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 measure, 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 of the blood-oxygen-level-dependent measure (BOLD

measure) are acquired over multiple time points across all voxels while a subject performs multiple tasks, yielding three dimensional tensor responses over time points. One could also examine a slice of an fMRI scan, resulting in a two-dimensional tensor for each time point. The BOLD measure 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 regions 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 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. 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 the 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 voxels within a tensor response together, many neuroscientists use what is referred to as the General Linear Model (GLM), which is not to be confused with the generalized linear model. The GLM 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). The GLM is conceptually simple and computationally efficient, though many multiple testing corrections do not take spatial relationships into account. Corrections that do take the spatial relationships between voxels into account require tuning and make strong assumptions about spatial structures (Poline et al., 1997). Additionally, neuroimaging data are usually pre-processed using a kernel convolution based spatial smoothing approach. using the GLM 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 retain some 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 cells (Zhang et al., 2015, 2014; Descombes et al., 1998; Zhu et al., 2014). 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 $p_1, \ldots, p_D$ 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. There is a parallel literature to model spatial dependence among regression coefficients induced by Markov random fields (MRF) (Smith and Fahrmeir, 2007; Zhang et al., 2015). MRF models are generally efficient in computation, though our proposed approach seems to yield somewhat better inference in comparison with a specific Gaussian MRF model (see Section 4). Intuitively, introducing tensor response in the regression without reshaping it seems to preserve the neighborhood information of the cells in the response. A more sophisticated MRF model may improve the inference, with perhaps added computational cost.

Recently, Li and Zhang (2017) propose a novel approach of regressing the tensor variate response on scalar predictors, where the *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 more seamlessly 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 (2017)), 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 (2017), 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" canonical decomposition/parallel factorization decomposition (CANDECOMP/PARAFAC, or CP), which is defined in Section 2.1. The model is also designed to be generalizable to any value of D for possible application in other research areas. 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 reduceing 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 recent literature on regressing a scalar response on a tensor covariate (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. Sections 4 and 5 show performance of the proposed model through simulation studies and brain activation data analysis, respectively. Section 6 concludes the paper. Details of posterior computation including the full MCMC implementation of the model can be found in the supplementary material.

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 follows that $vec(\boldsymbol{\gamma}_1 \circ \boldsymbol{\gamma}_2 \circ \cdots \circ \boldsymbol{\gamma}_D) = \boldsymbol{\gamma}_D \otimes \cdots \otimes \boldsymbol{\gamma}_1$, where $\otimes$ represents the Kronecker product. 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 products of each mode-k fibers of $\boldsymbol{\Gamma}$ with $\boldsymbol{a}$.

A D-way tensor $\boldsymbol{\Gamma} \in \otimes_{j=1}^{D} \Re^{p_j}$ assumes a rank-R parallel factorization (PARAFAC) decomposition (Kiers, 2000; Kolda and Bader, 2009) 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 $1 \leq r \leq R$. The 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). Part of the utility of this decomposition in many sparse tensor scenarios is that it allows for contiguous areas with the response to have little or no association with the covariate, which is reasonable in the context of the application in neuroscience (Li and Zhang, 2017). This is also visualized as the higher dimensional analogue to the singular value decomposition of matrices.

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 $\boldsymbol{v} = (v_1, ..., v_D)'$ represents the position of cell $\boldsymbol{v}$ in the D dimensional array of cells. Let $\boldsymbol{x}_t = (x_{1,t}, ..., x_{m,t})' \in \mathcal{X} \subset \Re^m$ 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 $t = 1, ..., T$. $\boldsymbol{\Gamma}_k \in \otimes_{j=1}^{D} \Re^{p_j}$, $k = 1, .., m$ is the tensor coefficient corresponding to the predictor $x_{k,t}$. To account for the temporal correlation of the response tensor at each cell $\boldsymbol{v}$, the error is assumed to follow a component-wise first-order autoregressive structure, $E_{t,\boldsymbol{v}} = \kappa_{\boldsymbol{v}}\, E_{t-1,\boldsymbol{v}} + \eta_{t,\boldsymbol{v}}$, where $\kappa_{\boldsymbol{v}} \in (-1, 1)$ is the autocorrelation coefficient corresponding to the cell $\boldsymbol{v}$. This parametrization allows varying temporal correlation in different voxels. However, $\kappa_{\boldsymbol{v}}$ parameters are weakly identifiable and thus introducing a large number of voxel specific $\kappa_{\boldsymbol{v}}$ parameters is perhaps not worth the extra burden both due to computation and due to estimation of these parameters. In fact, in many fMRI data application for brain imaging studies, one often transforms the voxel specific response vector and predictor matrices using the wavelet transform, and perform inference on the model parameters based on the transformed data. Wavelet transforms have the advantage of whitening the data, i.e., reducing the temporally correlated errors to i.i.d. errors (Bullmore et al., 2004; Fadili and Bullmore, 2002; Meyer, 2003; Zhang et al., 2015). In the light of this, we refrain from modeling $\boldsymbol{E}_t$ using complex temporal correlation structure and simplify the temporal correlation of $\boldsymbol{E}_t$ by setting $\kappa_{\boldsymbol{v}} = \kappa$, for all $\boldsymbol{v}$. Under this simplification, 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.

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 a 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, $1 \le r \le R$, $1 \le j \le D$ 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 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 $\mathbf{\Gamma}_k$.

Imposing this additional rank-R PARAFAC structure on $\mathbf{\Gamma}_k$ 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 predictors and predictor specific local parameters. The global parameter imposes shrinkage globally while local parameters carefully balance shrinkage for large and small coefficients.

The literature on the vector shrinkage priors provides an excellent starting point for studying multiway shrinkage priors on the tensor coefficient $\mathbf{\Gamma}_k$, though the latter presents a lot more challenges. Assuming that $\mathbf{\Gamma}_k$ admits a rank-R PARAFAC decomposition, proposing a prior on $\mathbf{\Gamma}_k$ is equivalent to specifying priors over tensor margins $\boldsymbol{\gamma}_{j,k}^{(r)}$. Given that every cell coefficient in $\mathbf{\Gamma}_k$ 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 tensor valued parameters 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 $\mathbf{\Gamma}_k$ leads to inaccurate estimation due to less desirable tail behavior of the distribution of $\Gamma_{\boldsymbol{v},k}$ parameters. The aberrant tail behavior results from the exchangeability of the rank-specific variance parameters across ranks in the M-DGDP prior, which results in convergence issues for the voxel parameters within $\mathbf{\Gamma}_k$, when using Markov Chain

Monte Carlo methods to draw from their posterior distributions. In contrast, we will propose a prior construction that prevents the rank-specific variance components from effectively switching values back and forth, leading to fast convergence of the voxel-specific parameters.

We propose a multiway stick breaking shrinkage prior on $\mathbf{\Gamma}_k$ 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 $\tau_k \sim$ Inverse Gamma(a_τ, b_τ). 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

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

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. The parameter 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 of $\boldsymbol{W}_{jr,k}$. 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 $\mathbf{\Gamma}_k$ for a relatively large number of cells compared to the ordinary spike and slab prior on cell coefficients.

An important aspect of these models is the selection of the model rank R. Since rank-R PARAFAC decomposition is a low-rank decomposition of the tensor, we can think about selecting R as selecting the truncation level of a large dimensional model. As long as R is large enough, the results should be robust to our choice (and further increasing R will lead to negligible changes in the results). Natural analogies are the truncation of a Dirichlet process mixture model, which results in an (approximate) finite mixture model (e.g., see Ishwaran and James (2001)), and the truncation of the stick-breaking construction of the Indian Buffet process (Teh et al., 2007). In practice, it is recommended that different rank-models be compared with some criterion in order to decide which rank model will be used to perform inference. Accordingly, in the real data application, we have fitted the model with a range of R values. The specific choice of R that corresponds to the lowest Deviance Information Criterion (DIC) (Gelman et al., 2014) is used.

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 Metropolis within Gibbs sampling algorithms are provided in the supplementary material.

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 coefficient $\boldsymbol{\Gamma}_{1,(T)}, .., \boldsymbol{\Gamma}_{m_{(T)},(T)}$. Define $\boldsymbol{\Gamma}$ as a $\Re^m \otimes_{j=1}^{D} \Re^{p_j}$ tensor with the $(v_1, .., v_D, k)$th cell 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 and 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. Indeed, assuming parameters σ^2 and κ to be known for the theoretical study is somewhat restrictive, but not very impractical as 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). The parameter κ belongs to $(-1, 1)$ and is assigned a uniform prior with bounded support. Although we assign inverse-gamma prior with an unbounded prior support for σ^2, truncating the prior distribution at a large value would not have perhaps altered the results much. Thus, this is not a strict restriction, but a mild assumption to simplify calculation. Moreover, assuming parameters within the variance structure fixed and known is a common practice in asymptotic studies, see Van der Vaart and Van Zanten (2011). 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 notation, 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 $\mathcal{F}_1 = \left\{\boldsymbol{h}_1 = (v_1, ..., v_D) : 1 \leq v_1 \leq p_{1,(T)}, ..., 1 \leq v_D \leq p_{D,(T)}\right\}$,

$\mathcal{F}_2 = \{h_2 = v_{D+1} : 1 \leq v_{D+1} \leq m_{(T)}\}$. 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 $\zeta_1^{(0)} = \{\boldsymbol{h}_1 \in \mathcal{F}_1 : \Gamma^{(0)}_{\boldsymbol{h}_1,h_2,(T)} \neq 0, \text{for some } h_2 \in \mathcal{F}_2\}$. 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 values 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 $a_T = \frac{\epsilon}{2p_{(T)}}$.

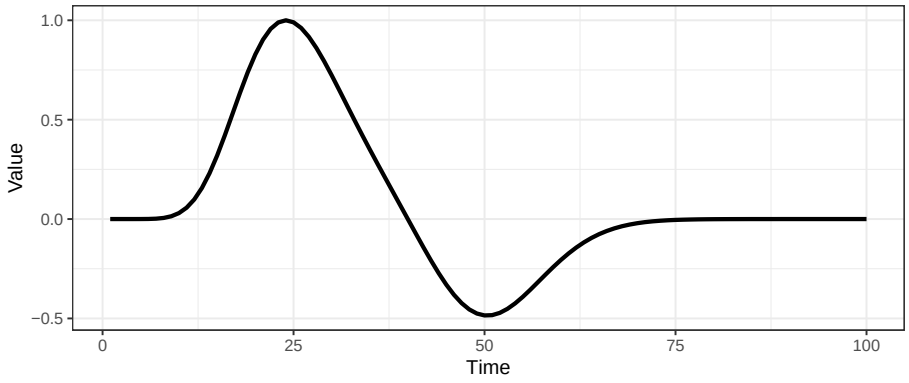

FIG 1. *Values taken by the simulated covariate through time when the number of total time steps was set to* $T = 100$.

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

4. 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 which closely mimic the real world fMRI data. Scalar predictors are simulated with the block experimental design. A single stimulus block is convolved with the canonical double-gamma haemodynamic response function. A more thorough discussion on how the covariate values are generated can be found below.

The block design consists of a single discrete epoch of activity and rest, with "activity" representing a period of stimulus presentation, and "rest" referring to a state of rest or baseline. The stimulus is assumed to take place at time $t = 0$ for a duration of one time step, with a stimulus value of 1. This is done to assure that results from simulated datasets with different choices of T (length of the time series) can be compared, as even data sets with small values of T would have a covariate that exhibits a peak in the stimulus function.

For a specific value of T, the covariate is generated using the `canonicalHRF` function in the `neuRosim` package in R (Welvaert et al., 2011) in which the delay of response relative to onset is $T \times 0.12$, the delay of undershoot is $T \times 0.5$, the dispersion of response is set to 2, the dispersion of undershoot is set to 1, and the scale of undershoot is set to 0.5. This setup is used so that simulations can be performed under different values of T without affecting the number of stimulus blocks in the simulated data and without changing the relative pattern of the simulated covariate values. An example of the values taken by the covariate for $T = 100$ can be seen in Figure 1.

The response tensor is simulated from (2.2) with $D = 2$, and $\sigma^{(0)2} = 1$. The true coefficient tensor $\mathbf{\Gamma}^{(0)}$ 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 a shared parameter μ. The nonzero elements of $\mathbf{\Gamma}^{(0)}$ are generated fixing a certain contrast-to-noise ratio η, which is defined as $\eta = \frac{\max_{\gamma^{(0)} \in \mathbf{\Gamma}^{(0)}} |\gamma^{(0)}|}{\sigma^{(0)}}$. In our simulations, the noise variance $\sigma^{(0)2}$ is set to 1. Different simulation scenarios are created by constructing a grid over different values for $T \in \{20, 50, 100, 200\}$, $\mu \in \{5, 10, 20, 30\}$, and $\eta = \{0.1, 0.25, 0.5, 0.75, 1, 1.5\}$.

Competitors. The model is fitted in each simulation scenario along with the classical General Linear Model (GLM), in which independent linear regressions are fitted with serially correlated errors using an AR(1) specification for each cell. Comparison with the maximum likelihood estimator of each cell of $\mathbf{\Gamma}$ thus obtained with the estimate of $\mathbf{\Gamma}$ from BTRR, will highlight the potential advantages of joint Bayesian modeling with tensor coefficients. In fitting the General Linear Model with AR(1) error structure for each cell, the `cochrane.orcutt` function in the `orcutt` package in R (Stefano et al., 2018) is used. It performs the iterative process necessary to estimate the values of $\mathbf{\Gamma}$ and κ. The estimates from the GLM are corrected by fixing the false positive rate to be 0.05 using the multiple testing correction porposed by Benjamini and Hochberg (1995). All coefficient estimates that are not deemed to be significantly different from zero are set equal to zero for any point estimates of the tensor response.

We also include another competitor that models spatial dependence among tensor cells through a Gaussian Markov Random Field (GMRF) (Zhang et al., 2015; Gössl et al., 2001; Quirós et al., 2010). More specifically, a GMRF prior

is assigned on the vectorized elements of $\mathbf{\Gamma}$ as following,

$$\begin{aligned} \text{vec}(\mathbf{\Gamma}) &\sim \text{N}(\mathbf{0}, (\lambda \mathbf{Q})^{-1}), \\ \lambda &\sim \text{Gamma}(a_\lambda, b_\lambda), \\ Q &= \begin{cases} n_v, & v = \ell \\ -1, & v \sim \ell \\ 0, & \text{otherwise} \end{cases}, \end{aligned}$$

where n_v is the number of neighbors of element v and $v \sim \ell$ denotes that elements v and ℓ are neighbors (Zhang et al., 2015). In fitting the GMRF model, a_λ and b_λ are set to 1 and 0.001 respectively, to match the noninformative prior in the BTRR model.

For the Bayesian models, the log-likelihood is examined in order to verify that the Markov chain converges. The models witness rapid convergence, so that only 1,100 draws are taken from the joint posterior distribution in each model fitting, out of which the first 100 draws are discarded as burn-ins. Average effective sample sizes shown in Figure 2 for the $1,000$ post burn-in samples calculated using the `coda` package in `R` confirm sufficiently uncorrelated post burn-in samples.

Point estimation of $\mathbf{\Gamma}$. Due to the continuous nature of the multiway stick breaking prior on $\mathbf{\Gamma}$, posterior mean estimate of all cells in $\mathbf{\Gamma}$ from BTRR turn out to be nonzero. In general, shrinkage priors in high dimensional regression are not designed to perform variable selection, and hence a post processing step on posterior samples of parameters is required to identify important and unimportant variables. Following Li and Pati (2017), we employ a two-stage variable selection algorithm that identifies zero and nonzero cells of $\mathbf{\Gamma}$ from its post burn-in MCMC samples. For cells of $\mathbf{\Gamma}$ identified as zero by the algorithm, the final point estimates are taken to be zero. We keep the posterior mean as point estimates for cells of $\mathbf{\Gamma}$ identified as nonzero. A comparison of the two stage estimates 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 it represents higher tensor dimensions and smaller sample size. As the signal-to-noise ratio tends to be very low for fMRI data (Welvaert and Rosseel, 2013), the model is attractive in this application if it performs better in such settings. The simulation settings under higher signal-to-noise are less realistic and are kept only to show relative performance of all models under varying signal-to-noise ratio. Also, since the main goal being the identification of activated cells, it is not primarily the estimated cell coefficients, but the contrast between coefficients corresponding to "active" and "inactive" voxels that is more informative. However, to show the effect of the shrinkage priors in the proposed Bayesian tensor response regression and the Gaussian Markov Random Field models, the scales are shown in Figure 3. The scales are kept separate due to the differing results in terms of both the contrast-to-noise ratio (η) and the effects of the different model treatments on the shrinkage of the activation estimates.

The true and the estimated activation maps in Figure 3 demonstrate desirable performance of BTRR in capturing the true activation pattern under

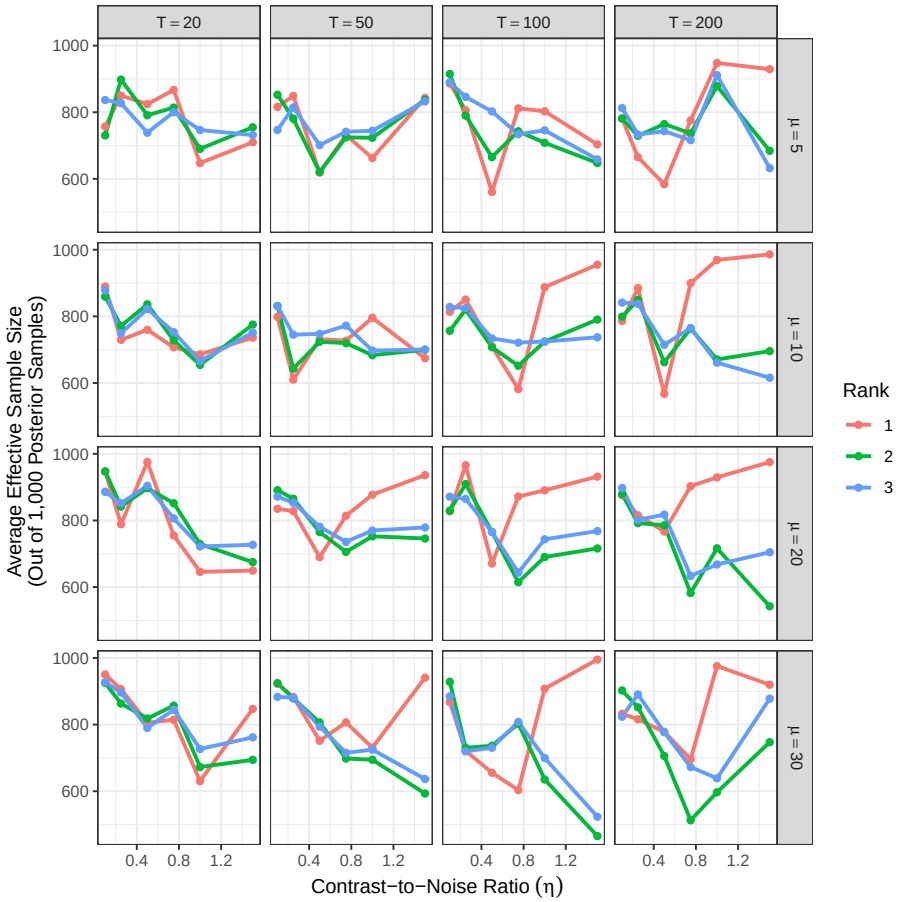

FIG 2. *The average effective sample size for elements of* $\mathbf{\Gamma}$ *under each of 288 scenarios.*

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 the 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 3 shows the improvement in the proposed model of shrinking insignificant coefficient estimates toward zero. This is done using a strong shrinkage prior, which also draws down the values for the significant coefficients. Such an observation is not entirely uncommon even with shrinkage priors in ordinary ultra-high dimensional regressions, see Bhattacharya et al. (2016). Moreover, as mentioned before, in the interest of detecting regions of nonzero coefficients, the contrast between the zero and nonzero estimates is more informative. In this regard, we argue in the next

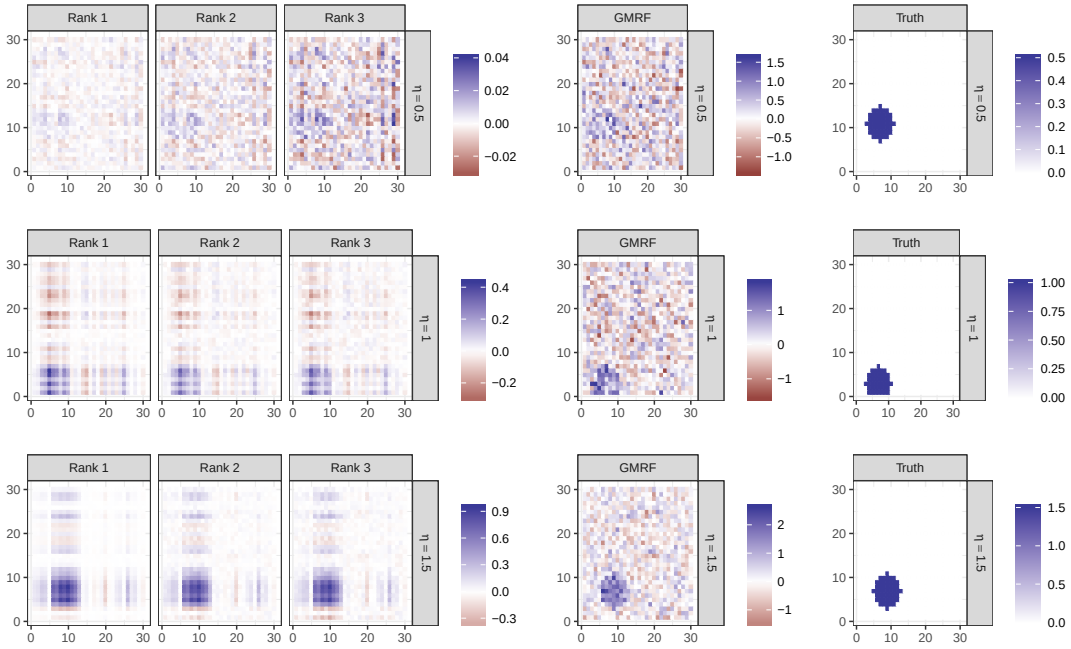

FIG 3. *Posterior mean and true values for* $\mathbf{\Gamma}^0$ *when* $\mu = 30$ *and* $T = 20$ *under different values for* R *and* η*. For comparison, the posterior mean estimate from a Gaussian Markov Random Field (GMRF) is also included.*

few paragraphs that the proposed model outperforms the GMRF both in terms of the overall estimation of cell coefficients and in identifying truely activated cells. The coefficient estimates from our approach can be somewhat improved by more involved choices of prior distribution on α_k. However, the little gain in the estimation of cell coefficients is perhaps not worth extra computational burden, since the inference on identifying activated cells does not change much with the more involved choices.

Figure 4 shows the standardized mean square error for estimates of $\mathbf{\Gamma}$, which is found using the formula

$$\text{sMSE} = \frac{1}{\text{var}\left(\mathbf{\Gamma}^{(0)}\right)} \times \frac{||\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}^{(0)}||^2}{\#\text{ voxels in } \mathbf{\Gamma}^{(0)}},$$

where $\text{var}\left(\mathbf{\Gamma}^{(0)}\right) = ||\mathbf{\Gamma}^{(0)} - \text{mean}(\mathbf{\Gamma}^{(0)})||^2/\#$ voxels in $\mathbf{\Gamma}^{(0)}$. In each scenario, BTRR model with ranks (R) 1, 2, and 3 are tested, and further testing suggests that additional ranks do not improve the performance. For the Bayesian models, that is, the proposed Bayesian tensor response regression (BTRR) and the Gaussian Markov random field (GMRF) models, the sequential 2-means method

of post-hoc variable selection method (Li and Pati, 2017) is used to produce the final point estimate. The General Linear Model (GLM) uses the Benjamini-Hochsberg multiple testing correction (Benjamini and Hochberg, 1995) to produce the final point estimate. In a real data application, the final rank used to fit a BTRR model can be selected using the Deviance Information Criterion (DIC) (Gelman et al., 2014). The model fitted with ranks 1, 2, and 3 are compared to the *General Linear Model* and the GMRF model described above.

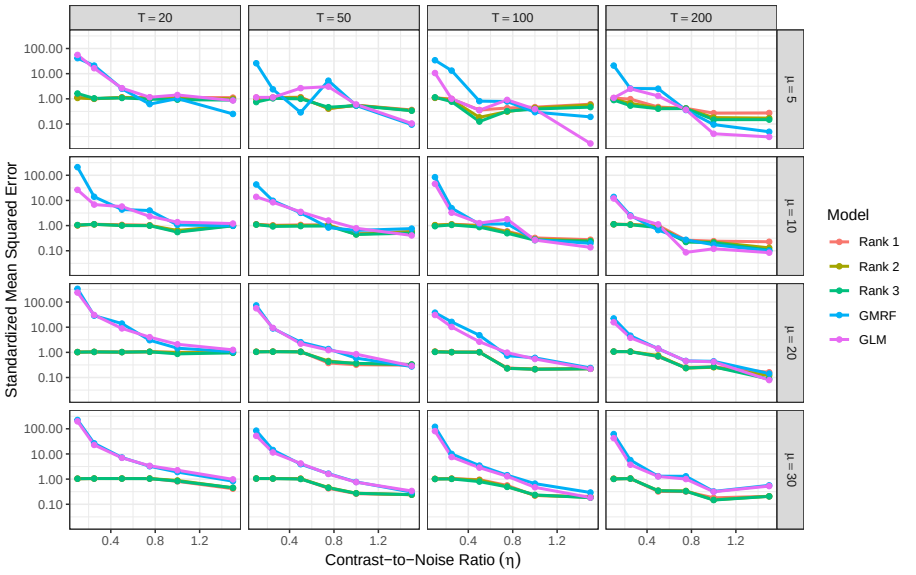

FIG 4. *The standardized mean square error from analyses on simulated data compared to results from the Gaussian Markov Random Field (GMRF) model and the General Linear Model (GLM).*

It is evident that the proposed model performs significantly better than all its competitors in terms of point estimation under low signal-to-noise ratio (see Figure 4), which perhaps justifies its usefulness for real fMRI data. In fact, the standardized MSE values in Figure 4 shows that the proposed method also performs better than its competitors under high signal-to-noise scenarios, except for cases where magnitude of the true simulated tensor has small number of cells (first two rows of Figure 4) and T is large. In fact, the only two specific scenarios under high signal-to-noise ratio where GLM does significantly better than BTRR can be found in the last two columns of the first row of Figure 4. Both these settings correspond to a large value of T and smaller tensor dimensions, which are favorable to asymptotic settings, and thus the carefully constructed shrinkage mechanism is no longer advantageous to the naive GLM. In fact, we have included these scenarios to provide an idea about the settings where the model does not necessarily have edge over its competitors.

Digging a bit deeper, Figure 5 presents the False Positive Rates (FPR) for

identifying nonzero cells in $\mathbf{\Gamma}$ from competing methods. Although the FPR values are generally low for all competing methods under all scenarios, BTRR tends to offer lower FPRs than the GMRF under scenarios with low-signal to noise ratios. The regularization imposed by the shrinkage prior helps in shrinking unimportant cell coefficients, which appears to yield advantages over GMRF, especially in presence of highly noisy data.

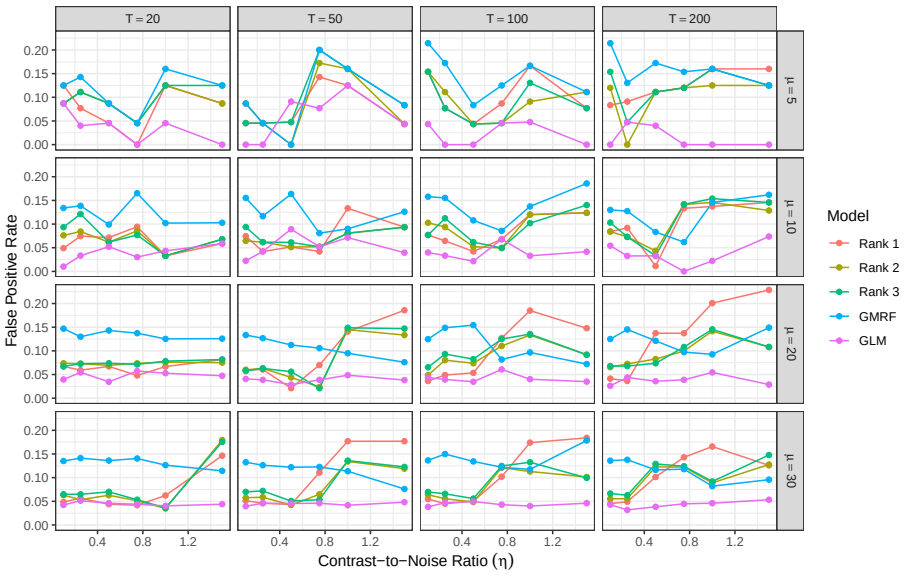

FIG 5. *False Positive Rates (FPR) for the GLM, GMRF model and BTRR models with different ranks.*

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 the cells of $\mathbf{\Gamma}$, shown in Figures 6 and 7, respectively. Given that the coverage of the 95% credible intervals in almost all the scenarios 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 sample size. 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 a lot of zero coefficients involves more uncertainty. Finally, there is a drop in coverage for the 95% posterior credible intervals as the contrast-to-noise ratio increases, especially for small tensor dimensions. This is due to the fact that the shrinkage priors, in their attempt to estimate cell coefficients with zero effects, lead to a little underestimation of the nonzero cell coefficients. When nonzero cell coefficients are higher in magnitude, this shrinkage effect becomes more prominent. The drop in coverage can be attributed to this phenomenon. It is also important to note that a rank-1 decomposition of the tensor coefficient is more restrictive structurally than rank-2

or rank-3 coefficients, and hence the drop in coverage is more severe in rank-1 decomposition.

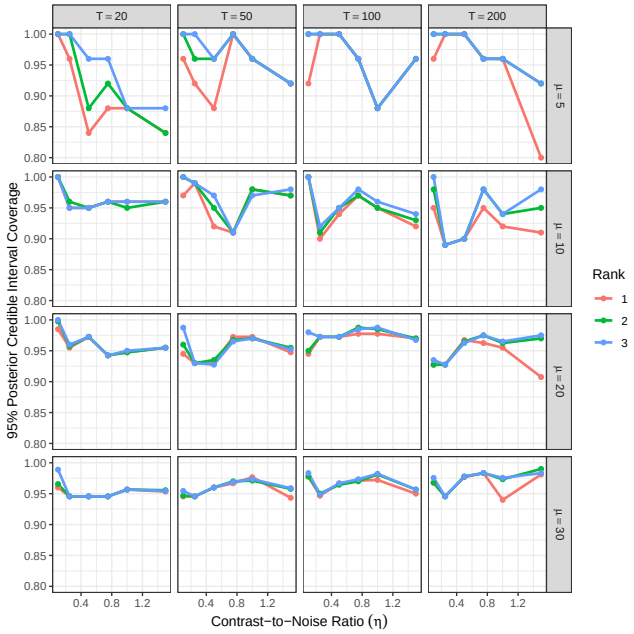

FIG 6. *The average coverage of the 95% posterior credible intervals for the posterior draws for the elements of* $\mathbf{\Gamma}$ *under varying conditions.*

Simulated data with model mis-specification. To assess performance of the BTRR model along with the GMRF competitor under model mis-specification, a simulated data set is created in which the nonzero-valued components within $\mathbf{\Gamma}$ are not sparse or spatially-contiguous. This would present a problem for the model, as the spatial correlation between elements in the tensor coefficient would essentially be meaningless, and the shrinkage imposed by the model would be an incorrect treatment of this data. In this example, the values within $\mathbf{\Gamma}$ are simulated independently from a Bernoulli(0.9) distribution, and the response in each cell is simulated independently from a linear regression model with errors auto-correlated in an AR(1) structure. The hyper-parameters used in the MCMC are the same used in other simulated data analyses. Figure 8 shows the final estimates of $\mathbf{\Gamma}$ from the GMRF model and the BTRR models with ranks 1, 2 and 3. The standardized Mean Squared Error (sMSE) for the BTRR models with ranks 1, 2, and 3 and the GMRF model are 1.30, 0.93, 0.92, and 1.02, respectively. This illustrates the point that while the BTRR models do not perform well estimating the dense, random coefficient tensor, the estimates are able to approximate results close to those of the mean model, which has an sMSE of 1.

Posterior estimates of autoregression parameter κ. Accurately estimating the

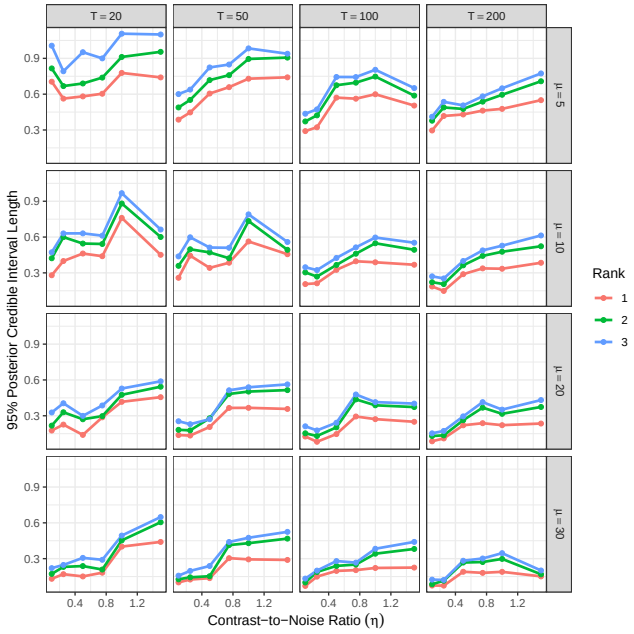

FIG 7. *The average length of the 95% posterior credible intervals for the posterior draws for the elements of* $\mathbf{\Gamma}$ *under varying conditions.*

posterior distribution of the autoregression parameter κ becomes essential since it captures the temporal correlation among tensor responses. We present posterior densities of κ for BTRR with ranks 1,2 and 3 for a representative simulation scenario with $T = 100$, $\mu = 30$ and $\eta = 1$ in Figure 9. The posterior mean of κ is found to be estimated close to the truth with a narrow 95% credible interval. The conclusion remains true in all other simulation scenarios.

The simulation study conclusively establishes 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 decisively outperforming the GLM estimates in smaller contrast to noise ratios reminiscent of real fMRI data. An application of the model to a dataset on brain activation study is explored in the following section.

5. Application to Balloon Analog Risk Taking Data

Neuroscientists at the University of California Los Angeles have 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) are subjects in an experiment with the following design. Each subject enters an fMRI machine with a computer display and a controller with two buttons. On the screen, the image of a balloon is shown,

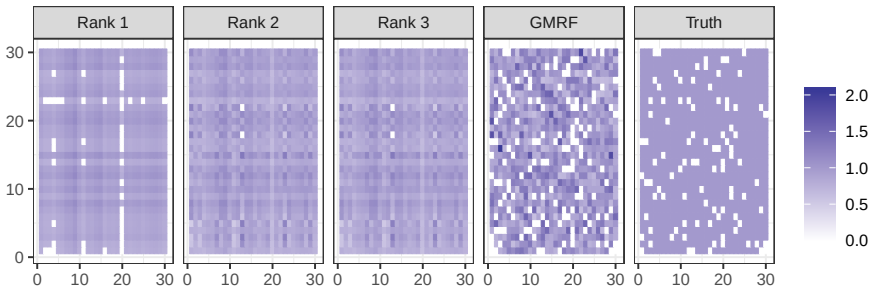

FIG 8. *Final estimates of* $\mathbf{\Gamma}$ *from BTRR and GMRF when* $\mathbf{\Gamma}^{(0)}$ *is constructed under a misspecified model.*

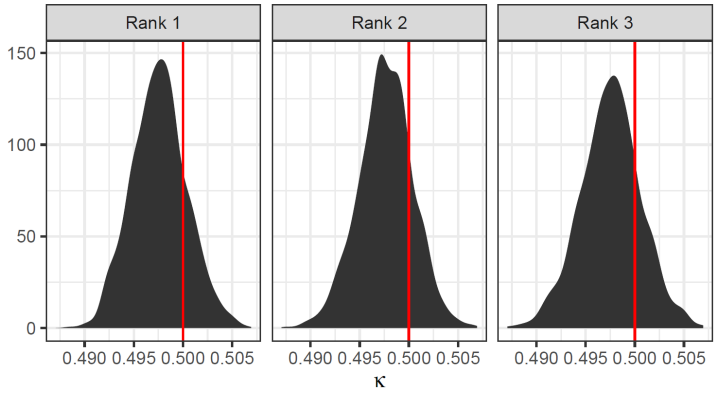

FIG 9. *Plots of the posterior densities of the autoregression coefficient for different ranks when data are simulated under for* $T = 100$, $\mu = 30$ *and* $\eta = 1$. *The true value for the autoregression coefficient is indicated with a red line.*

along with a payout amount, starting with a value of \$0.25. The buttons on the controller allow the subject to either inflate the balloon or take the payout. If the subject inflates the balloon, and the balloon does not explode, the payout amount increases by \$0.25. If the subject inflates the balloon, and the balloon explodes, no payout is received, the payout value is reset to \$0.25, and a new balloon is displayed. Balloons are assigned a number of pumps at which the balloon will 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 is red, green, or blue, respectively. A grey "control" balloon, offering no payout and an upper bound of 12 pumps before exploding, is also part of the trial to record a riskless scenario. Each subject participates in three runs. Each run consists of either 10 minutes, or 48 balloons exploding, whichever comes first. Note that in our real data application, we only work with data from a single run for one randomly chosen subject, since the proposed BTRR method is designed for single-subject scenarios.

Preprocessing was done using FSL (Smith et al., 2004) following what was done by Schonberg et al. (2012) as closely as possible. The fMRI have a repetition time (TR) of 2 seconds. In order to allow for T1 equilibrium effects, the first two images in the series were dropped. The echo planar imaging (EPI) scan was motion corrected, then high-pass filtered using a Gaussian least-squares linear fit with sigma = 50.0 seconds. While it is common for researchers to prewhiten their data in order to remove temporal autocorrelation, we followed the preprocessing steps followed by Schonberg et al. (2012), which did not include prewhitening. Thus, we make an assumption that the temporal autocorrelation does not vary significantly within the regions of interest in the study data. Brain extraction was done using the BET function in FSL. The anatomic (T1-weighted) scan was registered using an affine transformation to standard Montreal Neurological Institute (MNI) space, and the EPI scans were then registered to each subject's corresponding anatomic scan. Finally, the data were spatially smoothed using a Gaussian kernel with a 5mm full-width half-maximum (FWHM). To fit BTRR in a computationally efficient manner on a full 3D fMRI image response, the EPI scans were downsampled to have voxels with volume $8mm^3$. This analysis with low-resolution 3D fMRI images are used to find areas of increased activity within the entire brain in order to choose slices within the brain that can be analyzed at a higher resolution with voxels of volume $2mm^3$. For both cases, to facilitate parallelization of computational tasks and flexible choice of rank R for different brain structures, data were separated into one of 9 regions of interest based on the MNI structural atlas provided within FSL (Collins et al., 1995; Mazziotta et al., 2001). The MNI structural atlas is a hand-segmented atlas developed by Mazziotta et al. (2001) and Collins et al. (1995) and distributed within the FSL library of neuroimaging tools (Jenkinson et al., 2012).

As regions of interest are not hypercubic in nature, the smallest cuboid containing a region of interest is taken as the response tensor. Parts of the cuboid that are not a part of the region of interest are all assigned the value 0, and masking was performed on the coefficients of all of the models to avoid finding activations outside the brain. There is an established literature on brain

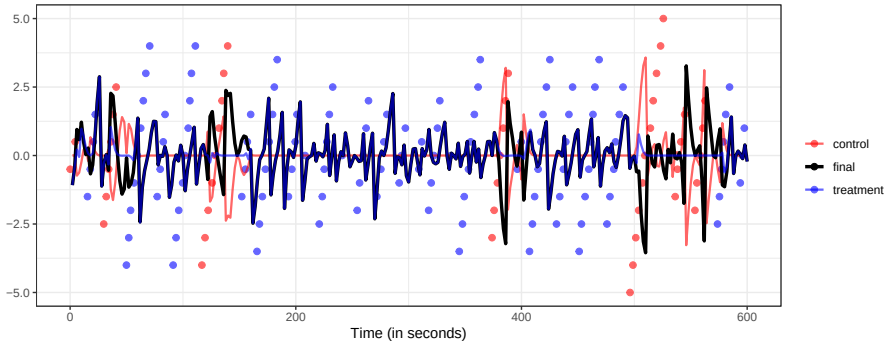

FIG 10. *The raw values of the demeaned number of pumps (points), their convolution with the double-gamma haemodynamic response function (light lines), and the final covariate resulting from their difference (heavy black line) for the subject analyzed.*

imaging that conceptualizes brain images or brain image cross-sections as 3− or 2-dimensional tensors respectively, after similar adjustments (please see Zhou et al. (2013) and Li and Zhang (2017)). Table 1 records the Regions of Interest (ROI) with the number of voxels in both the whole-brain low-resolution analysis and one of the single-slice high-resolution analyses.

The analysis of the whole-brain high-resolution scan for an individual is the ideal goal of these methods. However, due to the size of the data and availability of computer memory, the regions of interest were separated in order to draw from the posterior distribution via Markov Chain Monte Carlo simulations. As an added benefit, the parallelization keeps the computation time reasonable. A follow-up project has recently been published that extends the BTRR model to a tensor response mixed effect model. The proposed model incorporates region-specific random effects and infers jointly on voxel-level activation and region-level connectivity from multi-subject fMRI data, please see Spencer et al. (2020) for details.

To measure the level of risk to a subject at a given time, we modify the procedure used in Schonberg et al. (2012) slightly. First, 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 higher is the risk perceived by the subject. 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 10. Finally, the centered, convolved number of pumps on the control balloon is subtracted from the treatment series to provide a basis for comparison.

It is important to mention that the analysis performed by Schonberg et al. (2012) has some significant differences to the analysis shown here. To begin, their study is done on 16 subjects and across three runs for each subject, with

analyses performed separately for each subject and run. The parameters were then averaged across runs, and an analysis of variance (ANOVA) was performed in order to assess variance in activation within and between subjects. As our proposed model is intended for a single subject, the first run from a randomly-selected subject was used. The predictors for the treatment and control balloons were kept separate in Schonberg et al. (2012), and the activation found from the control balloons was subtracted from the activation found with the treatment balloons. We sought to effectively combine these steps by subtracting the predictors, which produces similar results, as the treatment and control balloons do not overlap temporally within the experiment. Z-score thresholding was used in Schonberg et al. (2012) in combination with a correction by Poline et al. (1997) that combines peak intensity testing corrections with active cluster size corrections. However, these corrections rely on the implicit assumption that the activation parameters follow a lattice approximation to a Gaussian field, which does not assume sparsity and spatial similarity in the same way as the proposed model. In addition, the parameters used to perform the testing corrections are not shared in Schonberg et al. (2012), and thus the results cannot be replicated. Schonberg et al. (2012) also used additional regressors in modeling, including a regressor for the average number of pumps in each balloon trial for both cash-out and explosion outcomes. These were not used in this analysis to focus on an application of the proposed model in a more accessible setting without including additional variables with different variable treatments, including slice timing and other nuisance variables. As a result, this analysis does not account for average BOLD responses under different trial conditions, such as different outcomes (cash-out or explosion) and stimuli (treatment or control balloons). However, inference from BTRR generally coincides with the conclusion presented by Schonberg et al. (2012) and the General Linear Model concerning the bilateral insula activation patterns, even though only one subject is analyzed. We will present them in the upcoming paragraphs.

As discussed earlier, in order to perform inference on high-resolution neuroimaging data, low-resolution inference is performed first on the entire three-dimensional volume of the regions of interest. Once general areas of activation are found in the three-dimensional regions of interest, slices of the subject's scan can be chosen in order to perform high-resolution inference. In order to illustrate this process, we first make 11,000 draws from the posterior distribution, discarding the first 1,000 draws as a burn-in using the low-resolution data. Point estimates for one of the axial slices of the low-resolution data after applying the sequential 2-means variable selection method (Li and Pati, 2017) and merging the regions of interest into a single image can be seen in Figure 11. This slice of the three-dimensional image was chosen for display in the figure because it shows information on an axial slice through the middle of the brain. Table 1 shows the tensor dimensions in each region, the number of ROI voxels within each response tensor, and which rank was chosen for each region using the Deviance Information Criterion (Gelman et al., 2014) for the low-resolution analysis. For reference, a figure showing the locations of each of these regions of interest within the Montreal Neurological Institute Atlas has been included in

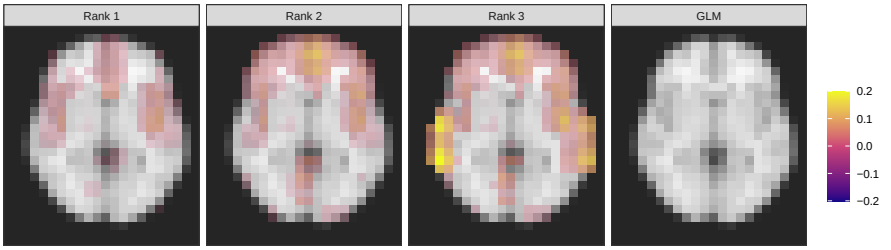

FIG 11. *A single slice of the tensor estimate for the whole brain overlaid on top of the reference image.*

the supplementary material. We then run the analysis following the same procedure on the high-resolution data on three slices that fall within the area shown in the downsampled slice in Figure 11. This high resolution 2D slices correspond to $z = (44, 45, 46)$. The tensor dimensions in each region, the number of ROI voxels within each response tensor, and which rank was chosen for each region using the Deviance Information Criterion (Gelman et al., 2014) for the high-resolution analysis for these slices are shown in Table 1. The point estimates for the high-resolution analysis for the single slice data with tensor ranks 1, 2 and 3 can be seen in Figure 12. The analysis shows no significant difference in terms of activation in these three slices and hence we only provide the final estimate of the coefficients in the high-resolution analysis allowing for different ranks to be used for different regions of interest based on the DIC for slice $z = 45$. This can be seen in Figure 13. For ease of visual comparison, all of the plots within figures 11, 12, and 13 were color-thresholded at the approximate 90th percentile value for nonzero coefficients across the four models being compared. The active coefficients identified in the single-slice analyses were between 0 and 0.37 for the Bayesian models and between -0.4 and 0.73 for the GLM. The active coefficients identified in the whole-brain analyses were between 0 and 0.59 for the Bayesian models and between -0.23 and 0 for the GLM.

The same general linear model maximum likelihood estimate described in section 4 is included for comparison. In the whole brain analysis, the Bayesian estimates for the tensor coefficient are coincident with each other, with higher rank models producing larger coefficient estimates. The BTRR model estimates suggest that there is activation in the subject's frontal and temporal lobes. The general linear model has estimates that are somewhat larger in amplitude in the high-resolution slice analysis than the Bayesian sparse tensor response regression estimates, though usually over smaller spatial regions. This is likely due to the shrinkage prior favoring smaller tensor coefficient values in the BTRR models. However, the final point estimates from the general linear model and the BTRR models show generally coincident voxel-level activations, though the areas of activation are larger in the BTRR inference. In addition, the activation patterns across the adjacent slices within the high-resolution are largely coincident with

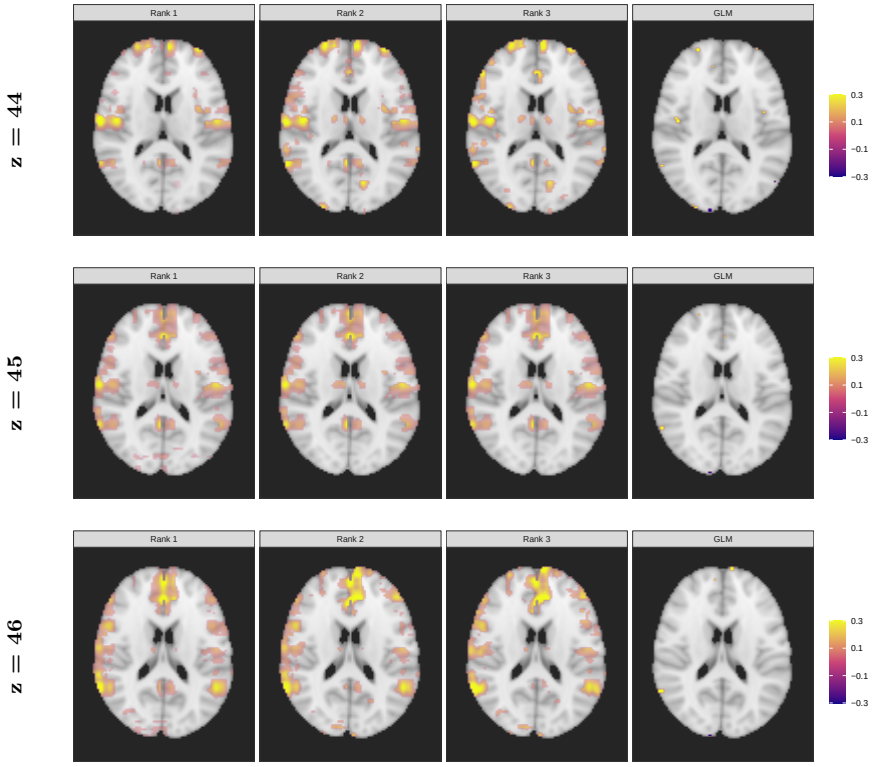

FIG 12. *The estimates for the tensor coefficient in high-resolution single-slice analyses for slices* $z = (44, 45, 46)$ *overlaid on reference images of the brain.*

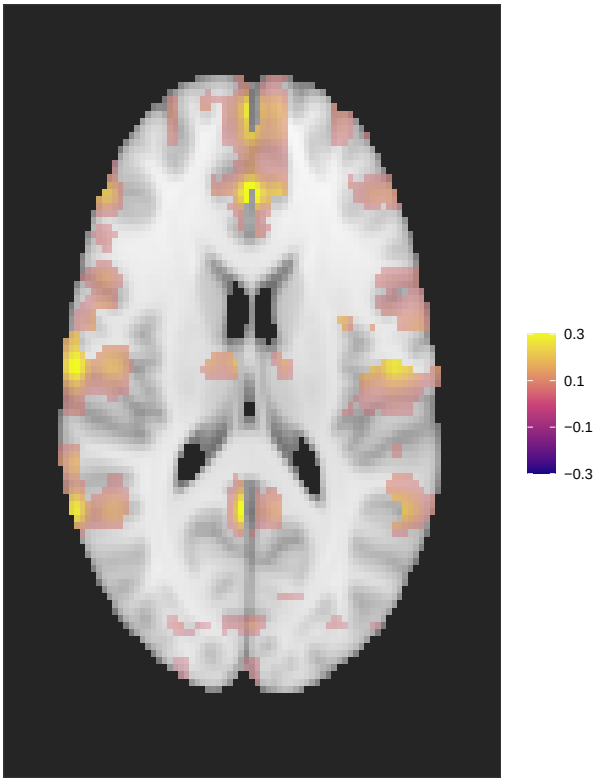

FIG 13. *The final estimate of the effect of increased perceived risk on BOLD measure response in different regions of the brain after selecting R for each region using the DIC.*

each other and the low-resolution while-brain analysis.

6. 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 fitted tensor coefficient to the true tensor coefficient as data size increases.

As an illustrative example, the present article focuses on the analysis of a single-subject fMRI data to detect voxels of the brain which exhibit neuronal activity in response to stimuli. The whole brain is analyzed first using low-

ROI	Rank Selected	p1	p2	p3	# voxels
Low-resolution					
Caudate	Rank 1	7	8	7	126
Cerebellum	Rank 3	15	11	10	691
Frontal Lobe	Rank 1	18	16	16	1667
Insula	Rank 1	13	9	6	148
Occipital Lobe	Rank 1	16	10	10	649
Parietal Lobe	Rank 1	19	12	13	1113
Putamen	Rank 2	9	7	5	103
Temporal Lobe	Rank 1	19	15	11	943
Thalamus	Rank 3	7	6	4	99
High-resolution slice ($z = 45$)					
Caudate	Rank 1	21	13		113
Frontal Lobe	Rank 2	68	42		952
Insula	Rank 3	41	16		83
Occipital Lobe	Rank 1	47	24		606
Parietal Lobe	Rank 1	70	37		761
Temporal Lobe	Rank 1	63	8		86
Thalamus	Rank 3	20	9		92

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.

resolution data from the three-dimensional brain volume, and then three slices of high-resolution data from the brain volume are analyzed to more specifically infer areas of activation. Analysis of simulated 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 as a Bayesian model. This facilitates its usage in presence of images with a fine resolution. The use of MCMC in the analysis does pose a challenge computationally, which necessitates the separation of the regions of interest in order to make the analysis parallelizable. Future work exploring alternative methods, such as using an expectation-maximization (EM) algorithm, may allow for fast analysis of the whole brain.

The core idea of the article 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 to investigate both voxel level activation and ROI level connectivity from multi-subject fMRI data has recently been published (Spencer et al., 2020). It also needs to be emphasized that the proposed tensor regression framework is developed along the principles of high dimensional variable selection/shrinkage prior literature where a new multi-way stick breaking shrinkage prior is developed to shrink unimportant cell coefficients close to zero. Introducing tensor variables in the analysis intuitively make use of the neighborhood information in the cells, though an explicit spatial modeling of elements in the tensor cells has not been considered here. Developing shrinkage priors for high dimensional parameters after accounting for their spatial correlations has been recognized to be an extremely challenging problem, and hence is much less addressed. We recognize the steep challenge involved in extending our approach to account for

spatial correlations of cell parameters and plan to tackle the problem in a future work.

7. Acknowledgement

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

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.

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 7.1. *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.

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

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 $E_{\mathbf{\Gamma}^{(0)}_{(T)}}(\Phi_T) \leq \exp(-T\lambda_0^2\epsilon^2/32)$.

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

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

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

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

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 7.2. *Let* $\mathcal{K}(\theta) = -\log\{\Pi_T(\boldsymbol{\Gamma}_{(T)} : ||\boldsymbol{\Gamma}_{(T)} - \boldsymbol{\Gamma}^{(0)}_{(T)}||_\infty < \theta)\}$ *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*

$$x(x + ||\tilde{\boldsymbol{\gamma}}^{(0)}_{2,k,v_2,(T)}||)\cdots(x + ||\tilde{\boldsymbol{\gamma}}^{(0)}_{D,k,v_D,(T)}||) + ||\tilde{\boldsymbol{\gamma}}^{(0)}_{1,k,v_1,(T)}||x(x + ||\tilde{\boldsymbol{\gamma}}^{(0)}_{2,k,v_2,(T)}||)\cdots(x + ||\tilde{\boldsymbol{\gamma}}^{(0)}_{D,k,v_D,(T)}||)$$
$$+ \cdots + x||\tilde{\boldsymbol{\gamma}}^{(0)}_{2,k,v_2,(T)}||\cdots||\tilde{\boldsymbol{\gamma}}^{(0)}_{D,k,v_D,(T)}|| - \theta = 0, \tag{7.1}$$

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 (7.1) 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 (7.1). By Cauchy's bound on the roots of polynomials, Eq. (7.1) 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 (7.1), 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.

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

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

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

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

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

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

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

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

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

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

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

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

where $Beta(m_1,m_2)$ is the integrating constant for the Beta density with parameters m_1 and m_2. Finally, integrating out τ_k, leads to

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

for some constant C. Hence

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

□

Under assumptions (a)-(f), the R.H.S is $o(T)$. Thus, we present the next theorem whose proof follows immediately from Theorem 7.2.

Theorem 7.3. *For any constant $\theta > 0$, under conditions (a)-(f) of Theorem 3.1, $\mathcal{K}(\theta) = o(T)$.*

Proof of Theorem 3.1

Proof.

$$\Pi_T(\mathcal{A}_T) = \frac{\int_{\mathcal{A}_T} f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\boldsymbol{\Gamma}_{(T)})\pi_T(\boldsymbol{\Gamma}_{(T)})}{\int f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\boldsymbol{\Gamma}_{(T)})\pi_T(\boldsymbol{\Gamma}_{(T)})} = \frac{\int_{\mathcal{A}_T} \frac{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\boldsymbol{\Gamma}_{(T)})}{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\boldsymbol{\Gamma}^{(0)}_{(T)})}\pi_T(\boldsymbol{\Gamma}_{(T)})}{\int \frac{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\boldsymbol{\Gamma}_{(T)})}{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\boldsymbol{\Gamma}^{(0)}_{(T)})}\pi_T(\boldsymbol{\Gamma}_{(T)})}$$
$$= \frac{\mathcal{N}_T}{\mathcal{D}_T} \leq \Phi_T + (1-\Phi_T)\frac{\mathcal{N}_T}{\mathcal{D}_T}, \tag{7.2}$$

where Φ_T is the exponentially consistent sequence of tests given by Lemma 7.1. 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,

$$\Phi_T \to 0 \quad a.s. \tag{7.3}$$

In addition, we have

$$E_{\boldsymbol{\Gamma}^{(0)}_{(T)}}((1-\Phi_T)\mathcal{N}_T) = \int (1-\Phi_T)\int_{\mathcal{A}_T} \frac{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\boldsymbol{\Gamma}_{(T)})}{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\boldsymbol{\Gamma}^{(0)}_{(T)})}\pi_T(\boldsymbol{\Gamma}_{(T)})f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\boldsymbol{\Gamma}^{(0)}_{(T)})$$
$$= \int_{\mathcal{A}_T}\int (1-\Phi_T)f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\boldsymbol{\Gamma}_{(T)})\pi_T(\boldsymbol{\Gamma}_{(T)}) \leq \sup_{\boldsymbol{\Gamma}_{(T)}\in\mathcal{A}_T} E_{\boldsymbol{\Gamma}_{(T)}}(1-\Phi_T) \leq \exp(-T\lambda_0^2\epsilon^2/16).$$

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

$$\exp(T\lambda_0^2\epsilon^2/32)(1-\Phi_T)\mathcal{N}_T \to 0 \quad a.s.. \tag{7.4}$$

Note that $\mathcal{D}_T = \int \frac{f(\boldsymbol{Y}_1..,\boldsymbol{Y}_T|\boldsymbol{\Gamma}_{(T)})}{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\boldsymbol{\Gamma}^{(0)}_{(T)})}\pi_T(\boldsymbol{\Gamma}_{(T)})$. Let $\tilde{b} = \lambda_0^2\epsilon^2/32$. Consider the set $\mathcal{H}_T = \left\{\boldsymbol{\Gamma}_{(T)} : \frac{1}{T}\log\left[\frac{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\boldsymbol{\Gamma}^{(0)}_{(T)})}{f(\boldsymbol{Y}_1,..,\boldsymbol{Y}_T|\boldsymbol{\Gamma}_{(T)})}\right] < \upsilon\right\}$, for $\upsilon = \tilde{b}/2$.

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

In view of (7.2), (7.3) and (7.4), it is enough to show that $-\log(\Pi_T(\mathcal{H}_T)) \leq T\tilde{b}/8$.

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

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

where the third line follows from assumption (e) of Theorem 3.1 and last inequality is immediate by applying Theorem 7.3. □

References

Armagan, A., Dunson, D. B., and Lee, J. (2013a). "Generalized double Pareto shrinkage." *Statistica Sinica*, 23(1): 119. 7, 8

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

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal statistical society: series B (Methodological)*, 57(1): 289–300. 12, 16

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). "Fast sampling with Gaussian scale mixture priors in high-dimensional regression." *Biometrika*, asw042. 14

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

Bullmore, E., Fadili, J., Maxim, V., Şendur, L., Whitcher, B., Suckling, J., Brammer, M., and Breakspear, M. (2004). "Wavelets and functional magnetic resonance imaging of the human brain." *Neuroimage*, 23: S234–S249. 6

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, asq017. 7

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. 11

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. 7

Collins, D. L., Holmes, C. J., Peters, T. M., and Evans, A. C. (1995). "Automatic 3-D model-based neuroanatomical segmentation." *Human brain mapping*, 3(3): 190–208. 21

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

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. 3

Dunson, D. B. and Xing, C. (2009). "Nonparametric Bayes modeling of multivariate categorical data." *Journal of the American Statistical Association*, 104(487): 1042–1051. 4

Fadili, M. and Bullmore, E. (2002). "Wavelet-generalized least squares: a new BLU estimator of linear regression models with 1/f errors." *NeuroImage*, 15(1): 217–232. 6

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. 8, 16, 23, 24

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.

7

Gerard, D. and Hoff, P. (2015). "Adaptive Higher-order Spectral Estimators." *arXiv preprint arXiv:1505.02114*. 2, 4

Gössl, C., Auer, D. P., and Fahrmeir, L. (2001). "Bayesian spatiotemporal inference in functional magnetic resonance imaging." *Biometrics*, 57(2): 554–562. 12

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. 4, 7

Hans, C. (2009). "Bayesian lasso regression." *Biometrika*, 96(4): 835–845. 7

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96(453): 161–173. 8

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). "Fsl." *Neuroimage*, 62(2): 782–790. 21

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, H. and Pati, D. (2017). "Variable selection using shrinkage priors." *Computational Statistics & Data Analysis*, 107: 107–119. 13, 16, 23

Li, L. and Zhang, X. (2017). "Parsimonious tensor response regression." *Journal of the American Statistical Association*, 112(519): 1131–1146. 3, 5, 22

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, 3

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

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., Woods, R., Paus, T., Simpson, G., Pike, B., et al. (2001). "A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM)." *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1412): 1293–1322. 21

Meyer, F. G. (2003). "Wavelet-based estimation of a semiparametric generalized linear model of fMRI time-series." *IEEE transactions on medical imaging*, 22(3): 315–322. 6

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

Poline, J.-B., Worsley, K. J., Evans, A. C., and Friston, K. J. (1997). "Combining spatial extent and peak intensity to test for activations in functional imaging." *Neuroimage*, 5(2): 83–96. 2, 23

Polson, N. G. and Scott, J. G. (2010). "Shrink globally, act locally: Sparse Bayesian regularization and prediction." *Bayesian Statistics*, 9: 501–538. 7

Quirós, A., Diez, R. M., and Gamerman, D. (2010). "Bayesian spatiotemporal model of fMRI data." *NeuroImage*, 49(1): 442–456. 12

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. 19, 21, 22, 23

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. 21

Song, Q. and Liang, F. (2017). "Nearly optimal Bayesian shrinkage for high dimensional regression." *arXiv preprint arXiv:1712.08964*. 4, 11, 29

Spencer, D., Guhaniyogi, R., and Prado, R. (2020). "Joint Bayesian Estimation of Voxel Activation and Inter-regional Connectivity in fMRI Experiments." *Psychometrika*, 1–25. 22, 27

Stefano, S., Quartagno, M., Tamburini, M., and Robinson, D. (2018). *orcutt: Estimate Procedure in Case of First Order Autocorrelation*.
URL https://CRAN.R-project.org/package=orcutt 12

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

Teh, Y. W., Grür, D., and Ghahramani, Z. (2007). "Stick-breaking construction for the Indian buffet process." In *Artificial Intelligence and Statistics*, 556–563. 8

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. 9

— (2011). "Information rates of nonparametric Gaussian process methods." *Journal of Machine Learning Research*, 12(Jun): 2095–2119. 9

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. 12

Welvaert, M. and Rosseel, Y. (2013). "On the definition of signal-to-noise ratio and contrast-to-noise ratio for fMRI data." *PloS one*, 8(11): e77089. 13

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. 3, 6, 12, 13

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. 3

Zhou, H. and Li, L. (2014). "Regularized matrix regression." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2): 463–483. 4

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. 4, 22

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. 3