(T),\tilde{\zeta}}||_2 \geq \epsilon/4\right)$$
$$\leq \sup_{\mathbf{\Gamma}_{(T)}\in\mathcal{A}_T} P\left(\sum_{\boldsymbol{h}\in\zeta_1} (\hat{\mathbf{\Gamma}}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}} - \mathbf{\Gamma}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}})'(\boldsymbol{X}'_{\zeta_{2,\boldsymbol{h}}}\boldsymbol{R}^{-1}\boldsymbol{X}_{\zeta_{2,\boldsymbol{h}}})(\hat{\mathbf{\Gamma}}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}} - \mathbf{\Gamma}_{(T),\boldsymbol{h},\zeta_{2,\boldsymbol{h}}}) > T\lambda_0^2\epsilon^2/16\right)$$
$$\leq \sup_{\mathbf{\Gamma}_{(T)}\in\mathcal{A}_T} P\left(\sum_{\boldsymbol{h}\in\zeta_1} \chi^2_{|\zeta_{2,\boldsymbol{h}}|} > T\lambda_0^2\epsilon^2/16\right)$$
$$\leq P\left(\chi^2_{|\zeta|} > T\lambda_0^2\epsilon^2/16\right) \leq \exp(-T\lambda_0^2\epsilon^2/16).$$

Hence Φ_T is a exponentially consistent sequence of tests. □

Next, we provide a bound on the discrepancy between the true and fitted tensor.

Theorem 8.2. *Let* $\mathcal{K}(\theta) = -\log\{\Pi_T(\mathbf{\Gamma}_{(T)} : ||\mathbf{\Gamma}_{(T)} - \mathbf{\Gamma}^{(0)}_{(T)}||_\infty < \theta)\}$ *and* $\tilde{\boldsymbol{\gamma}}_{j,k,v_j,(T)} = (\gamma^{(1)}_{j,k,v_j,(T)}, ..., \gamma^{(R)}_{j,k,v_j,(T)})'$, *and* $\tilde{\boldsymbol{\gamma}}^{(0)}_{j,k,v_j,(T)} = (\gamma^{0(1)}_{j,k,v_j,(T)}, ..., \gamma^{0(R)}_{j,k,v_j,(T)})'$, $\gamma^{0(r)}_{j,k,v_j,(T)} = 0$ *for* $r \in \{R_0+1, \ldots, R\}$, $R > R_0$. *For* $k = 1, ..., m_{(T)}$, *assume that* $\Delta_{\boldsymbol{v},k}$ *is a positive root of the equations given, for all* $\boldsymbol{v} \in \mathcal{F}_1 \times \mathcal{F}_2$, *by*

$$x(x+||\tilde{\boldsymbol{\gamma}}^{(0)}_{2,k,v_2,(T)}||)\cdots(x+||\tilde{\boldsymbol{\gamma}}^{(0)}_{D,k,v_D,(T)}||) + ||\tilde{\boldsymbol{\gamma}}^{(0)}_{1,k,v_1,(T)}||x(x+||\tilde{\boldsymbol{\gamma}}^{(0)}_{2,k,v_2,(T)}||)\cdots(x+||\tilde{\boldsymbol{\gamma}}^{(0)}_{D,k,v_D,(T)}||)$$
$$+\cdots+x||\tilde{\boldsymbol{\gamma}}^{(0)}_{2,k,v_2,(T)}||\cdots||\tilde{\boldsymbol{\gamma}}^{(0)}_{D,k,v_D,(T)}|| - \theta = 0, \tag{8.1}$$

and $\Delta = \min_{\boldsymbol{v},k}\Delta_{\boldsymbol{v},k}$. *Then, for some constant* C,

$$\mathcal{K}(\theta) \leq \left(Rm_{(T)}\sum_{j=1}^{D} p_{j,(T)}\right)\ln\{(2\pi R)^{1/2}/(2\Delta)\} - \ln(C) + Rm_{(T)}\sum_{j=1}^{D}\ln\{\Gamma(a_\lambda)/\Gamma(a_\lambda + p_{j,(T)})\}$$
$$+ \sum_{k=1}^{m_{(T)}}\sum_{j=1}^{D}\sum_{r=1}^{R_0}(a_\lambda + p_{j,(T)})\ln\left[b_\lambda + \sum_{v_j=1}^{p_{j,(T)}}\{(\gamma^{0(r)}_{j,k,v_j,(T)})^2 + 2\Delta^2\}^{1/2}\right]$$
$$+ (R - R_0)m_{(T)}\sum_{j=1}^{D}(a_\lambda + p_{j,(T)})\ln\big(b_\lambda + p_{j,(T)}2^{1/2}\Delta\big).$$

Proof.

$$|\Gamma_{\boldsymbol{v},k,(T)} - \Gamma^{(0)}_{\boldsymbol{v},k,(T)}| = \left|\sum_{r=1}^{R}\gamma^{(r)}_{1,k,v_1,(T)}\cdots\gamma^{(r)}_{D,k,v_D,(T)} - \sum_{r=1}^{R}\gamma^{0(r)}_{1,k,v_1,(T)}\cdots\gamma^{0(r)}_{D,k,v_D,(T)}\right|$$
$$= \left|\sum_{r=1}^{R}\left\{(\gamma^{(r)}_{1,k,v_1,(T)} - \gamma^{0(r)}_{1,k,v_1,(T)})\prod_{j\neq 1}\gamma^{(r)}_{j,k,v_j,(T)} + \cdots + (\gamma^{(r)}_{D,k,v_D,(T)} - \gamma^{0(r)}_{D,k,v_D,(T)})\prod_{j\neq D}\gamma^{0(r)}_{j,k,v_j,(T)}\right\}\right|$$
$$\leq ||\tilde{\gamma}_{1,k,v_1,(T)} - \tilde{\gamma}^{(0)}_{1,k,v_1,(T)}||_2\prod_{j\neq 1}||\tilde{\gamma}_{j,k,v_j,(T)}||_2 + \cdots + ||\tilde{\gamma}_{D,k,v_D,(T)} - \tilde{\gamma}^{(0)}_{D,k,v_D,(T)}||_2\prod_{j\neq D}||\tilde{\gamma}^{(0)}_{j,k,v_j,(T)}||_2,$$

Note that (8.1) can be written as $\mathrm{g}_{\boldsymbol{v},k}(x) = 0$, where

$$\mathrm{g}_{\boldsymbol{v},k}(x) = a_{D,k,\boldsymbol{v}}x^D + \cdots + a_{1,k,\boldsymbol{v}}x - a_{0,k,\boldsymbol{v}}$$

and the $a_{j,k,\boldsymbol{v}}$'s are suitably chosen to match the coefficient of x^j in (8.1). By Cauchy's bound on the roots of polynomials, Eq. (8.1) has only one positive root, namely the real $\Delta_{\boldsymbol{v},k}$ that satisfies $\Delta_{\boldsymbol{v},k} \leq 1 + \max_{j=0,\ldots,D}|a_{j,k,\boldsymbol{v}}|$, for all $\boldsymbol{v}$ and k. From (8.1), the fact that $||\tilde{\gamma}_{j,k,v_j,(T)} - \tilde{\gamma}^{(0)}_{j,k,v_j,(T)}|| < \Delta$ for all $v_j \in \{1,\ldots,p_{j,(T)}\}$, $j \in \{1,\ldots,D\}$ and $k \in \{1,...,m_{(T)}\}$ implies

$$|\Gamma_{\boldsymbol{v},k,(T)} - \Gamma^{(0)}_{\boldsymbol{v},k,(T)}| \leq \mathrm{g}_{\boldsymbol{v},k}(\Delta) + \theta \leq \mathrm{g}_{\boldsymbol{v},k}(\Delta_{\boldsymbol{v},k}) + \theta = \theta,$$

which leads to $||\boldsymbol{\Gamma}_{(T)} - \boldsymbol{\Gamma}^{(0)}_{(T)}||_\infty < \theta$. Hence

$$\Pi_T(\boldsymbol{\Gamma}_{(T)} : ||\boldsymbol{\Gamma}_{(T)} - \boldsymbol{\Gamma}^{(0)}_{(T)}||_\infty < \theta) \geq \Pi_T(\forall_{k\in\{1,\ldots,m_{(T)}\}}\,\forall_{j\in\{1,\ldots,D\}}\,\forall_{v_j\in\{1,\ldots,p_{j,(T)}\}}\,||\tilde{\gamma}_{j,k,v_j,(T)} - \tilde{\gamma}^{(0)}_{j,k,v_j,(T)}||_2 < \Delta).$$

We will bound the right-hand side from below.

$$\Pi_T\left(\forall_{k\in\{1,\ldots,m_{(T)}\}}\,\forall_{j\in\{1,\ldots,D\}}\,\forall_{v_j\in\{1,\ldots,p_{j,(T)}\}}\,||\tilde{\gamma}_{j,v_j,T} - \tilde{\gamma}^{(0)}_{j,v_j,T}||_2 < \Delta \mid \forall_{k\in\{1,\ldots,m_{(T)}\}}\,\{\phi_{r,k}\},\tau_k,\{W_{jr,k}\}\right)$$

$$= \prod_{k=1}^{m_{(T)}} \prod_{j=1}^{D} \prod_{v_j=1}^{p_{j,(T)}} \left[\exp\left\{ -\sum_{r=1}^{R} (\gamma^{0(r)}_{j,k,v_j,(T)})^2/(2w_{jr,k,v_j}\phi_{r,k}\tau_k) \right\} \Pi_T \left(||\tilde{\boldsymbol{\gamma}}_{j,k,v_j,(T)}|| < \Delta/2 \mid \{\phi_{r,k}\}, \tau_k, \{W_{jr,k}\} \right) \right]$$

$$\geq \prod_{k=1}^{m_{(T)}} \prod_{j=1}^{D} \prod_{v_j=1}^{p_{j,(T)}} \Big[\exp\left\{ -\sum_{r=1}^{R} (\gamma^{0(r)}_{j,k,v_j,(T)})^2/(2w_{jr,k,v_j}\phi_{r,k}\tau_k) \right\} \prod_{r=1}^{R} \Big[\exp\{-\Delta^2/(\phi_{r,k}\tau_k w_{jr,k,v_j})\}$$

$$(2\Delta)/(2\pi R\phi_{r,k}\tau_k w_{jr,k,v_j})^{1/2} \Big] \Big]$$

$$\geq \prod_{k=1}^{m_{(T)}} \prod_{j=1}^{D} \prod_{v_j=1}^{p_{j,(T)}} \prod_{r=1}^{R} \left[(2\Delta)/(2\pi R\phi_{r,k}\tau_k w_{jr,k,v_j})^{1/2} \exp\Big[-\{\Delta^2 + (\gamma^{0(r)}_{j,k,v_j,(T)})^2/2\}/(\phi_{r,k}\tau_k w_{jr,k,v_j}) \Big] \right],$$

where Step 2 follows from Anderson's lemma. Integrating out the w_{jr,k,v_j}'s, we obtain

$$\Pi \left(\forall_{k\in\{1,\ldots,m_{(T)}\}} \forall_{j\in\{1,\ldots,D\}} \forall_{v_j\in\{1,\ldots,p_{j,(T)}\}} ||\tilde{\boldsymbol{\gamma}}_{j,k,v_j,(T)} - \tilde{\boldsymbol{\gamma}}^{(0)}_{j,k,v_j,(T)}|| < \Delta \mid \tau_k, \{\phi_{r,k}\}, \{\lambda_{jr,k}\} \right)$$

$$\geq \prod_{k=1}^{m_{(T)}} \prod_{r=1}^{R} \prod_{j=1}^{D} \left[\{(2\Delta\lambda_{jr,k})/(R\phi_{r,k}\tau_k)^{1/2}\}^{p_{j,(T)}} \exp\Big[-\lambda_{jr,k} \sum_{v_j=1}^{p_{j,(T)}} \{(\gamma^{0(r)}_{j,k,v_j,(T)})^2 + 2\Delta^2\}^{1/2}/(\phi_{r,k}\tau_k)^{1/2} \Big] \right].$$

Integrating out the $\lambda_{jr,k}$'s, we then get

$$\Pi_T \left(\forall_{k\in\{1,\ldots,m_{(T)}\}} \forall_{j\in\{1,\ldots,D\}} \forall_{v_j\in\{1,\ldots,p_{j,(T)}\}} ||\tilde{\boldsymbol{\gamma}}_{j,k,v_j,(T)} - \tilde{\boldsymbol{\gamma}}^{(0)}_{j,k,v_j,(T)}|| < \Delta \mid \tau_k, \{\phi_{r,k}\} \right)$$

$$\geq \prod_{k=1}^{m_{(T)}} \prod_{r=1}^{R} \prod_{j=1}^{D} \left[\{(2\Delta)/(R\phi_{r,k}\tau_k)^{1/2}\}^{p_{j,(T)}} \frac{\Gamma(a_\lambda + p_{j,(T)})}{\Big[b_\lambda + \sum_{v_j=1}^{p_{j,(T)}} \{(\gamma^{0(r)}_{j,k,v_j,(T)})^2 + 2\Delta^2\}^{1/2} (\phi_{r,k}\tau_k)^{-1/2} \Big]^{a_\lambda + p_{j,(T)}}} \right]$$

$$\{b_\lambda^{a_\lambda}/\Gamma(a_\lambda)\}^{R(D)}$$

$$\geq \prod_{k=1}^{m_{(T)}} \prod_{r=1}^{R} \prod_{j=1}^{D} \left[\{(2\Delta)/(R\phi_{r,k}\tau_k)^{1/2}\}^{p_{j,(T)}} \{b_\lambda^{a_\lambda}/\Gamma(a_\lambda)\} \frac{\Gamma(a_\lambda + p_{j,(T)})(\phi_{r,k}\tau_k)^{(a_\lambda + p_{j,(T)})/2} \mathbf{1}\{\tau_k \in (0,1)\}}{\Big[b_\lambda + \sum_{v_j=1}^{p_{j,(T)}} \{(\gamma^{0(r)}_{j,k,v_j,(T)})^2 + 2\Delta^2\}^{1/2} \Big]^{a_\lambda + p_{j,(T)}}} \right]$$

Integrating our $\phi_{r,k}$'s together we obtain,

$$\Pi_T \left(\forall_{k\in\{1,\ldots,m_{(T)}\}} \forall_{j\in\{1,\ldots,D\}} \forall_{v_j\in\{1,\ldots,p_{j,(T)}\}} ||\tilde{\boldsymbol{\gamma}}_{j,k,v_j,(T)} - \tilde{\boldsymbol{\gamma}}^{(0)}_{j,k,v_j,(T)}|| < \Delta \mid \tau_k \right)$$

$$\geq \prod_{k=1}^{m_{(T)}} \prod_{r=1}^{R} \prod_{j=1}^{D} \left[\{(2\Delta)/(R\tau_k)^{1/2}\}^{p_{j,(T)}} \{b_\lambda^{a_\lambda}/\Gamma(a_\lambda)\} \frac{\Gamma(a_\lambda + p_{j,(T)}) \tau_k^{(a_\lambda + p_{j,(T)})/2} \mathbf{1}\{\tau_k \in (0,1)\}}{\Big[b_\lambda + \sum_{i_j=1}^{p_{j,(T)}} \{(\gamma^{0(r)}_{j,k,v_j,(T)})^2 + 2\Delta^2\}^{1/2} \Big]^{a_\lambda + p_{j,(T)}}} \right]$$

$$\prod_{r=1}^{R-1} \left[\frac{Beta(D, \alpha_k + D(R-r))}{Beta(1, \alpha_k)} \right],$$

where $Beta(m_1, m_2)$ is the integrating constant for the Beta density with pa-

rameters m_1 and m_2. Finally, integrating out τ_k, leads to

$$\Pi_T(\forall_{k\in\{1,\ldots,m_{(T)}\}}\ \forall_{j\in\{1,\ldots,D\}}\ \forall_{v_j\in\{1,\ldots,p_{j,(T)}\}}\ ||\tilde{\gamma}_{j,k,v_j,(T)} - \tilde{\gamma}^{(0)}_{j,k,v_j,(T)}|| < \Delta)$$
$$\geq \prod_{k=1}^{m_{(T)}}\prod_{j=1}^{D}\{\Gamma(a_\lambda + p_{j,(T)})/\Gamma(a_\lambda)\}^R \prod_{k=1}^{m_{(T)}}\prod_{j=1}^{D}\prod_{r=1}^{R}\left[b_\lambda + \sum_{v_j=1}^{p_{j,(T)}}\{(\gamma^{0(r)}_{j,k,v_j,(T)})^2 + 2\Delta^2\}^{1/2}\right]^{-a_\lambda - p_{j,(T)}}$$
$$\{2\Delta/(2\pi R)^{1/2}\}^{Rm_{(T)}\sum_{j=1}^{D} p_{j,(T)}} C^{-1},$$

for some constant C. Hence

$$\mathcal{K}(\theta) \leq -\log\left[\Pi_T(\forall_{k\in\{1,\ldots,m_{(T)}\}}\ \forall_{j\in\{1,\ldots,D\}}\ \forall_{v_j\in\{1,\ldots,p_{j,(T)}\}}\ ||\tilde{\gamma}_{j,k,v_j,(T)} - \tilde{\gamma}^{(0)}_{j,k,v_j,(T)}|| < \Delta)\right]$$
$$\leq \left(Rm_{(T)}\sum_{j=1}^{D} p_{j,(T)}\right)\ln\{(2\pi R)^{1/2}/(2\Delta)\} - \ln(C) + Rm_{(T)}\sum_{j=1}^{D}\ln\{\Gamma(a_\lambda)/\Gamma(a_\lambda + p_{j,(T)})\}$$
$$+ \sum_{k=1}^{m_{(T)}}\sum_{j=1}^{D}\sum_{r=1}^{R_0}(a_\lambda + p_{j,(T)})\ln\left[b_\lambda + \sum_{v_j=1}^{p_{j,(T)}}\{(\gamma^{0(r)}_{j,k,v_j,(T)})^2 + 2\Delta^2\}^{1/2}\right]$$
$$+ (R - R_0)m_{(T)}\sum_{j=1}^{D}(a_\lambda + p_{j,(T)})\ln\big(b_\lambda + p_{j,(T)}2^{1/2}\Delta\big).$$

□

Under assumptions (a)-(f), the R.H.S is $o(T)$. Thus, we present the next theorem whose proof follows immediately from Theorem 8.2.

Theorem 8.3. *For any constant $\theta > 0$, under conditions (a)-(f) of Theorem 3.1, $\mathcal{K}(\theta) = o(T)$.*

Proof of Theorem 3.1

Proof.

$$\Pi_T(\mathcal{A}_T) = \frac{\int_{\mathcal{A}_T} f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\boldsymbol{\Gamma}_{(T)})\pi_T(\boldsymbol{\Gamma}_{(T)})}{\int f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\boldsymbol{\Gamma}_{(T)})\pi_T(\boldsymbol{\Gamma}_{(T)})} = \frac{\int_{\mathcal{A}_T}\frac{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\boldsymbol{\Gamma}_{(T)})}{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\boldsymbol{\Gamma}^{(0)}_{(T)})}\pi_T(\boldsymbol{\Gamma}_{(T)})}{\int\frac{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\boldsymbol{\Gamma}_{(T)})}{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\boldsymbol{\Gamma}^{(0)}_{(T)})}\pi_T(\boldsymbol{\Gamma}_{(T)})}$$
$$= \frac{\mathcal{N}_T}{\mathcal{D}_T} \leq \Phi_T + (1 - \Phi_T)\frac{\mathcal{N}_T}{\mathcal{D}_T}, \tag{8.2}$$

where Φ_T is the exponentially consistent sequence of tests given by Lemma 8.1. Note that

$$P_{\boldsymbol{\Gamma}^{(0)}_{(T)}}\left(\Phi_T > \exp(-T\lambda_0^2\epsilon^2/64)\right) \leq E_{\boldsymbol{\Gamma}^{(0)}_{(T)}}(\Phi_T)\exp(T\lambda_0^2\epsilon^2/64) \leq \exp(-T\lambda_0^2\epsilon^2/64).$$

Therefore $\sum_{T=1}^{\infty} P_{\boldsymbol{\Gamma}^{(0)}_{(T)}}\left(\Phi_T > \exp(-T\lambda_0^2\epsilon^2/64)\right) < \infty$. Applying Borel-Cantelli lemma $P_{\boldsymbol{\Gamma}^{(0)}_{(T)}}\left(\Phi_T > \exp(-T\lambda_0^2\epsilon^2/64)\text{ infinitely often}\right) = 0$. Thus,

$$\Phi_T \to 0 \quad a.s. \tag{8.3}$$

In addition, we have

$$E_{\mathbf{\Gamma}^{(0)}_{(T)}}((1-\Phi_T)\mathcal{N}_T) = \int (1-\Phi_T)\int_{\mathcal{A}_T}\frac{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{\Gamma}_{(T)})}{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{\Gamma}^{(0)}_{(T)})}\pi_T(\mathbf{\Gamma}_{(T)})f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{\Gamma}^{(0)}_{(T)})$$

$$= \int_{\mathcal{A}_T}\int (1-\Phi_T)f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{\Gamma}_{(T)})\pi_T(\mathbf{\Gamma}_{(T)}) \leq \sup_{\mathbf{\Gamma}_{(T)}\in\mathcal{A}_T} E_{\mathbf{\Gamma}_{(T)}}(1-\Phi_T) \leq \exp(-T\lambda_0^2\epsilon^2/16).$$

Applying Borel-Cantelli lemma, $P_{\mathbf{\Gamma}^{(0)}_{(T)}}\left((1-\Phi_T)\mathcal{N}_T\exp(T\lambda_0^2\epsilon^2/32) > \exp(-T\lambda_0^2\epsilon^2/64) \text{ infinitely often}\right) = 0$ so

$$\exp(T\lambda_0^2\epsilon^2/32)(1-\Phi_T)\mathcal{N}_T \to 0 \quad a.s.. \tag{8.4}$$

Note that $\mathcal{D}_T = \int \frac{f(\boldsymbol{Y}_1..,\boldsymbol{Y}_T|\mathbf{\Gamma}_{(T)})}{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{\Gamma}^{(0)}_{(T)})}\pi_T(\mathbf{\Gamma}_{(T)})$. Let $\tilde{b} = \lambda_0^2\epsilon^2/32$. Consider the set

$\mathcal{H}_T = \left\{\mathbf{\Gamma}_{(T)} : \frac{1}{T}\log\left[\frac{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{\Gamma}^{(0)}_{(T)})}{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{\Gamma}_{(T)})}\right] < \upsilon\right\}$, for $\upsilon = \tilde{b}/2$.

$$\exp(\tilde{b}T)\mathcal{D}_T \geq \exp(\tilde{b}T)\int_{\mathcal{H}_T}\exp\left(-T\frac{1}{T}\log\frac{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{\Gamma}^{(0)}_{(T)})}{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{\Gamma}_{(T)})}\right)\pi_T(\mathbf{\Gamma}_{(T)}) \geq \exp((\tilde{b}-\tilde{b}/2)T)\Pi_T(\mathcal{H}_T).$$

In view of (8.2), (8.3) and (8.4), it is enough to show that $-\log(\Pi_T(\mathcal{H}_T)) \leq T\tilde{b}/8$.

$$\frac{1}{T}\log\left[\frac{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{\Gamma}^{(0)}_{(T)})}{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{\Gamma}_{(T)})}\right] = \frac{1}{T}\Bigg[-\frac{1}{2}\sum_{\boldsymbol{v}}(\boldsymbol{y}_{\boldsymbol{v}} - \sum_{k=1}^{m_{(T)}}\mathbf{\Gamma}^{(0)}_{\boldsymbol{v},k,(T)}\boldsymbol{x}_k)'\boldsymbol{R}^{-1}(\boldsymbol{y}_{\boldsymbol{v}} - \sum_{k=1}^{m_{(T)}}\mathbf{\Gamma}^{(0)}_{\boldsymbol{v},k,(T)}\boldsymbol{x}_k)$$
$$+\frac{1}{2}\sum_{\boldsymbol{v}}(\boldsymbol{y}_{\boldsymbol{v}} - \sum_{k=1}^{m_{(T)}}\mathbf{\Gamma}_{\boldsymbol{v},k,(T)}\boldsymbol{x}_k)'\boldsymbol{R}^{-1}(\boldsymbol{y}_{\boldsymbol{v}} - \sum_{k=1}^{m_{(T)}}\mathbf{\Gamma}_{\boldsymbol{v},k,(T)}\boldsymbol{x}_k)\Bigg].$$

$$\Pi_T\left(\mathbf{\Gamma}_{(T)} : \frac{1}{T}\log\left[\frac{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{\Gamma}^{(0)}_{(T)})}{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\mathbf{\Gamma}_{(T)})}\right] < \upsilon\right)$$
$$\geq \Pi_T\left(\mathbf{\Gamma}_{(T)} : |\frac{1}{2T}\sum_{\boldsymbol{v}}\sum_{k=1}^{m_{(T)}}(\mathbf{\Gamma}_{\boldsymbol{v},k,(T)} - \mathbf{\Gamma}^{(0)}_{\boldsymbol{v},k,(T)})'\boldsymbol{x}_k'\boldsymbol{R}^{-1}\boldsymbol{x}_k(\mathbf{\Gamma}_{\boldsymbol{v},k,(T)} - \mathbf{\Gamma}^{(0)}_{\boldsymbol{v},k,(T)})| < \upsilon\right)$$
$$\geq \Pi_T\left(\mathbf{\Gamma}_{(T)} : ||\mathbf{\Gamma}_{(T)} - \mathbf{\Gamma}^{(0)}_{(T)}||_2^2 < 2\upsilon/\lambda_1^2\right)$$
$$\geq \Pi_T\left(\mathbf{\Gamma}_{(T)} : ||\mathbf{\Gamma}_{(T)} - \mathbf{\Gamma}^{(0)}_{(T)}||_\infty < \sqrt{2\upsilon/\lambda_1^2}\right) \geq \exp(-T\tilde{b}/8),$$

where the third line follows from assumption (e) of Theorem 3.1 and last inequality is immediate by applying Theorem 8.3. □

References

Armagan, A., Dunson, D. B., and Lee, J. (2013a). "Generalized double Pareto shrinkage." *Statistica Sinica*, 23(1): 119. 6, 7

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

Belitser, E. and Nurushev, N. (2015). "Needles and straw in a haystack: robust confidence for possibly sparse sequences." *arXiv preprint arXiv:1511.01803*. 4

Bro, R. (2006). "Review on multiway analysis in chemistry2000–2005." *Critical reviews in analytical chemistry*, 36(3-4): 279–293. 2

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, asq017. 6

Castillo, I., Rousseau, J., et al. (2015a). "A Bernstein–von Mises theorem for smooth functionals in semiparametric models." *The Annals of Statistics*, 43(6): 2353–2383. 4

Castillo, I., Schmidt-Hieber, J., Van der Vaart, A., et al. (2015b). "Bayesian linear regression with sparse priors." *The Annals of Statistics*, 43(5): 1986–2018. 10

Castillo, I., van der Vaart, A., et al. (2012). "Needles and straw in a haystack: Posterior concentration for possibly sparse sequences." *The Annals of Statistics*, 40(4): 2069–2101. 4

Chen, K., Dong, H., and Chan, K.-S. (2013). "Reduced rank regression via adaptive nuclear norm penalization." *Biometrika*, 100(4): 901–920. 2

Chen, L. and Huang, J. Z. (2012). "Sparse reduced-rank regression for simultaneous dimension reduction and variable selection." *Journal of the American Statistical Association*, 107(500): 1533–1545. 2

Chumbley, J. R. and Friston, K. J. (2009). "False discovery rate revisited: FDR and topological inference using Gaussian random fields." *Neuroimage*, 44(1): 62–70. 2

Clyde, M., Desimone, H., and Parmigiani, G. (1996). "Prediction via orthogonalized model mixing." *Journal of the American Statistical Association*, 91(435): 1197–1208. 6

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

Daubechies, I., Roussos, E., Takerkart, S., Benharrosh, M., Golden, C., D'ardenne, K., Richter, W., Cohen, J., and Haxby, J. (2009). "Independent component analysis for brain fMRI does not select for independence." *Proceedings of the National Academy of Sciences*, 106(26): 10415–10422. 17

Descombes, X., Kruggel, F., and Von Cramon, D. Y. (1998). "Spatio-temporal fMRI analysis using Markov random fields." *Medical Imaging, IEEE Transactions on*, 17(6): 1028–1039. 2

Dunson, D. B. and Xing, C. (2009). "Nonparametric Bayes modeling of multivariate categorical data." *Journal of the American Statistical Association*, 104(487): 1042–1051. 3

Friston, K. J., Ashburner, J., Frith, C. D., Poline, J.-B., Heather, J. D., Frackowiak, R. S., et al. (1995). "Spatial registration and normalization of images." *Human brain mapping*, 3(3): 165–189. 2

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL. 11

Genovese, C. R., Lazar, N. A., and Nichols, T. (2002). "Thresholding of statistical maps in functional neuroimaging using the false discovery rate." *Neuroimage*, 15(4): 870–878. 2

George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88(423): 881–889. 6

Gerard, D. and Hoff, P. (2015). "Adaptive Higher-order Spectral Estimators." *arXiv preprint arXiv:1505.02114*. 2, 3

Guhaniyogi, R. (2017). "Convergence rate of Bayesian supervised tensor modeling with multiway shrinkage priors." *Journal of Multivariate Analysis*, 160: 157–168. 4

Guhaniyogi, R., Qamar, S., and Dunson, D. B. (2017). "Bayesian tensor regression." *The Journal of Machine Learning Research*, 18(1): 2733–2763. 3, 7

Hans, C. (2009). "Bayesian lasso regression." *Biometrika*, 96(4): 835–845. 6

Kiers, H. A. (2000). "Towards a standardized notation and terminology in multiway analysis." *Journal of Chemometrics: A Journal of the Chemometrics Society*, 14(3): 105–122. 5

Kiers, H. A. and Mechelen, I. V. (2001). "Three-way component analysis: Principles and illustrative application." *Psychological methods*, 6(1): 84. 2

Kolda, T. G. and Bader, B. W. (2009). "Tensor decompositions and applications." *SIAM review*, 51(3): 455–500. 5

Li, L. and Zhang, X. (2015). "Parsimonious Tensor Response Regression." *arXiv preprint arXiv:1501.07815*. 3

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 (Statistical Methodology)*, 73(4): 559–578. 2

Martin, R., Mess, R., Walker, S. G., et al. (2017). "Empirical Bayes posterior concentration in sparse high-dimensional linear models." *Bernoulli*, 23(3): 1822–1847. 4

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R., and Wang, P. (2010). "Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer." *The annals of applied statistics*, 4(1): 53. 2

Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., and Nichols, T. E. (2011). *Statistical parametric mapping: the analysis of functional brain images: the analysis of functional brain images*. Academic press. 2

Polson, N. G. and Scott, J. G. (2010). "Shrink globally, act locally: Sparse Bayesian regularization and prediction." *Bayesian Statistics*, 9: 501–538. 6

Schonberg, T., Fox, C. R., Mumford, J. A., Congdon, E., Trepel, C., and Poldrack, R. A. (2012). "Decreasing ventromedial prefrontal cortex activity during sequential risk-taking: an fMRI investigation of the balloon analog risk

task." *Frontiers in neuroscience*, 6: 80. 13

Similä, T. and Tikka, J. (2007). "Input selection and shrinkage in multiresponse linear regression." *Computational Statistics & Data Analysis*, 52(1): 406–422. 2

Smith, M. and Fahrmeir, L. (2007). "Spatial Bayesian variable selection with application to functional magnetic resonance imaging." *Journal of the American Statistical Association*, 102(478): 417–431. 3

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., et al. (2004). "Advances in functional and structural MR image analysis and implementation as FSL." *Neuroimage*, 23: S208–S219. 15

Song, Q. and Liang, F. (2017). "Nearly optimal Bayesian shrinkage for high dimensional regression." *arXiv preprint arXiv:1712.08964*. 4, 9, 10, 20

Sun, W. W. and Li, L. (2017). "STORE: sparse tensor response regression and neuroimaging analysis." *The Journal of Machine Learning Research*, 18(1): 4908–4944. 3

Van Der Pas, S., Kleijn, B., Van Der Vaart, A., et al. (2014). "The horseshoe estimator: Posterior concentration around nearly black vectors." *Electronic Journal of Statistics*, 8(2): 2585–2618. 4

Van der Vaart, A. W. and Van Zanten, H. (2009). "Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth." *The Annals of Statistics*, 37(5B): 2655–2675. 8

— (2011). "Information rates of nonparametric Gaussian process methods." *Journal of Machine Learning Research*, 12(Jun): 2095–2119. 8

Wei, R. and Ghosal, S. (2017). "Contraction properties of shrinkage priors in logistic regression." *Preprint at http://www4. stat. ncsu. edu/˜ ghoshal/papers*. 4

Welvaert, M., Durnez, J., Moerkerke, B., Verdoolaege, G., and Rosseel, Y. (2011). "neuRosim: An R package for generating fMRI data." *Journal of Statistical Software*, 44(10): 1–18. 11

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007). "Dimension reduction and coefficient estimation in multivariate linear regression." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3): 329–346. 2

Zhang, L., Guindani, M., and Vannucci, M. (2015). "Bayesian models for functional magnetic resonance imaging data analysis." *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(1): 21–41. 2

Zhang, L., Guindani, M., Versace, F., and Vannucci, M. (2014). "A spatio-temporal nonparametric Bayesian variable selection model of fMRI data for clustering correlated time courses." *NeuroImage*, 95: 162–175. 2

Zhou, H. and Li, L. (2014). "Regularized matrix regression." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2): 463–483. 3

Zhou, H., Li, L., and Zhu, H. (2013). "Tensor regression with applications in neuroimaging data analysis." *Journal of the American Statistical Association*, 108(502): 540–552. 3

Zhu, H., Fan, J., and Kong, L. (2014). "Spatially varying coefficient model

for neuroimaging data with jump discontinuities." *Journal of the American Statistical Association*, 109(507): 1084–1098. 2