Journal of Machine Learning Research 18 (2017) 1-31 Submitted 7/16; Published 8/17

Bayesian Tensor Regression

Rajarshi Guhaniyogi[*] RGUHANIY@UCSC.EDU
Department of Applied Mathematics & Statistics
University of California
Santa Cruz, CA 95064, USA

Shaan Qamar[*] SIQAMAR@GMAIL.COM
Google Inc.
Mountain View, CA 94043, USA

David B. Dunson DUNSON@DUKE.EDU
Department of Statistical Science
Duke University
Durham, NC 27708-0251, USA

Editor: Robert McCulloch

Abstract

We propose a Bayesian approach to regression with a scalar response on vector and tensor covariates. Vectorization of the tensor prior to analysis fails to exploit the structure, often leading to poor estimation and predictive performance. We introduce a novel class of multiway shrinkage priors for tensor coefficients in the regression setting and present posterior consistency results under mild conditions. A computationally efficient Markov chain Monte Carlo algorithm is developed for posterior computation. Simulation studies illustrate substantial gains over existing tensor regression methods in terms of estimation and parameter inference. Our approach is further illustrated in a neuroimaging application.

Keywords: Multiway Shrinkage Prior, Magnetic Resonance Imaging (MRI), Parafac Decomposition, Posterior Consistency, Tensor Regression

1. Introduction

In many application areas, it is common to collect predictors that are structured as a multiway array or tensor. For example, the elements of this tensor may correspond to voxels in a brain image (Lindquist, 2008; Lazar, 2008; Hinrichs et al., 2009; Ryali et al., 2010). Existing approaches for quantifying associations between an outcome and such tensor predictors mostly fall within two groups. The first approach assesses the association between each cell (for brain images referred to as voxel) and the response independently, providing a p-value 'map' (Lazar, 2008). The p-values can be adjusted for multiple comparisons to identify 'significant' sub-regions of the tensor. Although this approach is widely used and appealing in its simplicity, clearly such independent screening approaches have key disadvantages relative to methods that take into account the joint impact of the overall

. [*]**These authors contributed equally** tensor simultaneously. Unfortunately, the literature on simultaneous analysis approaches is sparse.

One naive approach is to simply vectorize the tensor and then use existing methods for high-dimensional regression. Such vectorization fails to preserve spatial structure, making it more difficult to learn low-dimensional relationships with the response. Efficient learning is of critical importance, as the sample size is typically massively smaller than the total number of cells. Alternative approaches within the regression framework include functional regression and two stage approaches. The former views the tensor as a discretization of a continuous functional predictor. Most of the literature on functional predictors focuses on 1D functions; Reiss and Ogden (2010) consider the 2D case, but substantial challenges arise in extensions to 3D due to dimensionality and collinearity among cells. Recently Wang et al. (2014) considered 3D regularized functional regression with Haar wavelet basis. The article is essentially frequentist in nature with simulation studies showing only the mean squared error and the percentage of correctly identified zero and nonzero elements. Additionally, the article reveals that functional regression is largely affected by the choice of proper basis functions. The second set of approaches, i.e. Two stage approaches first conduct a dimension reduction step, commonly using PCA, and then fit a model using lower dimensional predictors (Caffo et al., 2010). A clear disadvantage of such approaches is that the main principal components driving variability in the random tensor may have relatively limited impact on the response variable. Potentially, supervised PCA could be used, but it is not clear how to implement such an approach in 3D or higher dimensions.

Zhou et al. (2013) propose extending generalized linear regression to include a tensor structured parameter corresponding to the measured tensor predictor. To circumvent difficulties with extensions to higher order tensor predictors, they impose additional structure on the tensor parameter, supposing it decomposes as a rank-R parafac sum (see Section 2.1). This massively reduces the effective number of parameters to be estimated. They develop a penalized likelihood approach where adaptive lasso penalties are be imposed on individual margins of the parafac decomposition, focusing on good point estimation for the tensor parameter. However, their method relies heavily on cross-validation for selecting tuning parameters which are sensitive to the tensor dimension, the signal-to-noise ratio (degree of sparsity) and the parafac rank. Given that there is no automatic selection procedure for the tuning parameters provided in Zhou et al. (2013), they have to be fed manually by the end user which is problematic for an unknown tensor regression problem.

Of practical interest is a “self calibrating” procedure which adapts the complexity of the model to the data. We propose a principled method to effectively shrink unimportant cell coefficients to zero while maintaining accuracy in estimating important cell coefficients. Our framework gives rise to the automatic selection of tuning parameters, with carefully constructed shrinkage priors that naturally induce sparsity within and across components in the tensor factorization of the tensor coefficient for optimal region selection. In addition, the need for valid measures of uncertainty on parameter (predictive) estimates is crucial, especially in settings with low or moderate sample sizes, which naturally motivates our Bayesian approach. Recently Suzuki (2015) proposed a Bayesian tensor regression approach with naive Gaussian prior on the components of the tensor factorization of the tensor coefficients. In contrast, our proposed prior on the tensor coefficient is more sophisticated in the sense that it imparts shrinkage in three ways: at a global level, at a local level of individual parameters, and also provides shrinkage towards low rank decomposition of the tensor coefficient. Similarly, Bayesian tensor regression framework proposed in Goldsmith et al. (2014) uses binary indicators to determine whether a cell in the tensor predictor is predictive of the response. For a tensor predictor with $30 \times 30 \times 30$ cells, such an approach requires to update 27000 binary indicators in each MCMC iteration and is deemed unsatisfactory due to mixing issues and poor inference.

Our approach differs from image reconstruction literature as we do not model the distribution of the tensor $\boldsymbol{X}$ (Qiu, 2007). There is a considerable recent literature on frequentist tensor modeling in which one typically encounters time series (generally to study social networks or images evolving over time) with response at every time point is an array/tensor (Gerard and Hoff, 2015; Hoff et al., 2015). There is also a Bayesian literature that facilitates joint modeling of a large number of unordered categorical variables (Zhou et al., 2015). Our framework is fundamentally different from these approaches in the sense that *these are all unsupervised tensor modeling approach while we propose a framework for supervised tensor regression.* To the best of our knowledge, we are the first to propose a novel multiway shrinkage prior in Bayesian tensor regression framework (with scalar response on a tensor predictor) that accommodates shrinkage of the tensor coefficient for the appropriate identification of important cells in the tensor predictor. Besides, we offer strong posterior consistency results on Bayesian tensor regression framework with multiway shrinkage prior.

Remainder of the manuscript evolves as follows. In Section 2, we propose the basic framework of the tensor regression model with a scalar response, vector predictors and a tensor predictor. Section 3 characterizes desirable criteria for a multiway shrinkage prior and proposes a novel multiway shrinkage prior on the tensor coefficient. Sections 4 and 5 provide theoretical results on the convergence of posterior distribution under the mutiway shrinkage prior and details on posterior computation respectively. Various simulation studies with 2D and 3D tensor predictors are presented in Sections 6 and 7 respectively to study effectiveness of the Bayesian tensor regression under various degrees of sparsity and signal strength. Section 8 is devoted to a real brain connectome data analysis using the proposed Bayesian tensor regression model along with its competitors. The manuscripts ends with a discussion.

2. Tensor Regression

This section provides details on the tensor regression model.

2.1 Basic Notation

Let $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1p_1})'$ and $\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{2p_2})'$ be vectors of length p_1 and p_2, respectively. The vector outer product $\boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_2$ is a $p_1 \times p_2$ matrix with (i, j)-th entry $\beta_{1i}\beta_{2j}$. A D-way outer product between vectors $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp_j})$, $1 \le j \le D$, is a $p_1 \times \cdots \times p_D$ multi-dimensional array denoted $\boldsymbol{B} = \boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_2 \circ \cdots \circ \boldsymbol{\beta}_D$ with entries $(\boldsymbol{B})_{i_1,\ldots,i_D} = \prod_{j=1}^{D} \beta_{ji_j}$. Define a vec$(\boldsymbol{B})$ operator as stacking elements of this D-way tensor into a column vector of length $\prod_{j=1}^{D} p_j$. From the definition of outer products, it is easy to see that vec$(\boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_2 \circ \cdots \circ \boldsymbol{\beta}_D) = \boldsymbol{\beta}_D \otimes \cdots \otimes \boldsymbol{\beta}_1$. As a higher order generalization of matrix singular value decomposition, Tucker decomposition of a D-way tensor $\boldsymbol{B} \in \otimes_{j=1}^{D} \Re^{p_j}$ is often considered. The Tucker decomposition (Kolda and Bader, 2009) can be expressed as

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

where $\boldsymbol{\beta}_j^{(r_j)}$ is a p_j dimensional vector, $1 \leq j \leq D$, and $\boldsymbol{\Lambda} = (\lambda_{r_1,\ldots,r_D})_{r_1,\ldots,r_D=1}^{R_1,\ldots,R_D}$ is referred to as the *core tensor*. If one considers $\{\boldsymbol{\beta}_j^{(r_j)}; 1 \leq r_j \leq R_j, 1 \leq j \leq D\}$ as "factor loadings" and $\lambda_{r_1,\ldots,r_D}$ to be the corresponding coefficients, then the Tucker decomposition may be thought of as a multiway analogue to factor modeling.

A rank-R parafac decomposition emerges as a special case of Tucker decomposition 1 when $R_1 = R_2 = \cdots = R_D = R$ and $\lambda_{r_1,\ldots,r_D} = I(r_1 = r_2 = \cdots = r_D)$. In particular, $\boldsymbol{B} \in \otimes_{j=1}^D \Re^{p_j}$ assumes a rank-R parafac decomposition if

$$\boldsymbol{B} = \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} \circ \cdots \circ \boldsymbol{\beta}_D^{(r)} \tag{2}$$

where $\boldsymbol{\beta}_j^{(r)}$, $1 \leq j \leq D$ and $1 \leq r \leq R$ are the p_j dimensional 'margins'. The parafac decomposition is more widely used due to its relative simplicity.

2.2 Model Framework

Let $y \in \mathcal{Y}$ denotes a response variable, with $\boldsymbol{z} \in \mathcal{X} \subset \Re^p$ and $\boldsymbol{X} \in \otimes_{j=1}^D \Re^{p_j}$ scalar and tensor predictors, respectively. We consider a tensor regression model having a general form

$$y \sim f\left(\alpha + \boldsymbol{z}'\boldsymbol{\gamma} + \langle \boldsymbol{X}, \boldsymbol{B}\rangle, \sigma\right), \quad \langle \boldsymbol{X}, \boldsymbol{B}\rangle = \text{vec}(\boldsymbol{X})'\text{vec}(\boldsymbol{B}), \tag{3}$$

where $f(\mu, \sigma)$ is a family of distributions having location μ and scale σ, $\boldsymbol{\gamma}$ is a $p \times 1$ coefficient for scalar preditors and $\boldsymbol{B} \in \otimes_{j=1}^D \Re^{p_j}$ is the tensor parameter corresponding to measured tensor predictor $\boldsymbol{X}$. We focus more specifically on the Gaussian linear model case with

$$y = \alpha + \boldsymbol{z}'\boldsymbol{\gamma} + \langle \boldsymbol{X}, \boldsymbol{B}\rangle + \epsilon,\ \epsilon \sim N(0, \sigma^2). \tag{4}$$

The coefficient tensor $\boldsymbol{B}$ has $\prod_{j=1}^D p_j$ elements, necessitating substantial dimensionality reduction. A rank-1 parafac decomposition assumes $\boldsymbol{B} = \boldsymbol{\beta}_1 \circ \cdots \circ \boldsymbol{\beta}_D$ and $\text{vec}(\boldsymbol{B}) = \boldsymbol{\beta}_D \otimes \cdots \otimes \boldsymbol{\beta}_1$. This reduces to modeling $y = \alpha + \boldsymbol{z}'\boldsymbol{\gamma} + \boldsymbol{\beta}_1'\boldsymbol{X}\boldsymbol{\beta}_2$ when $D = 2$, corresponding to the bilinear model considered in Hung and Wang (2013). Since only the single parameter vector $\boldsymbol{\beta}_j$ captures signal along the jth dimension, a rank-1 assumption severely limits flexibility, ruling out interactions among dimensions. Following Zhou et al. (2013), we use a more flexible rank-R parafac decomposition for $\boldsymbol{B} = \sum_{r=1}^R \boldsymbol{\beta}_1^{(r)} \circ \cdots \circ \boldsymbol{\beta}_D^{(r)}$ introduced in (2) with $\boldsymbol{\beta}_j^{(r)} \in \Re^{p_j}$, $1 \leq j \leq D$, and $1 \leq r \leq R$. Expression (4) then becomes

$$\begin{aligned} y &= \alpha + \boldsymbol{z}'\boldsymbol{\gamma} + \left\langle \boldsymbol{X}, \sum_{r=1}^R \boldsymbol{\beta}_1^{(r)} \circ \cdots \circ \boldsymbol{\beta}_D^{(r)} \right\rangle + \epsilon \\ &= \alpha + \boldsymbol{z}'\boldsymbol{\gamma} + \sum_{(i_1,\ldots,i_D)} (\boldsymbol{X})_{i_1,\ldots,i_D}(\boldsymbol{B})_{i_1,\ldots,i_D} + \epsilon \end{aligned} \tag{5}$$

where voxel $(\boldsymbol{X})_{i_1,\ldots,i_D}$ of the tensor predictor has corresponding parameter

$$(\boldsymbol{B})_{i_1,\ldots,i_D} = \sum_{r=1}^{R} \prod_{j=1}^{D} \beta_{j,i_j}^{(r)}, \quad (i_1,\ldots,i_D) \in \mathcal{V}_{\boldsymbol{B}} = \otimes_{j=1}^{D}\{1,\ldots,p_j\}. \tag{6}$$

The model is therefore nonlinear in the parameters defining $\boldsymbol{B}$. A hierarchical specification is completed by placing priors over unknown model parameters. While placing priors over α and $\boldsymbol{\gamma}$ is straightforward, Section 3.2 focuses on specification of the prior over tensor parameters which is nontrivial and one of the main contributions of this work.

Under the assumed rank-R parafac decomposition for $\boldsymbol{B}$, model (5) requires estimating $p+2+R\sum_{j=1}^{D} p_j$ as opposed to $p+2+\prod_{j=1}^{D} p_j$ parameters for the unstructured vectorized (saturated) model. As we are interested in identifying geometric sub-regions of the tensor across which coefficients are not close to zero, with the remaining elements being very close to zero, one wonders whether such dramatic dimension reduction retains sufficient flexibility. Finally, we would like to accurately estimate coefficient values in these sub-regions. Consistent with our theoretical analysis in Section 4, extensive simulation studies in Section 6 confirm our ability to accomplish these goals.

2.3 Model Identifiability

From model (5) it is clear that only voxel-level coefficients are identified and not the individual tensor margins defining their product-sum given in (6). In the tensor setting, identifiability restrictions are understood in light of the following indeterminacies:

1. *Scale indeterminacy*: for each $r = 1,\ldots,R$, define $\boldsymbol{\lambda}_r = (\lambda_{1r},\ldots,\lambda_{Dr})$ such that $\prod_{j=1}^{D}\lambda_{jr} = 1$. Then replacing $\boldsymbol{\beta}_j^{(r)}$ by $\lambda_{jr}\boldsymbol{\beta}_j^{(r)}$ leaves the tensor parameter $\boldsymbol{B}$ unaltered.

2. *Permutation indeterminacy*: $\sum_{r=1}^{R} \circ_{j=1}^{D}\boldsymbol{\beta}_j^{(r)} = \sum_{r=1}^{R} \circ_{j=1}^{D}\boldsymbol{\beta}_j^{(P(r))}$ for any permutation $P(\cdot)$ of $\{1,2,\ldots,R\}$. In particular, this implies that $\circ_{j=1}^{D}\boldsymbol{\beta}_j^{(r)}$ are not identifiable for $r = 1,\ldots,R$.

3. *Orthogonal transformation indeterminacy* ($D = 2$ only): for any orthonormal matrix $\boldsymbol{O}$, one has $(\boldsymbol{\beta}_1^{(r)}\boldsymbol{O}) \circ (\boldsymbol{\beta}_2^{(r)}\boldsymbol{O}) = \boldsymbol{\beta}_1^{(r)} \otimes \boldsymbol{\beta}_2^{(r)}$.

For $D > 2$, imposing the following $(D-1)R$ constraints ensures identifiability of the margin parameters comprising the rank-R parafac decomposition:

$$\beta_{j,1}^{(r)} = 1,\ 1 \le j < D,\ 1 \le r \le R, \qquad \text{and} \qquad \beta_{D,1}^{(1)} > \cdots > \beta_{D,1}^{(R)}. \tag{7}$$

In the context of Bayesian tensor regression, our emphasis is on accurate estimation and inference on tensor parameter $\boldsymbol{B}$, and on state-of-the-art predictive performance. Importantly, $\boldsymbol{B}$ is always identifiable even if the tensor margins $\boldsymbol{\beta}_j^{(r)}$'s are not, hence we avoid imposing identifiability restrictions on the latter. As is evident from our simulation studies, non-identifiability of $\boldsymbol{\beta}_j^{(r)}$ does not appear to cause convergence issues and in-fact, simplifies the design of efficient computational algorithms.

3. Multiway Shrinkage Priors

This Section outlines the novel multiway shrinkage prior on the tensor coefficient.

3.1 Vector Shrinkage Priors

There has been recent interest in high-dimensional regression with vector predictors, choosing priors which shrink small coefficients towards zero while minimizing shrinkage of large coefficients. Many of these priors can be expressed as a global-local (GL) scale mixtures (Polson and Scott, 2012) with

$$\theta_j \sim \mathrm{N}(0, \psi_j \tau), \quad \psi_j \sim g, \quad \tau \sim h, \tag{8}$$

where $(\theta_1, \ldots, \theta_p)$ is a coefficient vector, τ is a global scale and ψ_j is a local-scale. When g is a mixture of two components, with one concentrated near zero and the other away from zero, a spike and slab prior is obtained. Many other choices of g and h have been considered. Although the GL family is widely used and versatile, Bhattacharya et al. (2015) note advantages in drawing the local scales jointly. In particular, they propose to let

$$\theta_j \sim \mathrm{DE}(\cdot|\phi_j \tau), \quad (\phi_1, ..., \phi_p) \sim \mathrm{Dirichlet}(a, \ldots, a), \quad \tau \sim h.$$

where $\mathrm{DE}(\cdot)$ denote the double-exponential distribution. For small a and large p, the $\mathrm{Dirichlet}(a, \ldots, a)$ prior has the property of favoring many values close to zero with a few much larger values, but with $\sum_j \phi_j = 1$. Though we draw motivation from literature on vector shrinkage priors, our goal of proposing a shrinkage prior on tensor parameter $\boldsymbol{B}$ is fundamentally more challenging as discussed in forthcoming sections.

3.2 Multiway Priors

We propose a new class of multiway shrinkage priors in the generalized linear model setting with tensor valued predictors. Assuming tensor parameter $\boldsymbol{B}$ admits a rank-R parafac decomposition, model (5) results in cell-level coefficients that are a nonlinear function of the corresponding tensor margin parameters (see (6)). Moreover, this implies simultaneous shrinkage on each of the $\prod_{j=1}^{D} p_j$ cell coefficients as imposed by the prior over $R \sum_{j=1}^{D} p_j$ parameters. This necessitates careful prior specification on the tensor margins $\{\boldsymbol{\beta}_j^{(r)}; 1 \leq j \leq D, 1 \leq r \leq R\}$ such that the induced cell-level prior has adequate tails so as to prevent over shrinkage.

There are a number of desirable characteristics for a multiway prior on the tensor margins. The proposed multiway shrinkage prior must have a structure that facilitates efficient and reliable model fitting. In addition, it is important to ensure that

1. For each $r = 1, \ldots, R$, $\left(\beta_{1,i_1}^{(r)}, \ldots, \beta_{D,i_D}^{(r)}\right)$ and $\left(\beta_{1,k_1}^{(r)}, \ldots, \beta_{D,k_D}^{(r)}\right)$ are equal in distribution, for any $(i_1, \ldots, i_D), (k_1, \ldots, k_D) \in \mathcal{V}_{\boldsymbol{B}} \times \mathcal{V}_{\boldsymbol{B}}$ and $(i_1, \ldots, i_D) \neq (k_1, \ldots, k_D)$. This is to ensure that $(B)_{i_1, \ldots, i_D}$ and $(B)_{k_1, \ldots, k_D}$ have the same distribution apriori.

2. Shrinkage towards a low rank decomposition, with the model adapting to the complexity and signal in the data, effectively deleting unnecessary dimensions.

3. The prior should favor recovery of contiguous geometric subregions of the tensor across which the cell observations are predictive of the response.

3.3 The Multiway Dirichlet GDP Prior

There are many ways of specifying priors over tensor margins $\boldsymbol{\beta}_j^{(r)}$ to satisfy the criteria listed. We propose a particular choice called the *multiway Dirichlet generalized double Pareto* (M-DGDP) prior. This prior induces shrinkage across components in an exchangeable way, with global scale $\tau \sim \mathrm{Ga}(a_\tau, b_\tau)$ adjusted in each component as $\tau_r = \phi_r \tau$ for $r = 1, \ldots, R$, where $\Phi = (\phi_1, \ldots, \phi_R) \sim \mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_R)$ encourages shrinkage towards lower ranks in the assumed parafac decomposition. In addition, $\boldsymbol{W}_{jr} = \mathrm{diag}(w_{jr,1}, \ldots, w_{jr,p_j})$, $j = 1, \ldots, D$ and $r = 1, \ldots, R$ are margin-specific scale parameters for each component. The hierarchical margin-level prior is given by

$$\boldsymbol{\beta}_j^{(r)} \sim \mathrm{N}\big(0, (\phi_r \tau)\boldsymbol{W}_{jr}\big), \quad w_{jr,k} \sim \mathrm{Exp}(\lambda_{jr}^2/2), \quad \lambda_{jr} \sim \mathrm{Ga}(a_\lambda, b_\lambda). \tag{9}$$

Collapsing over element-specific scales, notice that $\beta_{j,k}^{(r)} | \lambda_{jr}, \phi_r, \tau \overset{\text{iid}}{\sim} \mathrm{DE}(\lambda_{jr}/\sqrt{\phi_r \tau})$, $1 \le k \le p_j$. Prior (9) induces a GDP prior on the individual margin coefficients which in turn has the form of an adaptive Lasso penalty (Armagan et al., 2013a). Flexibility in estimating $\boldsymbol{B}_r = \{\boldsymbol{\beta}_j^{(r)}; 1 \le j \le D\}$ is accommodated by modeling within-margin heterogeneity via element-specific scaling $w_{jr,k}$. Common rate parameter λ_{jr} shares information between margin elements, encouraging shrinkage at the local scale.

The framework adpoted by Zhou et al. (2013) for Frequentist tensor regression starts by assuming the true parafac rank and relying on shrinkage through a global parameter for estimating the tensor coefficient. In contrast, M-DGDP prior proposes joint shrinkage on the global and local component parameters to achieve improved inference and estimation. Though rank-selection is not our purview, our prior also accomodates dimension reduction by favoring low-rank factorizations as discussed below.

4. Posterior Consistency for Tensor Regression

This Section details out theoretical properties of the tensor regression framework.

4.1 Notation and Framework

We establish convergence results for tensor regression model (5) under the simplifying assumptions that the intercept is omitted by centering the response and the error variance is $\sigma^2 = 1$. Since our main focus is on the tensor coefficient, we assume coefficients for ordinary scalar covariates to be known. Without loss of generality, we assume $\boldsymbol{\gamma} = (0, \ldots, 0)$. We consider an asymptotic setting in which the dimensions of the tensor grow with n. This paradigm attempts to capture the fact that tensor dimension $\prod_j p_{j,n}$ is typically substantially larger than sample size. This creates theoretical challenges, related to (but distinct from) those faced in showing posterior consistency for high dimensional regression (Armagan et al., 2013b) and multiway contingency tables (Zhou et al., 2015).

Suppose the data generating model comes from the same class of models where the fitted model belongs to, i.e., having true tensor parameter $\boldsymbol{B}_n^0 \in \otimes_{j=1}^D \Re^{p_{j,n}}$, error variance $\sigma_0^2 = 1$. We also assume that the true tensor coefficient $\boldsymbol{B}_n^0$ generating the data admits a rank-R PARAFAC decomposition as below

$$\boldsymbol{B}_n^0 = \sum_{r=1}^R \boldsymbol{\beta}_{1,n}^{0(r)} \circ \cdots \circ \boldsymbol{\beta}_{D,n}^{0(r)}, \qquad \boldsymbol{\beta}_{j,n}^{0(r)} = (\beta_{j,n,1}^{0(r)}, \ldots, \beta_{j,n,p_{j,n}}^{0(r)})' \in \Re^{p_{j,n}}.$$

In addition, define $\boldsymbol{F}_n, \boldsymbol{F}_n^0 \in \Re^{R\sum_{j=1}^D p_{j,n}}$ as the vectorized parameters:

$$\begin{aligned} \boldsymbol{F}_n &= \mathrm{vec}\big(\boldsymbol{\beta}_{1,n}^{(1)}, \cdots, \boldsymbol{\beta}_{1,n}^{(R)}, \cdots, \boldsymbol{\beta}_{D,n}^{(1)}, \cdots, \boldsymbol{\beta}_{D,n}^{(R)}\big) \\ \boldsymbol{F}_n^0 &= \mathrm{vec}\big(\boldsymbol{\beta}_{1,n}^{0(1)}, \cdots, \boldsymbol{\beta}_{1,n}^{0(R)}, \cdots, \boldsymbol{\beta}_{D,n}^{0(1)}, \cdots, \boldsymbol{\beta}_{D,n}^{0(R)}\big). \end{aligned}$$

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

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

Denote $\mathrm{KL}(f(y_i|\boldsymbol{B}_n^0), f(y_i|\boldsymbol{B}_n))$ as KL_i. The KL-distance between $\mathrm{N}(\mu_1, \sigma_1^2)$ and $\mathrm{N}(\mu_2, \sigma_2^2)$ is $\log(\sigma_2/\sigma_1) + \big(\sigma_1^2 + (\mu_1 - \mu_2)^2\big)/2\sigma_2^2 - \frac{1}{2}$, so it follows $\mathrm{KL}_i = \mathrm{KL}\left(\mathrm{N}\big(\langle \boldsymbol{X}_i, \boldsymbol{B}_n\rangle, 1\big), \mathrm{N}\big(\langle \boldsymbol{X}_i, \boldsymbol{B}_n^0\rangle, 1\big)\right)$ $= \frac{1}{2}\left(\langle \boldsymbol{X}_i, \boldsymbol{B}_n^0\rangle - \langle \boldsymbol{X}_i, \boldsymbol{B}_n\rangle\right)^2$. Hence, a KL neighborhood of radius ϵ around $\boldsymbol{B}_n^0$ can be re-expressed as $\mathcal{B}_n = \left\{\boldsymbol{B}_n : \frac{1}{2n}\sum_{i=1}^n \left(\langle \boldsymbol{X}_i, \boldsymbol{B}_n^0\rangle - \langle \boldsymbol{X}_i, \boldsymbol{B}_n\rangle\right)^2 < \epsilon\right\}$. Further, let π_n and Π_n denote prior and posterior densities with n observations, respectively, and

$$\Pi_n(\mathcal{B}_n^c) = \frac{\int_{\mathcal{B}_n^c} f(\boldsymbol{y}_n|\boldsymbol{B}_n)\pi_n(\boldsymbol{F}_n)}{\int f(\boldsymbol{y}_n|\boldsymbol{B}_n)\pi_n(\boldsymbol{F}_n)},$$

with $\boldsymbol{y}_n = (y_1, \ldots, y_n)'$ and $f(\boldsymbol{y}_n|\boldsymbol{B}_n)$ is the density of $\boldsymbol{y}_n$ under model (5). Posterior consistency is established by showing that

$$\Pi_n\left(\mathcal{B}_n^c\right) \to 0 \quad \text{under } \boldsymbol{B}_n^0 \quad \text{a.s. as } n \to \infty. \tag{10}$$

4.2 Main Result

Our main theorem is that (10) holds under a simple sufficient condition on the prior and the tensor predictors.

Theorem 1 *Let $\zeta_n = n^{\frac{1+\rho_3}{2}}$ ($\rho_3 > 0$), $M_n = \frac{1}{n}\sqrt{\sum_{i=1}^n ||\boldsymbol{X}_i||_2^2}$. Given Lemma 6 in Appendix A, for any $\epsilon > 0$, $\Pi_n(\boldsymbol{B}_n : \frac{1}{n}\sum_{i=1}^n KL_i > \epsilon) \to 0$ a.s. under $\boldsymbol{B}_n^0$, for the prior $\pi_n(\boldsymbol{B}_n)$ that satisfies*

$$\pi_n\left(\boldsymbol{B}_n : ||\boldsymbol{B}_n - \boldsymbol{B}_n^0||_2 < \frac{2\eta}{3M_n\zeta_n}\right) > \exp(-dn), \ \textit{for all large}\, n \tag{11}$$

for any $d > 0$ and $\eta < \frac{\epsilon}{32} - d$. That is, the model is posterior consistent when (11) holds.

Lemma 6 in Appendix A verifies the existence of exponentially-consistent tests. The proof of the Lemma and Theorem are provided in the Appendix A. The proposed multiway shrinkage prior satisfies (11) and hence leads to posterior consistency under the following Theorem.

Theorem 2 *For fixed constants H_1, H_2, M_1, ρ_1 and $\rho_2 > 0$, the M-DGDP prior* (9) *on $\boldsymbol{B}_n$ satisfies (11), i.e. yields posterior consistency under conditions:*

(a) $H_1 n^{\rho_1} < M_n < H_2 n^{\rho_2}$

(b) $\sup_{l=1,\dots,p_{j,n}} |\boldsymbol{\beta}^{0(r)}_{j,n,l}| < M_1 < \infty$*, for all* $j = 1, \dots, D$, $r = 1, \dots, R$

(c) $\sum_{j=1}^{D} p_{j,n} \log(p_{j,n}) = o(n)$.

Remark 3 *Condition (a) in Theorem 2 gives an upper and lower bound on the sum of the Frobenius norms of tensor predictors. In tensor predictors with $\{0, 1\}$ entries (often observed for white/grey matter fMRI data), this condition simply imposes a restriction on the minimum and maximum number of* 1*s in a tensor predictor as a function of the sample size n. Condition (b) is mild, assuming the supremum of all entries in the tensor margins are bounded. Finally, condition (c) in Theorem 2 requires that $\sum_{j=D} p_{j,n}$ grows sub-linearly with sample size n. However, note that the number of cells $\prod_{j=1}^{D} p_{j,n}$ in the tensor can grow at a rate much faster than the sample size n; hence, the modeling framework allows large tensor covariates even for moderate sample sizes. Of course, condition (c) trivially holds when the dimension of the tensor is kept fixed as n grows.*

Remark 4 *Our Section on posterior consistency is based on the assumption that $\boldsymbol{\gamma} = (0, \dots, 0)'$ and error variance* 1*, however the results are trivially extendable to cases with unknown $\boldsymbol{\gamma}$ and σ^2. In such a generalization, it is important to note that one would need to assume the dim($\boldsymbol{\gamma}$) is fixed and not growing with n.*

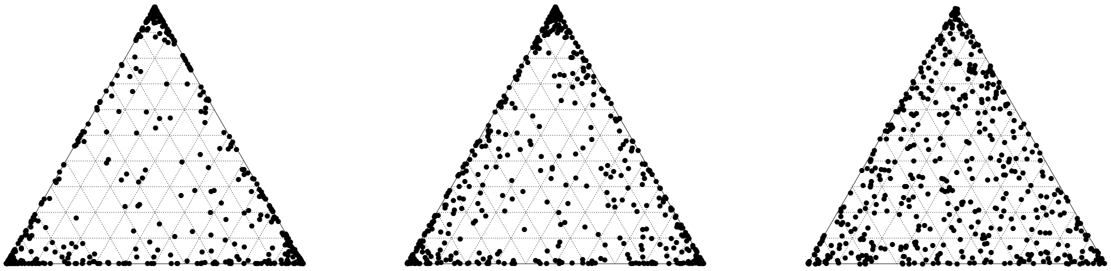
4.3 Prior Hyper-parameter Elicitation

The marginal distribution of cell coefficients (6) under the proposed M-DGDP prior (9) is not available in closed-form. To assess how a shrinkage prior on the margins induces prior on cell coefficients of the tensor, we turn to an expression for the cell-level variance:

$$\begin{aligned}
\mathrm{var}(\boldsymbol{B}_{i_1,\dots,i_D}) &= \mathbb{E}\bigg(\mathrm{var}\bigg\{\sum_{r=1}^{R}\prod_{j=1}^{D}\boldsymbol{\beta}^{(r)}_{j,i_d} \mid \boldsymbol{W}, \Lambda, \Phi, \tau\bigg\}\bigg) \\
&= \mathbb{E}_{\Phi}\bigg(\sum_{r=1}^{R}\phi_r^D\,\mathbb{E}_{\tau}\{\tau^D\}\,\mathbb{E}_{\Lambda_{\cdot,r}}\bigg\{\mathbb{E}_{\boldsymbol{W}_r|\boldsymbol{\Lambda}_r}\bigg(\prod_{j=1}^{D} w_{jr,i_j}\bigg)\bigg\}\bigg) \\
&= \frac{\Gamma(\alpha_0 + D)}{\Gamma(\alpha_0)\, b_\tau^D}(2C_\lambda)^D\,\mathbb{E}_{\Phi}\bigg(\sum_{r=1}^{R}\phi_r^D\bigg).
\end{aligned}$$

The following Lemma provides lower and upper bounds on the variance that can be useful for elicitation of default hyperparameter values.

Lemma 5 *Under M-DGDP shrinkage prior* (9) *and for $D > 1$, if $\alpha_1 = \cdots = \alpha_R = c/R$, $c \in \mathbb{N}_+$, with constants $C_\lambda = b_\lambda^2/\big((a_\lambda - 1)(a_\lambda - 2)\big)$, $a_\lambda > 2$, $A_\tau = \exp((D^2 - 3D)/2)$, then the cell-level variance is bounded below by $R\alpha_1^D(2C_\lambda/b_\tau)^D$ and above by $A_\tau(2C_\lambda/b_\tau)^D \exp(\alpha_1 RD)$.*



(a) $\boldsymbol{\alpha} = (0.2, 0.2, 0.2)$

(b) $\boldsymbol{\alpha} = (0.3, 0.3, 0.3)$

(c) $\boldsymbol{\alpha} = (0.5, 0.5, 0.5)$

Figure 1: Visualization of points in the $\mathcal{S}^2$ probability simplex for 500 independent realizations of Dirichlet($\boldsymbol{\alpha}$). As $\alpha \downarrow 0$, points increasingly tend to concentrate around vertices of the $\mathcal{S}^{R-1}$ simplex. This notion of sparsity is made precise in Yang and Dunson (2014).

Hyperparameters of the Dirichlet component in multiway prior (9) play a key role in controlling dimensionality of the model, with smaller values favoring more component-specific scales $\tau_r \approx 0$, thus effectively collapsing on a low-rank tensor factorization. Figure 1 plots realizations from the Dirichlet distribution when $R = 3$ for different concentration parameters[2] α.

A discrete uniform prior is placed on α over a grid, $\mathcal{A}$. By default, grid values are chosen to be 10 equally spaced values in $[R^{-D}, R^{-0.10}]$, letting the data tune this parameter according to the degree of sparisty present. Armagan et al. (2013a) study various choices of $(a_\lambda, \zeta = b_\lambda / a_\lambda)$ that lead to desirable shrinkage properties, such as Cauchy-like tails for $\beta_{j,k}^{(r)}$ while retaining Laplace-like shrinkage near zero. Empirical results from simulation studies across a variety of settings in Section 6 reveal no strong sensitivity to choices for hyper-parameters a_λ, b_λ. From Lemma 5, setting $a_\lambda = 3$ and $b_\lambda = \sqrt[2D]{a_\lambda}$ avoids overly narrow variance of the induced prior on tensor elements, $\boldsymbol{B}_{i_1,\ldots,i_D}$. Table 1 provides various quantiles of the induced prior on these elements under these default hyperparameter settings as a function of the parafac rank-R and tensor dimension D.

5. Posterior Computation and Model Fitting

Letting $y \in \Re$ denote a response, and $\boldsymbol{z} \in \Re^p, \boldsymbol{X} \in \otimes_{j=1}^{D} \Re^{p_j}$ predictors, we let

$$\begin{aligned} y \mid \boldsymbol{\gamma}, \boldsymbol{B}, \sigma &\sim \mathrm{N}\big(\boldsymbol{z}'\boldsymbol{\gamma} + \langle \boldsymbol{X}, \boldsymbol{B} \rangle, \sigma^2\big) \\ \boldsymbol{B} &= \sum_{r=1}^{R} \boldsymbol{B}_r, \ \boldsymbol{B}_r = \boldsymbol{\beta}_1^{(r)} \circ \cdots \circ \boldsymbol{\beta}_D^{(r)} \\ \sigma^2 &\sim \pi_\sigma, \ \boldsymbol{\gamma} \sim \pi_\gamma, \ \boldsymbol{\beta}_j^{(r)} \sim \pi_{\boldsymbol{\beta}}. \end{aligned} \tag{12}$$

[2] For simplicity we assume $\alpha_1 = \cdots = \alpha_R = \alpha$.

	R	5%	25%	50%	75%	95%
$D=2$	1	0.001	0.011	0.057	0.254	1.729
	5	0.004	0.040	0.164	0.595	3.332
	10	0.005	0.058	0.237	0.852	4.635
$D=3$	1	0.000	0.001	0.010	0.072	0.917
	5	0.000	0.009	0.061	0.341	3.382
	10	0.001	0.017	0.111	0.608	5.996

Table 1: Percentiles for $|\boldsymbol{B}_{i_1,\dots,i_D}|$ under the M-DGDP prior with default $a_\lambda = 3$, $b_\lambda = \sqrt[2D]{a_\lambda}$, $b_\tau = \alpha R^{1/D}$ $(v=1)$ and $\alpha = 1/R$. Statistics are displayed as the parafac rank-R vary and dimension D of the tensor vary.

The noise variance is modeled using a conjugate inverse-gamma prior, $\sigma^2 \sim \mathrm{IG}(v/2, vs_0^2/2)$, with $v = 2$ and s_0^2 chosen by default so $\Pr(\sigma^2 \leq 1) = 0.95$ assuming a centered and scaled response. Regression coefficients are given a conjugate normal prior $\boldsymbol{\gamma} \sim \mathrm{N}(0, \sigma^2 \boldsymbol{\Sigma}_{0\boldsymbol{\gamma}})$ and the tensor predictor is normalized over all cells to have mean zero and variance 1, allowing one to assume default values for hyper-parameters in the proposed multiway prior.

5.1 Posterior Computation

The proposed multiway prior (9) leads to Gibbs sampling scheme for most parameters of the tensor regression model (12). We rely on marginalization and blocking to reduce autocorrelation for $\big\{\big(\boldsymbol{\beta}_j^{(r)}, w_{jr}; 1 \leq j \leq D, 1 \leq r \leq R\big), (\Phi, \tau), (\boldsymbol{\gamma}, \sigma)\big\}$, drawing in sequence from $[\alpha, \Phi, \tau | \boldsymbol{B}, W]$, $[\boldsymbol{B}, W | \Phi, \tau, \boldsymbol{\gamma}, \sigma, \boldsymbol{y}]$ and $[\boldsymbol{\gamma}, \sigma | \boldsymbol{B}, \boldsymbol{y}]$ as follows:

(1) Sample $[\alpha, \Phi, \tau | \boldsymbol{B}, \boldsymbol{W}]$ compositionally as $[\alpha | \boldsymbol{B}, \boldsymbol{W}][\Phi, \tau | \alpha, \boldsymbol{B}, \boldsymbol{W}]$:

  (a) Sample from the conditional distribution of Dirichlet concentration parameter $[\alpha | \boldsymbol{B}, \boldsymbol{W}]$ via griddy-Gibbs: form a reference set by drawing M samples from $[\Phi, \tau | \alpha, \boldsymbol{B}, \boldsymbol{W}]$ for each $\alpha \in \mathcal{A}$. Set $w_{j,l} = \pi(\boldsymbol{B} | \alpha, \Phi_l, \tau_l, \boldsymbol{W}) \pi(\Phi_l, \tau_l | \alpha)$, $1 \leq l \leq M$, $p(\alpha | \boldsymbol{B}, \boldsymbol{W}) = \pi(\alpha) \sum_{l=1}^{M} w_{j,l}/M$, and $\Pr(\alpha = \alpha_j | -) = p(\alpha_j | \boldsymbol{B}, \boldsymbol{W}) / \sum_{\alpha \in \mathcal{A}} p(\alpha | \boldsymbol{B}, \boldsymbol{W})$.

  (b) Sample component-specific scales as $[\Phi, \tau | \alpha^*, \boldsymbol{B}, \boldsymbol{W}] = [\Phi | \boldsymbol{B}, \boldsymbol{W}][\tau | \Phi, \boldsymbol{B}, \boldsymbol{W}]$; define $p_0 = \sum_{j=1}^{D} p_j$, and recall $a_\tau = \sum_{r=1}^{R} \alpha_r = R\alpha$ and $b_\tau = \alpha (R/v)^{1/D}$ (see Section 3.3), then

  - draw $\psi_r \sim \mathrm{giG}(\alpha - p_0/2, 2b_\tau, 2C_r)$, $C_r = \sum_{j=1}^{D} \boldsymbol{\beta}_j^{(r)\,T} \boldsymbol{W}_{jr}^{-1} \boldsymbol{\beta}_j^{(r)}$, and set $\phi_r = \psi_r / \sum_{l=1}^{R} \psi_l$ in parallel for $1 \leq r \leq R$ (see Appendix A for definition of 'giG')
  - draw $\tau \sim \mathrm{giG}(a_\tau - Rp_0/2, 2b_\tau, 2\sum_{r=1}^{R} D_r)$, $D_r = C_r/\phi_r$.

(2) Sample from $\big\{(\boldsymbol{\beta}_j^{(r)}, w_{jr}, \lambda_{jr}); 1 \leq j \leq D, 1 \leq r \leq R\big\} | \Phi, \tau, \boldsymbol{\gamma}, \sigma, \boldsymbol{y}$ using a back-fitting procedure to produce a sequence of draws from the margin-level conditional distributions across components. For $r = 1, \dots, R$ and $j = 1, \dots, D$, sample from conditional distribution $[(\boldsymbol{\beta}_j^{(r)}, w_{jr}, \lambda_{jr}) | \boldsymbol{\beta}_{-j}^{(r)}, \boldsymbol{B}_{-r}, \Phi, \tau, \boldsymbol{\gamma}, \sigma, \boldsymbol{y}]$, where $\boldsymbol{\beta}_{-j}^{(r)} = \{\boldsymbol{\beta}_l^{(r)}, l \neq j\}$ and $\boldsymbol{B}_{-r} = \boldsymbol{B} \setminus \boldsymbol{B}_r$;

  (a) draw $[w_{jr}, \lambda_{jr} | \boldsymbol{\beta}_j^{(r)}, \phi_r, \tau] = [w_{jr} | \lambda_{jr}, \boldsymbol{\beta}_j^{(r)}, \phi_r, \tau][\lambda_{jr} | \boldsymbol{\beta}_j^{(r)}, \phi_r, \tau]$:

- draw $\lambda_{jr} \sim \text{Ga}\big(a_\lambda + p_j, b_\lambda + ||\boldsymbol{\beta}_j^{(r)}||_1/\sqrt{\phi_r \tau}\big)$; and
- draw $w_{jr,k} \sim \text{giG}\big(\frac{1}{2}, \lambda_{jr}^2, \beta_{j,k}^{2\,(r)}/(\phi_r \tau)\big)$ independently for $1 \leq k \leq p_j$

(b) draw $\boldsymbol{\beta}_j^{(r)} \sim \text{N}(\boldsymbol{\mu}_{jr}, \boldsymbol{\Sigma}_{jr})$: define $h_{i,j,k}^{(r)} = \sum_{d_1=1,\ldots,d_D=1}^{p_1,\ldots,p_D} I(d_j = k)\, x_{d_1,\ldots,d_D} \big(\prod_{l \neq j} \beta_{l,i_l}^{(r)}\big)$, $\boldsymbol{H}_{i,j}^{(r)} = (h_{i,j,1}^{(r)}, \ldots, h_{i,j,p_j}^{(r)})'$, $\tilde{y}_i = y_i - \boldsymbol{z}_i'\boldsymbol{\gamma} - \sum_{l \neq r} \langle \boldsymbol{X}_i, \boldsymbol{B}_l \rangle$ for $1 \leq i \leq n$; then $\boldsymbol{\Sigma}_{jr} = \big(\boldsymbol{H}_j^{(r)\,T} \boldsymbol{H}_j^{(r)}/\sigma^2 + \boldsymbol{W}_{jr}^{-1}/(\phi_r \tau)\big)^{-1}$, $\boldsymbol{\mu}_{jr} = \boldsymbol{\Sigma}_{jr} \boldsymbol{H}_j^{(r)} \tilde{\boldsymbol{y}}/\sigma^2$

(3) Sample $[\boldsymbol{\gamma}, \sigma | \boldsymbol{B}, \boldsymbol{y}] = [\boldsymbol{\gamma} | \sigma, \tilde{\boldsymbol{y}}][\sigma^2 | \tilde{\boldsymbol{y}}]$; define $\tilde{y}_i = y_i - \langle \boldsymbol{X}_i, \boldsymbol{B} \rangle$ for $1 \leq i \leq n$, then

(a) draw $\sigma^2 \sim \text{IG}(a_\sigma, b_\sigma)$, $a_\sigma = (n+v)/2$, $b_\sigma = \big(v s_0^2 + ||\tilde{\boldsymbol{y}}||_2^2 - \tilde{\boldsymbol{y}}^T \boldsymbol{Z} \boldsymbol{\mu}_{\boldsymbol{\gamma}}\big)/2$

(b) draw $\boldsymbol{\gamma} \sim \text{N}\big(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \sigma^2 \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\big)$, $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = \big(\boldsymbol{Z}^T \boldsymbol{Z} + \boldsymbol{\Sigma}_{0\boldsymbol{\gamma}}^{-1}\big)^{-1}$, $\boldsymbol{\mu}_{\boldsymbol{\gamma}} = \boldsymbol{\Sigma}_{\boldsymbol{\gamma}} \boldsymbol{Z}^T \tilde{\boldsymbol{y}}$.

6. Simulation Studies

To illustrate finite-sample performance of the proposed multiway priors, we show results from a simulation study with various dimensionality (p, R) and define $\bar{b} = \max |\boldsymbol{B}_{i_1,\ldots,i_D}^0|$ as the maximum signal size. Throughout, set $p_j = p$, true error variance $\sigma_0^2 = 1$ and $\bar{b} = 1$ for convenience. In addition, we set the true vector coefficient $\boldsymbol{\gamma}_0 = (0, \ldots, 0)$ and focus exclusively on inference for tensor parameter $\boldsymbol{B}$. The following simulated setups are considered:

1. "Generated" tensor: We construct tensor parameters having rank $R_0 = \{3, 5\}$ with $p = \{64, 100\}$ and $D = 2$.

2. "Ready made" tensor: We use three tensor (2D) images without generating them from a parafac decomposition with known rank.

Five replicated datasets with $n = 1000$ are generated according to (12) with $x_{i_1,\ldots,i_D} \sim \text{N}(0, 1)$. The tensor parameters considered are shown in Figure 2, where the magnitude of the non-zero cells is $\bar{b} = 1$. Examples are chosen to demonstrate recovery of cell-level coefficients across varying degrees of complexity (dimension, parafac rank) and sparsity (% of non-zero cells; see Figure 2). The performance of our method with M-DGDP prior (9) (BTR) is compared with (i) frequentist tensor regression with penalization (FTR)(Zhou et al., 2013); and (ii) Lasso (on the vectorized tensor predictor). Comparisons are based on (a) cell mean squared estimation error (true non-zero, true zero, and overall); and (b) frequentist coverage (and length) of 95% credible intervals.

By default, BTR uses $R = 10$ as an upper bound on the tensor parafac rank, minimizing effects of extra dimensions automatically and concentrating on a lower rank coefficient tensor as MCMC proceeds. We ran FTR with various choices of R and found equivalent performance for R between 3 to 15, thus the default value is set to $R = 10$ to ensure comparability with the BTR fitting. MCMC for BTR was run for 1300 iterations, with a 300 iteration burn-in and remaining samples thinned by 5. The latter was chosen to keep runtime between BTR and FTR similar for 3D simulation studies in Section 7. There, the total runtime using non-optimized R code on an x86×64 Intel(R) Core(TM) i7-3770 is between 6.2 - 7.5 hours. In simulation studies, the tuning parameter in FTR is selected

		R3-ex	R5-ex	Shapes	Eagle	Palmtree	Horse
$\lvert\text{cell}_0\rvert > 0$	BTR	$\mathbf{0.023}_{0.00}$	$\mathbf{0.021}_{0.00}$	$\mathbf{0.243}_{0.01}$	$\mathbf{0.226}_{0.02}$	$\mathbf{0.316}_{0.01}$	$\mathbf{0.278}_{0.01}$
	FTR	$0.035_{0.00}$	$0.030_{0.00}$	$0.415_{0.03}$	$0.354_{0.03}$	$0.435_{0.02}$	$0.391_{0.03}$
	Lasso	$0.628_{0.02}$	$0.822_{0.03}$	$0.619_{0.07}$	$0.665_{0.03}$	$0.698_{0.03}$	$0.888_{0.01}$
$\lvert\text{cell}_0\rvert = 0$	BTR	$\mathbf{0.011}_{0.00}$	$\mathbf{0.014}_{0.00}$	$\mathbf{0.071}_{0.00}$	$\mathbf{0.085}_{0.00}$	$\mathbf{0.100}_{0.01}$	$\mathbf{0.137}_{0.00}$
	FTR	$0.022_{0.00}$	$0.020_{0.00}$	$0.127_{0.02}$	$0.163_{0.03}$	$0.159_{0.00}$	$0.215_{0.02}$
	Lasso	$0.090_{0.00}$	$0.098_{0.02}$	$0.081_{0.01}$	$0.097_{0.00}$	$0.094_{0.01}$	$0.155_{0.02}$
Overall	BTR	**0.013**	**0.015**	**0.093**	**0.102**	**0.131**	**0.172**
	FTR	0.023	0.021	0.164	0.184	0.196	0.257
	Lasso	0.187	0.288	0.179	0.204	0.217	0.407

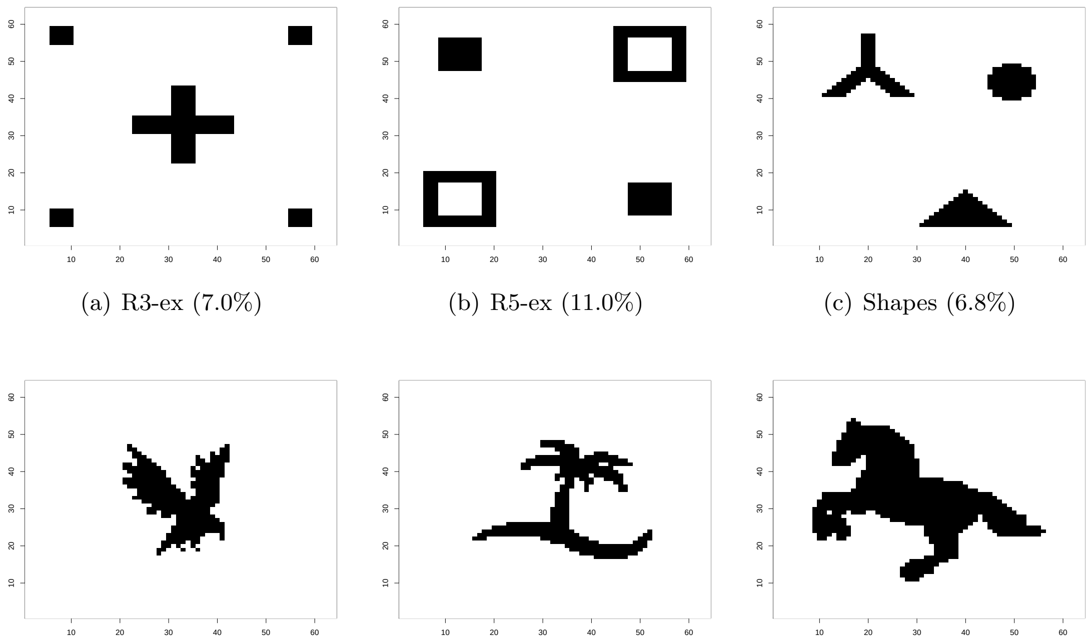
Table 2: Comparison of cell estimation as measured by root mean squared error (RMSE) for the six 2D tensor images portrayed in Figure 2. Results from FTR (Zhou et al., 2013) use $R = 10$. For BTR, $R = 10$ is used as an upper bound to the tensor parafac rank. Subscript shows the standard error over a few replicated simulations.

over a grid of values to minimize RMSE for the tensor predictor[3]. In real applications, cross validation instead would be used to select the tuning parameter that results in lowest hold-out predictive RMSE for FTR. Assuming 10-fold cross validation were used over a vector of 20 tuning parameters, FTR would have a runtime of approximately 8 hours. It also needs to be mentioned that the convergence of parameters in BTR is extremely rapid with an average effective sample size (ESS) ≈ 600 over 1000 iterations.

Cell-level RMSE reported in Table 2 demonstrates that our method (BTR) consistently out performs FTR. When the tensor parameter has a low-rank parafac decomposition ('R3-ex' and 'R5-ex'), BTR and FTR perform best, with BTR having lower RMSE on both true zero and non-zero cells. This validates empirically prior (9) along with our suggested default hyper-parameter choices in Section 3. In particular, the tensor coefficient in BTR has three different types of shrinkage parameters: global, local and shrinkage across ranks. Such a careful construction of shrinkage prior on $\boldsymbol{B}$ adapts to varying degrees of sparsity, shrinking many tensor coefficients close to zero while accurately estimating nonzero cells. FTR shrinkage being dependent on only local parameters suffers in terms of both inferential and predictive performances.

Table 3 demonstrate that BTR yields 95% credible intervals with good frequentist coverage across each of the simulated settings, both overall as well as on the true non-zero coefficients. Our method is one of the first to offer uncertainty quantification for tensor valued predictors.Finally, Table 4 provides evidence of the robustness of our method to increasing predictor dimension using two of the simulated examples. In both cases, RMSE for FTR worsens considerably on the true zero coefficients. For the true nonzero cells, RMSE increases for both methods as the margin dimension increases; however on a relative basis, FTR worsens considerably more, while on an absolute scale, BTR remains the clear winner.

[3] To choose initial values, a preliminary analysis was run with a coarsened 16×16 image.



(a) R3-ex (7.0%)

(b) R5-ex (11.0%)

(c) Shapes (6.8%)

(d) Eagle (7.4%)

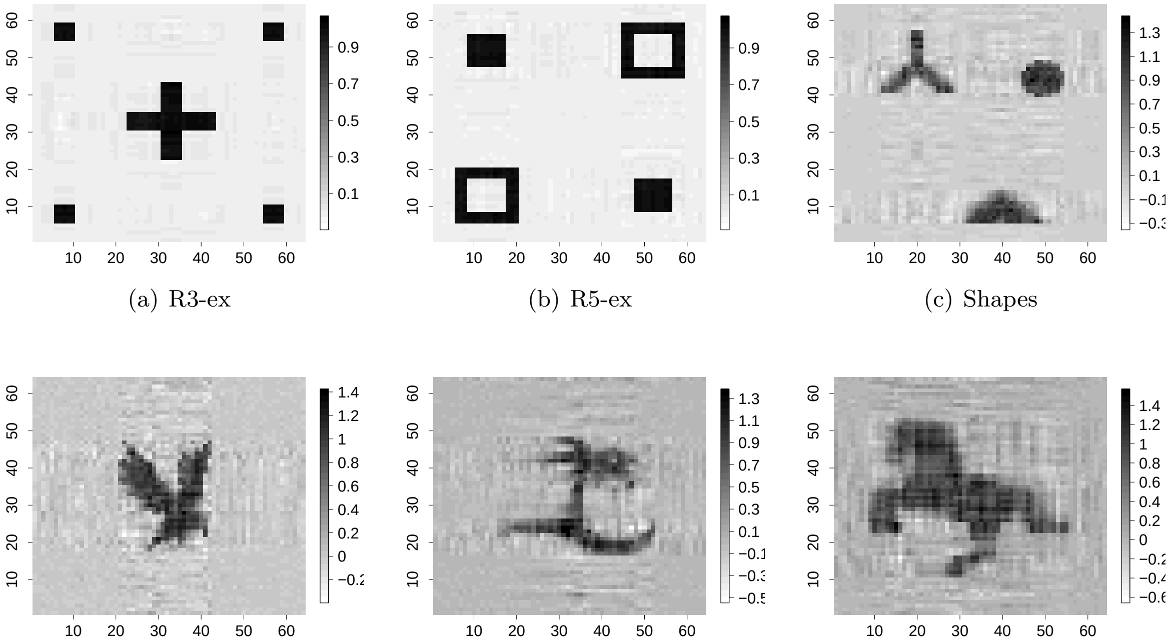
(e) Palmtree (8.0%)

(f) Horse (18.6%)

Figure 2: Simulated data with 64 × 64 2D tensor images ($p = 64$, $D = 2$). Row 1: The first two images (from left) have a rank-3 and rank-5 parafac decomposition; the third image is "regular", although does not have a low-rank parafac decomposition. Row 2: All three images are irregular, and do not have a low-rank parafac decomposition. Sparsity (% non-zero cells) are displayed in sub-captions.

		R3-ex	R5-ex	Shapes	Eagle	Palmtree	Horse
$\lvert \text{cell}_0 \rvert > 0$	coverage	$0.986_{0.02}$	$0.946_{0.02}$	$0.747_{0.01}$	$0.731_{0.04}$	$0.677_{0.04}$	$0.795_{0.02}$
Overall	coverage	$0.995_{0.01}$	$0.970_{0.01}$	$0.965_{0.00}$	$0.940_{0.02}$	$0.948_{0.02}$	$0.927_{0.01}$
	length	$0.066_{0.01}$	$0.061_{0.01}$	$0.290_{0.00}$	$0.301_{0.03}$	$0.410_{0.03}$	$0.566_{0.02}$

Table 3: Row 1: Average coverage of 95% posterior credible intervals for all the cells of $\boldsymbol{B}$ for which the true cell coefficient is nonzero. Row 2: Average coverage of 95% posterior credible intervals for all the cells of $\boldsymbol{B}$. Row 3: Average length of 95% posterior credible intervals for all the cells of $\boldsymbol{B}$. Subscripts show standard errors over replicated simulations.



(a) R3-ex

(b) R5-ex

(c) Shapes

(d) Eagle

(e) Palmtree

(f) Horse

Figure 3: Recovered images for the 64×64 2D tensor images in Figure 2 using our proposed BTR method. Here, $R = 10$ is used as an upper bound to the tensor parafac rank.

		$\lvert\text{cell}_0\rvert$	R5-ex		Shapes	
			64	100	64	100
BTR	coverage	> 0	$0.946_{0.02}$	$0.991_{0.01}$	$0.747_{0.01}$	$0.590_{0.06}$
	length	> 0	$0.061_{0.01}$	$0.069_{0.01}$	$0.290_{0.00}$	$0.247_{0.01}$
	rmse	> 0	$0.021_{0.00}$	$0.032_{0.01}$	$0.243_{0.01}$	$0.320_{0.03}$
	rmse	$= 0$	$0.014_{0.00}$	$0.014_{0.00}$	$0.071_{0.00}$	$0.063_{0.00}$
FTR	rmse	> 0	$0.030_{0.00}$	$0.369_{0.06}$	$0.415_{0.03}$	$0.586_{0.14}$
	rmse	$= 0$	$0.020_{0.00}$	$0.111_{0.02}$	$0.127_{0.02}$	$0.135_{0.02}$

Table 4: Sensitivity analysis of cell estimation error (RMSE) as the tensor dimension increases; here $p_j = p \in \{64, 100\}$ for the 2D tensor images 'R5-ex' and 'Shapes'.

7. Simulated response with a real 3D brain image

We analyze data containing 3D MRI images for 550 adolescents, with information such as age and sex available. Age and sex are treated as ordinary scalar covariates while 3D MRI images act as tensor covariates. Let $\boldsymbol{X}$ denote a $30\times30\times30$ 3D MRI image, Z_1 be the age and Z_2 be the sex of an individual. The response is simulated using $y \sim \mathrm{N}\left(\boldsymbol{Z}'\boldsymbol{\gamma} + \langle \boldsymbol{X}, \boldsymbol{B}_0\rangle, \sigma^2\right)$, where $\boldsymbol{Z}$ denotes $(Z_1, Z_2)'$, $\boldsymbol{\gamma} \in \mathcal{R}^2$ and $\boldsymbol{B}_0 \in \mathcal{R}^{30\times30\times30}$.

We assume the true $\boldsymbol{B}_0$ is a rank 2 tensor, with $\boldsymbol{B}_0 = \boldsymbol{a}_1 \circ \boldsymbol{a}_2 \circ \boldsymbol{a}_3 + \boldsymbol{b}_1 \circ \boldsymbol{b}_2 \circ \boldsymbol{b}_3$. Initialization and standardization of predictors follow exactly as prescribed in Section 5. By varying $\boldsymbol{a}_i$'s and $\boldsymbol{b}_i$'s, the following cases with varying degrees of sparsity in the tensor parameter $\boldsymbol{B}_0$ are considered:
Case 1: $\boldsymbol{b}_1 = \boldsymbol{b}_2 = (0, \ldots, 0, \sin((1:15)*\pi/4))$, $\boldsymbol{b}_3 = (\sin((1:10)*\pi/4), 0, \ldots, 0)$,
$\boldsymbol{a}_1 = (0, \ldots, 0, \sin((1:10)*\pi/4))$, $\boldsymbol{a}_2 = (0, \ldots, 0, \cos((1:15)*\pi/4))$,
$\boldsymbol{a}_3 = (\sin((1:15)*\pi/4), 0, \ldots, 0)$.
Case 2: $\boldsymbol{b}_1 = \boldsymbol{b}_2 = (0, \ldots, 0, \sin((1:15)*\pi/6))$, $\boldsymbol{b}_3 = (\sin((1:20)*\pi/6), 0, \ldots, 0)$,
$\boldsymbol{a}_1 = (0, \ldots, 0, \sin((1:15)*\pi/4))$, $\boldsymbol{a}_2 = (0, \ldots, 0, \cos((1:10)*\pi/6))$,
$\boldsymbol{a}_3 = (\sin((1:15)*\pi/6), 0, \ldots, 0)$.
Case 3: $\boldsymbol{b}_1 = \boldsymbol{b}_2 = (0, \ldots, 0, \sin((1:20)*\pi/6))$, $\boldsymbol{b}_3 = (\sin((1:20)*\pi/6), 0, \ldots, 0)$,
$\boldsymbol{a}_1 = (0, \ldots, 0, \sin((1:10)*\pi/4))$, $\boldsymbol{a}_2 = (0, \ldots, 0, \cos((1:20)*\pi/4))$,
$\boldsymbol{a}_3 = (\sin((1:20)*\pi/6), 0, \ldots, 0)$.

We implement BTR, FTR, and Lasso on the vectorized tensor. As before, we present results for FTR with $R = 10$ (See additional discussion in Section 6 on FTR default setup). Point estimates for coefficients corresponding to age and sex covariates are provided in Table 6. Table 5 summarizes RMSEs for the estimated tensor coefficients for each method. BTR shows at least a 15% improvement over FTR on simulated cases considered. Evidently BTR tends to outperform FTR and vectorized lasso by a greater margin in less sparse settings as well. Importantly, every parameter in BTR is auto-tuned, while the `TensorReg` toolbox used for FTR (Zhou et al., 2013) requires calibrating tuning parameter values specific to each setting. Note that rather than using cross validation, tuning parameters in these experiments were chosen to provide the lowest possible (most optimistic) RMSE for the tensor coefficient. FTR fixes $R = 10$ based on findings discussed in Section 6 while BTR sets $R = 10$ as an upper bound, concentrating on a lower dimension parafac rank via adaptive shrinkage. While vectorized lasso and FTR do not come equipped with parameter uncertainty estimates, Table 7 demonstrates how BTR consistently provides over 95% coverage across examples with varying degrees of sparsity.

Finally, Table 8 provides a measure of mixing efficiency for a single MCMC run in each of the simulated cases considered (post burn-in over the remaining 1000 MCMC samples). All reported RMSE and coverage statistics were computed over these draws as well (thinning by 5 as previously discussed in Section 6).

8. Brain Connectome Data Analysis

Brain connectome data are known by neuroscientists to have low signal-to-noise ratio, and effective modeling is often hindered as the sample size is often very limited compared to the

		Case 1	Case 2	Case 3
	BTR	**0.39**	**0.30**	**0.34**
$\lvert\text{cell}_0\rvert > 0$	FTR	0.46	0.41	0.43
	Lasso	0.46	0.42	0.44
	BTR	**0.04**	**0.14**	**0.10**
$\lvert\text{cell}_0\rvert = 0$	FTR	0.00	0.00	0.00
	Lasso	0.01	0.03	0.02
	BTR	**0.13**	**0.20**	**0.17**
Overall	FTR	0.15	0.22	0.18
	Lasso	0.15	0.23	0.18

Table 5: Comparison of cell estimation as measured by root mean squared error (RMSE) for the coefficients in case 1,2, 3 corresponding to 3D tensor images. Results from both BTR and FTR (Zhou et al., 2013) use $R = 10$.

		Case 1	Case 2	Case 3
γ_1 (truth = 0.5)	BTR	**0.57**	**0.54**	**0.33**
	FTR	0.46	0.85	0.95
γ_2 (truth = 2.0)	BTR	**2.00**	**2.04**	**1.86**
	FTR	1.87	0.22	3.30

Table 6: Point estimates for age and sex coefficients under BTR and FTR Zhou et al. (2013). True parameter values are also provided.

	Case 1	Case 2	Case 3
Coverage	0.98	0.96	0.99
Length	0.54	0.87	2.16

Table 7: Length and coverage of 95% credible intervals for BTR. Values are reported as averages over all voxels of the tensor coefficient.

		τ^2	Scalar predictor γ_i (Ave.)	Tensor predictor β_{ijk} (Ave.)
Case 1 (88% sparsity)	lag-2	0.04	0.34	0.45
	lag-4	0.01	0.11	0.22
Case 2 (82% sparsity)	lag-2	0.08	0.33	0.46
	lag-4	0.03	0.12	0.23
Case 3 (70% sparsity)	lag-2	0.09	0.43	0.53
	lag-4	0.02	0.30	0.40

Table 8: MCMC autocorrelation of the proposed BTR method on data studies of Section 7 generated using 3D brain MRI scans.

number of cells in the tensor predictor. In this setting, developing well calibrated predictive models is thus of key importance.

To investigate the performance of competing methods outside the class of fMRI brain image data, we present an analysis using a brain connectome dataset on structural connectivity. Data are extracted from diffusion tensor imaging (DTI) and consist of estimates of the number of “fibers” connecting pairs of brain regions for 109 individuals. For each individual, brain connections among 70 brain regions (following desikan atlas) are encoded by a 70×70 weighted adjacency matrix. The (i, j)-th off-diagonal entry in the adjacency matrix is the estimated number of fiber tracts connecting the i-th and j-th brain region. The data also provides 10 clinical covariates for every individual, including sex, age, openness, agreeableness and conscientiousness.

The focus of this study is on developing a predictive model with Creativity Composite Index (CCI) as a response fitted against clinical covariates and a tensor covariate (i.e., the weighted adjacency matrix). Implementing FTR on this data using the `TensorReg` package was attempted, however, functions in the toolbox require $n > R \times p$. In this example, because $n = 109$ and $p = 70$, it is only possible therefore to fit FTR with $R = 1$, which has previously been found to perform poorly by Zhou et al. (2013). We therefore compare our proposed method (BTR) to the vectorized Lasso on the basis of their predictive performance. To assess the predictive performance, the sample of $n = 109$ individuals are divided into 10 folds. Both vectorized lasso and BTR are fitted on 9 folds as training data and the remaining fold as the hold out sample. This is carried out for each of the 10 folds and predictive inferences are obtained for both vectorized lasso and BTR.

Table 9 reports the root mean squared error (RMSE) and correlation between observed and predicted responses, here average is over the 10 crossvalidated folds. For reference, average RMSE of the null model is 10.03 with a standard deviation of 2.40 across the

Method	avg(RMSE)	sd(RMSE)	avg(cov.)	sd(cov.)	avg(cor.)	sd(cor.)
Lasso	9.21	2.18	63%	20%	0.31	0.11
BTR	9.03	1.64	91%	10%	0.32	0.13

Table 9: mean and standard deviation of RMSEs and cor(y_{obs}, y_{pred}) of Lasso and BTR over 10 folds of the data. It also provides mean and standard deviation of coverages of Lasso and BTR over 10 folds of the data

folds. Given the very high degree of sparsity in the connectome adjacency matrix, it is not surprising that the Lasso is competitive to BTR. However, note that BTR detects this signal with far fewer effective parameters as compared to vectorized lasso. Finally, we measure coverage of 95% predictive intervals for all competitors. The latter is of course a byproduct of our fully Bayesian approach (BTR), while for the Lasso we use a two-staged approach. First we estimate the regression coefficients and subsequently construct approximate 95% predictive intervals based on the normal response-model centered on the predictive mean with variance equal to the estimated residual variance.

9. Discussion

This work develops a novel class of prior distributions on tensor valued predictors which substantially reduces dimensionality relative to vectorizing, providing a multiway analogue of vector shrinkage priors, and enabling high dimensional region selection. The prior on tensor coefficient constructed here imparts shrinkage of the tensor components at global and local levels, while also encouraging shrinkage towards low rank tensor decomposition. In contrast, existing penalization framework on the tensor coefficient shrinks only at the global level. Strong theoretical results are proved for the proposed class of multiway shrinkage priors and a computationally efficient MCMC algorithm is developed in the regression setting. We plan to extend methods developed here to settings where the measured response for each subject is binary (e.g., indicator of a heath outcome) or multivariate, i.e., $\boldsymbol{y} = (y_1, ..., y_d)$. Also, the current framework of Bayesian tensor regression fixes rank R of the PARAFAC decomposition at reasonably large value. It might be of interest to learn the PARAFAC rank R by adding a discrete prior distribution on R.

In various longitudinal studies, monitoring the evolution in the predictor response relationship (i.e., changes to the scalar and tensor parameters) is of fundamental interest. One application involves subjects receiving various treatments (e.g., chemotherapy), with tensor valued predictors corresponding to mRI (fMRI) scans obtained at regular intervals over a period of time. In such settings it is of crucial importance to monitor the progression of the disease in response to the treatment being administered. We plan to extend our method to such settings.

Acknowledgement

The authors would like to thank Joshua T. Vogelstein from Johns Hopkins university for gracefully allowing us to utilize their brain connectome dataset in the real data analysis of our proposed BTR method.

Appendix A

MCMC algorithm

The following derivations concern the M-DGDP prior (9) and the sampling algorithm outlined in Section 5.1.

For step (1b) Recall from Section 3.3 that $\tau \sim \text{Ga}(a_\tau, b_\tau)$ and $\Phi \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_R)$ and denote $p_0 = \sum_{j=1}^{D} p_j$. Then,

$$\begin{aligned}
\pi(\Phi|\boldsymbol{B}, \boldsymbol{\omega}) &\propto \pi(\Phi) \int_0^\infty \pi(\boldsymbol{B}|\boldsymbol{\omega}, \Phi, \tau)\pi(\tau) d\tau \\
&\propto \Big[\prod_{r=1}^R \phi_r^{\alpha_r - 1}\Big] \int_0^\infty \prod_{r=1}^R \Big[(\tau\phi_r)^{-p_0/2} \exp\Big(-\frac{1}{\tau\phi_r}\sum_{j=1}^d ||\boldsymbol{\beta}_{jr}||^2/(2\omega_{jr})\Big)\Big] \tau^{a_\tau - 1}\exp(-b_\tau\tau) d\tau \\
&\propto \Big[\prod_{r=1}^R \phi^{\alpha_r - \frac{p_0}{2} - 1}\Big] \int_0^\infty \tau^{a_\tau - R\frac{p_0}{2} - 1} \prod_{r=1}^R \exp\Big(-\frac{C_r}{\tau\phi_r} - b_\tau(\tau\phi_r)\Big)\, d\tau
\end{aligned}$$

with $C_r = \sum_{j=1}^d ||\boldsymbol{\beta}_{jr}||^2/(2\omega_{jr})$. When $a_\tau = \sum_{r=1}^R \alpha_r$, this contains the kernel of a generalized inverse Gaussian (gIG) distribution for $(\tau\phi_r)$. Recall: $X \sim f_X(x) = \text{giG}(p, a, b) \propto x^{p-1}\exp(-(ax + b/x)/2)$. Following Lemma 9 in the Appendix B, for independent random variable $T_r \sim f_r$ on $(0, \infty)$, the joint density of $\{\phi_r = T_r / \sum_{\tilde{r}} T_{\tilde{r}} : r = 1, \dots, R\}$ has support on $\mathcal{S}^{R-1}$. In particular,

$$f(\phi_1, \dots, \phi_{R-1}) = \int_0^\infty t^{R-1} \prod_{r=1}^R f_r(\phi_r t)\; dt, \quad \phi_R = 1 - \sum_{r<R} \phi_r.$$

Substituting $f_r(x) \propto x^{-\delta_r}\exp(-C_r/x)\exp(-b_\tau x)$ in the above expression yields

$$\begin{aligned}
f(\phi_1, \dots, \phi_{R-1}) &\propto \int_0^\infty \tau^{R-1} \prod_{r=1}^R (\phi_r\tau)^{-\delta_r} \exp\Big(-\frac{C_r}{(\phi_r\tau)} - b_\tau(\phi_r\tau)\Big) d\tau \\
&= \Big[\prod_{r=1}^R \phi^{-\delta_r}\Big] \int_0^\infty \tau^{R - \sum_r \delta_r - 1} \prod_{r=1}^R \exp\Big(-\frac{C_r}{(\phi_r\tau)} - b_\tau(\phi_r\tau)\Big) d\tau.
\end{aligned}$$

Matching exponents between this expression and the preceding one implies (1) $a_\tau - R(p_0/2) - 1 = R - \sum_r \delta_r - 1$, and (2) $\delta_r = 1 + p_0/2 - \alpha_r$. Then,

$$a_\tau = R(1 + p_0/2) - \sum_r \delta_r = R(1 + p_0/2) - (R + Rp_0/2 - \sum_r \alpha_r) = \sum_r \alpha_r$$

as previously noted. Hence, draws from $[\Phi|\alpha, \boldsymbol{B}, \boldsymbol{W}]$ are obtained by sampling $T_r \sim f_r = \text{giG}(\alpha_r - p_0/2, 2b_\tau, 2C_r)$ independently for $r = 1, \dots, R$, and renormalizing.

Proof of lemma 5

Proof Using priors defined in (9), one has $C_\lambda = \mathbb{E}_\lambda(1/\lambda^2) = \frac{b_\lambda^2}{(a_\lambda-1)(a_\lambda-2)}$ for any $a_\lambda > 2$. In addition, the following inequalities are useful to bound the latter quantity:

- If $\alpha_1 = c/R$, $c \in \mathbb{N}_+$, $\Gamma(\alpha_0+D)/\Gamma(\alpha_0) = \alpha_0(\alpha_0+1)\cdots(\alpha_0+D-1)$. Using the fact that $\log(x+1) \leq x$, $x \geq 0$, one has $\log(\alpha_0)+\cdots+\log(\alpha_0+D-1) \leq \alpha_0 D - 1 + \sum_{k=1\vee D-2}^{D-2} k$. Then $\alpha_0^D \leq \Gamma(\alpha_0+D)/\Gamma(\alpha_0) \leq A_\tau \exp(\alpha_0 D)$ where $A_\tau = \exp(-1+\sum_{k=1\vee D-2}^{D-2} k) = \exp\big((D^2-3D)/2\big)$, $D \geq 2$.

- Let $||x||_r$ denote the L^rth norm. Trivially, $||\Phi||_D^D \leq 1$; in addition, by Hölder's inequality, for any $x \in \Re^k$ and $0 < r < p$, one has $||x||_p \geq k^{-\left(\frac{1}{r}-\frac{1}{p}\right)}||x||_r$. In our setting, $D \geq 2$. Taking $r = 1$ in the latter yields $||\Phi||_D^D \geq R^{-(D-1)}$.

Recall $\alpha_0 = \sum_{r=1}^R \alpha_r = \alpha_1 R$. This leads to the lower and upper bounds for the prior voxel-level variance:

$$\begin{aligned}
\mathrm{var}(B_{i_1,\ldots,i_D}) &\geq (2C_\lambda)^D\,(\alpha_1 R)^D R^{-(D-1)}/b_\tau^D \;= (2C_\lambda)^D\,\alpha_1^D R/b_\tau^D \\
\mathrm{var}(B_{i_1,\ldots,i_D}) &\leq A_\tau (2C_\lambda)^D\,\exp(\alpha_1\,RD)/b_\tau^D.
\end{aligned}$$

■

Consistency proofs

The proof of Theorem 1 relies in part on the existence of exponentially consistent tests.

Definition An exponentially consistent sequence of test functions $\Phi_n = I(\boldsymbol{y}_n \in \mathcal{C}_n)$ for testing $H_0 : \boldsymbol{B}_n = \boldsymbol{B}_n^0$ vs. $H_1 : \boldsymbol{B}_n \neq \boldsymbol{B}_n^0$ satisfies

$$\mathbb{E}_{\boldsymbol{B}_n^0}(\Phi_n) \leq c_1 \exp(-b_1 n), \qquad \sup_{\boldsymbol{B}_n \in \mathcal{B}_n^c} \mathbb{E}_{\boldsymbol{B}_n}(1-\Phi_n) \leq c_2 \exp(-b_2 n)$$

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

Lemma 6 *There exist an exponentially consistent sequence of tests Φ_n for testing H_0 : $\boldsymbol{B}_n = \boldsymbol{B}_n^0$ vs. $H_1 : \boldsymbol{B}_n \neq \boldsymbol{B}_n^0$.*

Proof We begin by stating that $\sum_{i=1}^n \big(y_i - \langle \boldsymbol{X}_i, \boldsymbol{B}_n^0 \rangle\big)^2 \sim \chi_n^2$ under $\boldsymbol{B}_n^0$. We choose the critical region of the test Φ_n as $\mathcal{C}_n = \Big\{\boldsymbol{B}_n : \frac{1}{n}\sum_{i=1}^n \big(y_i - \langle \boldsymbol{X}_i, \boldsymbol{B}_n^0 \rangle\big)^2 > \epsilon/4\Big\}$. Note that

$$\mathbb{E}_{\boldsymbol{B}_n^0}(\Phi_n) = P_{\boldsymbol{B}_n^0}\Big(\sum_{i=1}^n \big(y_i - \langle \boldsymbol{X}_i, \boldsymbol{B}_n^0 \rangle\big)^2 > n\epsilon/4\Big) \leq \exp\Big(-\frac{n\epsilon}{16}\Big), \text{ for large n,}$$

where the last line follows by simplifying Lemma 1 in Laurent and Massart (2000).

Now we will use the fact that

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(y_i - \langle \boldsymbol{X}_i, \boldsymbol{B}_n^0\rangle)^2\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \langle \boldsymbol{X}_i, \boldsymbol{B}_n\rangle\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(\langle \boldsymbol{X}_i, \boldsymbol{B}_n - \boldsymbol{B}_n^0\rangle\right)^2 + \frac{2}{n}\sum_{i=1}^{n}\left(y_i - \langle \boldsymbol{X}_i, \boldsymbol{B}_n\rangle\right)\left(\langle \boldsymbol{X}_i, \boldsymbol{B}_n - \boldsymbol{B}_n^0\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\langle \boldsymbol{X}_i, \boldsymbol{B}_n - \boldsymbol{B}_n^0\rangle\right)^2 + \frac{1}{n}\sum_{i=1}^{n} KL_i + \frac{2}{n}\sum_{i=1}^{n}(y_i - \langle \boldsymbol{X}_i, \boldsymbol{B}_n\rangle)\langle \boldsymbol{X}_i, \boldsymbol{B}_n - \boldsymbol{B}_n^0\rangle.
\end{aligned}
$$

Note that, under $\boldsymbol{B}_n$,

$$
\frac{2}{n}\sum_{i=1}^{n}(y_i - \langle \boldsymbol{X}_i, \boldsymbol{B}_n\rangle)\langle \boldsymbol{X}_i, \boldsymbol{B}_n - \boldsymbol{B}_n^0\rangle \sim N(0, \frac{4}{n^2}\sum_{i=1}^{n} KL_i),
$$

so that, $\frac{2}{n}\sum_{i=1}^{n}(y_i - \langle \boldsymbol{X}_i, \boldsymbol{B}_n\rangle)\langle \boldsymbol{X}_i, \boldsymbol{B}_n - \boldsymbol{B}_n^0\rangle = \sqrt{\frac{4}{n}\sum_{i=1}^{n} KL_i}\frac{Z}{\sqrt{n}}$, where $Z \sim N(0,1)$. Thus,

$$
\begin{aligned}
&\sup_{\boldsymbol{B}_n \in \mathcal{B}_n^c} \mathbb{E}_{\boldsymbol{B}_n}(1-\Phi_n) = \sup_{\boldsymbol{B}_n \in \mathcal{B}_n^c} P_{\boldsymbol{B}_n}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \langle \boldsymbol{X}_i, \boldsymbol{B}_n^0\rangle)^2 \le \epsilon/4\right)\\
&\le \sup_{\boldsymbol{B}_n \in \mathcal{B}_n^c} P_{\boldsymbol{B}_n}\left(\left|\left|\sqrt{\frac{4}{n}\sum_{i=1}^{n} KL_i}\frac{Z}{\sqrt{n}} + \frac{1}{n}\sum_{i=1}^{n} KL_i\right| - \left|\frac{1}{n}\sum_{i=1}^{n}(y_i - \langle \boldsymbol{X}_i, \boldsymbol{B}_n\rangle)^2\right|\right| \le \epsilon/4\right)\\
&\le \sup_{\boldsymbol{B}_n \in \mathcal{B}_n^c} P_{\boldsymbol{B}_n}\left(\left|\sqrt{\frac{4}{n}\sum_{i=1}^{n} KL_i}\frac{Z}{\sqrt{n}} + \frac{1}{n}\sum_{i=1}^{n} KL_i\right| - \epsilon/4 \le \left|\frac{1}{n}\sum_{i=1}^{n}(y_i - \langle \boldsymbol{X}_i, \boldsymbol{B}_n\rangle)^2\right|\right)\\
&\le \sup_{\boldsymbol{B}_n \in \mathcal{B}_n^c} P_{\boldsymbol{B}_n}\left(\left|\frac{1}{n}\sum_{i=1}^{n} KL_i + \sqrt{\frac{4\sum_{i=1}^{n} KL_i}{n}}\frac{Z}{\sqrt{n}}\right| - \epsilon/4 \le \left|\frac{1}{n}\sum_{i=1}^{n}(y_i - \langle \boldsymbol{X}_i, \boldsymbol{B}_n\rangle)^2\right|\right)
\end{aligned}
$$

Let $\mathcal{T}_n = \left\{\left|\sqrt{\frac{4\sum_{i=1}^{n} KL_i}{n}}\frac{Z}{\sqrt{n}}\right| \le \frac{1}{2n}\sum_{i=1}^{n} KL_i\right\}$. Using this fact we have

$$
\begin{aligned}
&\sup_{\boldsymbol{B}_n \in \mathcal{B}_n^c} \mathbb{E}_{\boldsymbol{B}_n}(1-\Phi_n)\\
&\le \sup_{\boldsymbol{B}_n \in \mathcal{B}_n^c} P_{\boldsymbol{B}_n}\left(\left\{\left|\frac{1}{n}\sum_{i=1}^{n} KL_i + \sqrt{\frac{4\sum_{i=1}^{n} KL_i}{n}}\frac{Z}{\sqrt{n}}\right| - \epsilon/4 \le \left|\frac{1}{n}\sum_{i=1}^{n}(y_i - \langle \boldsymbol{X}_i, \boldsymbol{B}_n\rangle)^2\right|\right\} \cap \mathcal{T}_n\right) + \sup_{\boldsymbol{B}_n \in \mathcal{B}_n^c} P_{\boldsymbol{B}_n}\left(\mathcal{T}_n\right)\\
&\le \sup_{\boldsymbol{B}_n \in \mathcal{B}_n^c} P_{\boldsymbol{B}_n}\left(\frac{1}{2n}\sum_{i=1}^{n} KL_i - \epsilon/4 \le \left|\frac{1}{n}\sum_{i=1}^{n}(y_i - \langle \boldsymbol{X}_i, \boldsymbol{B}_n\rangle)^2\right|\right) + \sup_{\boldsymbol{B}_n \in \mathcal{B}_n^c} P_{\boldsymbol{B}_n}\left(\left|\frac{Z}{\sqrt{n}}\right| \ge \frac{1}{4}\sqrt{\frac{1}{n}\sum_{i=1}^{n} KL_i}\right)\\
&\le P_{\boldsymbol{B}_n}\left(\frac{3\epsilon}{4} \le \left|\frac{1}{n}\sum_{i=1}^{n}(y_i - \langle \boldsymbol{X}_i, \boldsymbol{B}_n\rangle)^2\right|\right) + P_{\boldsymbol{B}_n}\left(|Z| \ge \frac{1}{4}\sqrt{n\epsilon}\right)\\
&\le P_{\boldsymbol{B}_n}\left(\frac{3n\epsilon}{4} \le \chi_n^2\right) + P_{\boldsymbol{B}_n}\left(\chi_1^2 \ge \frac{n\epsilon}{4}\right) \le \exp\left(-\frac{3n\epsilon}{16}\right) + \exp\left(-\frac{n\epsilon}{64}\right) \le 2\exp\left(-\frac{n\epsilon}{64}\right),
\end{aligned}
$$

where the last line requires an application of Lemma 1 in Laurent and Massart (2000). ■

Theorem 1
Proof Under Lemma 6 one has

$$\Pi_n(\mathcal{B}_n^c) = \frac{\int_{\mathcal{B}_n^c} f(\boldsymbol{y}_n|\boldsymbol{B}_n)\pi_n(\boldsymbol{F}_n)}{\int f(\boldsymbol{y}_n|\boldsymbol{B}_n)\pi_n(\boldsymbol{F}_n)} = \frac{\int_{\mathcal{B}_n^c} \frac{f(\boldsymbol{y}_n|\boldsymbol{B}_n)}{f(\boldsymbol{y}_n|\boldsymbol{B}_n^0)}\pi_n(\boldsymbol{F}_n)}{\int \frac{f(\boldsymbol{y}_n|\boldsymbol{B}_n)}{f(\boldsymbol{y}_n|\boldsymbol{B}_n^0)}\pi_n(\boldsymbol{F}_n)} = \frac{N}{D} \leq \Phi_n + (1-\Phi_n)\frac{N}{D}.$$

Note that we have

$$P_{\boldsymbol{B}_n^0}\left(\Phi_n > \exp(-b_1 n/2)\right) \leq \mathbb{E}_{\boldsymbol{B}_n^0}\left(\Phi_n\right)\exp(b_1 n/2) \leq c_1\exp(-b_1 n/2).$$

Therefore $\sum_{n=1}^{\infty} P_{\boldsymbol{B}_n^0}\left(\Phi_n > \exp(-b_1 n/2)\right) < \infty$. Using Borel-Cantelli lemma $P_{\boldsymbol{B}_n^0}\left(\Phi_n > \exp(-b_1 n/2) i.o.\right) = 0$. It follows that

$$\Phi_n \to 0 \quad a.s. \tag{13}$$

In addition, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{B}_n^0}((1-\Phi_n)N) &= \int (1-\Phi_n)\int_{\mathcal{B}_n^c} \frac{f(\boldsymbol{y}_n|\boldsymbol{B}_n)}{f(\boldsymbol{y}_n|\boldsymbol{B}_n^0)}\pi_n(\boldsymbol{F}_n) f(\boldsymbol{y}_n|\boldsymbol{B}_n^0) \\
&= \int_{\mathcal{B}_n^c}\int (1-\Phi_n) f(\boldsymbol{y}_n|\boldsymbol{B}_n)\pi_n(\boldsymbol{F}_n) \\
&\leq \sup_{\boldsymbol{B}_n \in \mathcal{B}_n^c} \mathbb{E}_{\boldsymbol{B}_n}(1-\Phi_n) \leq c_2\exp(-b_2 n).
\end{aligned}$$

Using a similar technique as above, $P_{\boldsymbol{B}_n^0}\left((1-\Phi_n)N\exp(nb_2/2) > \exp(-nb_2/4) i.o.\right) = 0$ so

$$\exp(bn)(1-\Phi_n)N \to 0 \quad a.s.. \tag{14}$$

By Lemma 6 and (13)-(14) it is enough to show that $M = \exp(\tilde{b}n)\int \frac{f(\boldsymbol{y}_n|\boldsymbol{B}_n)}{f(\boldsymbol{y}_n|\boldsymbol{B}_n^0)}\pi_n(\boldsymbol{F}_n) \to \infty$ for some $\tilde{b} \leq b = \frac{\epsilon}{256}$. We choose $\tilde{b} = b$. Consider the set $\mathcal{H}_n = \left\{\boldsymbol{B}_n : \frac{1}{n}\log\left[\frac{f(\boldsymbol{y}_n|\boldsymbol{B}_n^0)}{f(\boldsymbol{y}_n|\boldsymbol{B}_n)}\right] < \eta\right\}$, for some η which is chosen later.

$$\begin{aligned}
M &\geq \exp(\tilde{b}n)\int\limits_{\mathcal{H}_n} \exp\left(-n\frac{1}{n}\log\frac{f(\boldsymbol{y}_n|\boldsymbol{B}_n^0)}{f(\boldsymbol{y}_n|\boldsymbol{B}_n)}\right)\pi_n(\boldsymbol{F}_n) \\
&\geq \exp((\tilde{b}-\eta)n)\pi_n(\mathcal{H}_n).
\end{aligned}$$

Note that

$$\begin{aligned}
&\frac{1}{n}\log\left[\frac{f(\boldsymbol{y}_n|\boldsymbol{B}_n)}{f(\boldsymbol{y}_n|\boldsymbol{B}_n^0)}\right] \\
&= \frac{1}{n}\left[-\frac{1}{2}\sum_{i=1}^{n}(y_i - \langle \boldsymbol{X}_i, \boldsymbol{B}_n\rangle)^2 + \frac{1}{2}\sum_{i=1}^{n}(y_i - \langle \boldsymbol{X}_i, \boldsymbol{B}_n^0\rangle)^2\right].
\end{aligned}$$

Let $\boldsymbol{y}_n = (y_1, ..., y_n)'$, $\boldsymbol{H}_n = (\langle \boldsymbol{X}_1, \boldsymbol{B}_n\rangle, \dots, \langle \boldsymbol{X}_n, \boldsymbol{B}_n\rangle)$ and $\boldsymbol{H}_n^0 = \big(\langle \boldsymbol{X}_1, \boldsymbol{B}_n^0\rangle, \dots, \langle \boldsymbol{X}_n, \boldsymbol{B}_n^0\rangle\big)$. Then

$$\begin{aligned}
&\pi_n\Big(\boldsymbol{B}_n : \frac{1}{n}\left[-||\boldsymbol{y}_n - \boldsymbol{H}_n^0||^2 + ||\boldsymbol{y}_n - \boldsymbol{H}_n||^2\right] < 2\eta\Big)\\
&\geq \pi_n\Big(\boldsymbol{B}_n : \frac{1}{n}\left|2||\boldsymbol{y}_n - \boldsymbol{H}_n^0||\left(||\boldsymbol{y}_n - \boldsymbol{H}_n|| - ||\boldsymbol{y}_n - \boldsymbol{H}_n^0||\right) + \left(||\boldsymbol{y}_n - \boldsymbol{H}_n|| - ||\boldsymbol{y}_n - \boldsymbol{H}_n^0||\right)^2\right| < 2\eta\Big)\\
&\geq \pi_n\Big(\boldsymbol{B}_n : \frac{1}{n}\left|2||\boldsymbol{y}_n - \boldsymbol{H}_n^0||||\boldsymbol{H}_n^0 - \boldsymbol{H}_n|| + ||\boldsymbol{H}_n - \boldsymbol{H}_n^0||^2\right| < 2\eta\Big)\\
&\geq \pi_n\Big(\boldsymbol{B}_n : \frac{1}{n}||\boldsymbol{H}_n^0 - \boldsymbol{H}_n|| < \frac{2\eta}{3\zeta_n}, ||\boldsymbol{y}_n - \boldsymbol{H}_n^0||^2 < \zeta_n^2\Big)\\
&\geq \pi_n(\mathcal{A}_{1n} \cap \mathcal{A}_{2n})
\end{aligned}$$

where $\mathcal{A}_{1n} = \left\{\frac{1}{n}||\boldsymbol{H}_n - \boldsymbol{H}_n^0|| < \frac{2\eta}{3\zeta_n}\right\}$, $\mathcal{A}_{2n} = \left\{||\boldsymbol{y}_n - \boldsymbol{H}_n^0||^2 < \zeta_n^2\right\}$.

We will show that $P_{\boldsymbol{B}_n^0}(\mathcal{A}_{2n}) = 1$ for all large n. Assume $\zeta_n = n^{(1+\rho_3)/2}$, $\rho_3 > 0$ so that $\zeta_n^2 > 8n$ for all large n. Then,

$$P_{\boldsymbol{B}_n^0}(\mathcal{A}_{2n}') = P_{\boldsymbol{B}_n^0}(\chi_n^2 > \zeta_n^2) \leq \exp(-\zeta_n^2/2).$$

Therefore, using Borel-Cantelli lemma $P_{\boldsymbol{B}_n^0}(\mathcal{A}_{2n}'\ i.o.) = 0$. Hence $P_{\boldsymbol{B}_n^0}(\mathcal{A}_{2n}) = 1$ for all large n. It is enough to bound $\pi_n(\mathcal{A}_{1n})$. Let $M_n = \frac{1}{n}\sqrt{\sum_{i=1}^n ||\boldsymbol{X}_i||_2^2}$. Now use the fact that $\frac{1}{n}||\boldsymbol{H}_n - \boldsymbol{H}_n^0|| = \frac{1}{n}\sqrt{\sum_{i=1}^n(\langle \boldsymbol{X}_i, \boldsymbol{B}_n - \boldsymbol{B}_n^0\rangle)^2} \leq \left(\frac{1}{n}\sqrt{\sum_{i=1}^n ||\boldsymbol{X}_i||_2^2}\right)||\boldsymbol{B}_n - \boldsymbol{B}_n^0||_2$ to conclude

$$\left\{||\boldsymbol{B}_n - \boldsymbol{B}_n^0||_2 < \frac{2\eta}{3M_n\zeta_n}\right\} \subseteq \mathcal{A}_{1n}. \tag{15}$$

By (11) one has $\pi_n(\mathcal{A}_{1n}) \geq \pi_n\Big(||\boldsymbol{B}_n - \boldsymbol{B}_n^0||_2 < \frac{2\eta}{3M_n\zeta_n}\Big) \geq \exp(-dn)$ and hence $M \geq \exp\big((\tilde{b} - \eta - d)n\big) \to \infty$ as $n \to \infty$ proving the result. ■

Theorem 2
Proof Define $g : \mathbb{R} \to \mathbb{R}$ s.t.

$$g(\kappa) = R\kappa^D + \kappa^{D-1}\sum_{j=1}^D\sum_{r=1}^R ||\boldsymbol{\beta}_{j,n}^{0(r)}||_2 + \cdots + \kappa\sum_{j=1}^D\sum_{r=1}^R\prod_{l\neq j}||\boldsymbol{\beta}_{l,n}^{0(r)}||_2.$$

Let $\kappa_n > 0$ be s.t. $g(\kappa_n) = \frac{2\eta}{3M_n\zeta_n}$. Note that by Decarte's rule of sign, the equation $g(\kappa) - \frac{2\eta}{3M_n\zeta_n} = 0$ has a unique positive root. Further

$$\frac{1}{\kappa_n} < 1 + \max_{i=1,\dots,D}\left|\frac{3\sum_{j_1\neq\cdots\neq j_i}\sum_{r=1}^R\prod_{l=1}^{i}||\boldsymbol{\beta}_{j_l,n}^{0(r)}||_2}{2\eta/M_n\zeta_n}\right| \tag{16}$$

$$\kappa_n < 1 + \max\left\{\frac{2\eta}{3M_n\zeta_n R}, \max_{i=1,\dots,D}\left|\frac{\sum_{j_1\neq\cdots\neq j_i}\sum_{r=1}^R\prod_{l=1}^{i}||\boldsymbol{\beta}_{j_l,n}^{0(r)}||_2}{R}\right|\right\} \tag{17}$$

by Lemma 8 in Appendix B.

Using Lemma 7 in Appendix B, it is easy to see that

$$\left\{||\boldsymbol{\beta}_{j,n}^{(r)} - \boldsymbol{\beta}_{j,n}^{0(r)}||_2 \leq \kappa_n,\ j = 1, ..., D;\ r = 1, ..., R\right\} \subseteq \left\{||\boldsymbol{B}_n - \boldsymbol{B}_n^0||_2 < \frac{2\eta}{3M_n\zeta_n}\right\}. \tag{18}$$

Using (15), $\pi_n\left(||\boldsymbol{B}_n - \boldsymbol{B}_n^0||_2 < \frac{2\eta}{3M_n\zeta_n}\right) \geq \pi_n\left(\left\{||\boldsymbol{\beta}_{j,n}^{(r)} - \boldsymbol{\beta}_{j,n}^{0(r)}||_2 \leq \kappa_n,\ j = 1, ..., D;\ r = 1, ..., R\right\}\right)$. Note that

$$\begin{aligned}
&\pi_n\left(\left\{||\boldsymbol{\beta}_{j,n}^{(r)} - \boldsymbol{\beta}_{j,n}^{0(r)}||_2 \leq \kappa_n,\ j = 1, ..., D;\ r = 1, ..., R\right\} |\{w_{jr,l}\}_{l=1}^{p_{j,n}}, \{\lambda_{jr}\}_{j,r=1}^{D,R-1}, \{\phi_r\}_{r=1}^{R-1}, \tau\right) \\
&\geq \left[\prod_{j=1}^{D}\prod_{r=1}^{R} \pi_n\left(||\boldsymbol{\beta}_{j,n}^{(r)} - \boldsymbol{\beta}_{j,n}^{0(r)}||_2 \leq \kappa_n|\{w_{jr,l}\}_{l=1}^{p_{j,n}}, \{\lambda_{jr}\}_{j,r=1}^{D,R-1}, \{\phi_r\}_{r=1}^{R-1}, \tau\right)\right].
\end{aligned}$$

Therefore, it is enough to bound $\pi_n(||\boldsymbol{\beta}_{j,n}^{(r)} - \boldsymbol{\beta}_{j,n}^{0(r)}|| \leq \kappa_n, j = 1, ..., D; r = 1, ..., R)$. For $j = 1, ..., D$, $r = 1, ..., R$,

$$\begin{aligned}
&\pi_n(||\boldsymbol{\beta}_{j,n}^{(r)} - \boldsymbol{\beta}_{j,n}^{0(r)}|| \leq \kappa_n|\{w_{jr,l}\}_{l=1}^{p_{j,n}}, \lambda_{jr}, \{\phi_r\}_{r=1}^{R-1}, \tau) \\
&\quad \geq \prod_{l=1}^{p_{j,n}} \pi_n\left(|\beta_{j,n,l}^{(r)} - \beta_{j,n,l}^{0(r)}| \leq \frac{\kappa_n}{\sqrt{p_{j,n}}}|\{w_{jr,l}\}_{l=1}^{p_{j,n}}, \lambda_{jr}, \{\phi_r\}_{r=1}^{R-1}, \tau\right) \\
&\quad \geq \prod_{l=1}^{p_{j,n}} \left\{\left(\frac{2\kappa_n}{\sqrt{2p_{j,n}\pi w_{jr,l}\phi_r\tau}}\right)\exp\left(-\frac{|\beta_{j,n,l}^{0(r)}|^2 + \kappa_n^2/p_{j,n}}{w_{jr,l}\phi_r\tau}\right)\right\},
\end{aligned}$$

where the last step follows from the fact that $\int_a^b e^{-x^2/2}dx \geq e^{-(a^2+b^2)/2}(b-a)$. Thus,

$$\begin{aligned}
&\pi_n(||\boldsymbol{\beta}_{j,n}^{(r)} - \boldsymbol{\beta}_{j,n}^{0(r)}|| \leq \kappa_n|\lambda_{jr}, \{\phi_r\}_{r=1}^{R-1}, \tau) \\
&= \mathbb{E}\left[\pi_n(||\boldsymbol{\beta}_{j,n}^{(r)} - \boldsymbol{\beta}_{j,n}^{0(r)}|| \leq \kappa_n|\{w_{jr,l}\}_{l=1}^{p_{j,n}}, \lambda_{jr}, \{\phi_r\}_{r=1}^{R-1}, \tau)\right] \\
&\geq \left(\frac{2\kappa_n}{\sqrt{2p_{j,n}\pi\phi_r\tau}}\right)^{p_{j,n}} \prod_{l=1}^{p_{j,n}} E\left\{\frac{1}{\sqrt{w_{jr,l}}}\exp\left(-\frac{|\beta_{j,n,l}^{0(r)}|^2 + \kappa_n^2/p_{j,n}}{w_{jr,l}\phi_r\tau}\right)\right\} \\
&\geq \left(\frac{2\kappa_n\lambda_{jr}^2}{2\sqrt{2p_{j,n}\pi\phi_r\tau}}\right)^{p_{j,n}} \prod_{l=1}^{p_{j,n}} \int_{w_{jr,l}} \left\{\frac{1}{\sqrt{w_{jr,l}}}\exp\left(-\frac{|\beta_{j,n,l}^{0(r)}|^2 + \kappa_n^2/p_{j,n}}{w_{jr,l}\phi_r\tau} - \frac{\lambda_{jr}^2 w_{jr,l}}{2}\right)dw_{jr,l}\right\}.
\end{aligned} \tag{19}$$

Use the change of variable $\frac{1}{w_{jr,l}} = z_{jr,l}$ and the normalizing constant from the inverse Gaussian density to deduce

$$\begin{aligned}
&\int_{w_{jr,l}} \left\{ \frac{1}{\sqrt{w_{jr,l}}} \exp\left(-\frac{|\beta^{0(r)}_{j,n,l}|^2 + \kappa_n^2/p_{j,n}}{w_{jr,l}\phi_r\tau} - \frac{\lambda_{jr}^2 w_{jr,l}}{2} \right) dw_{jr,l} \right\} \\
&\quad = \int_{z_{jr,l}} \left\{ \frac{1}{\sqrt{z^3_{jr,l}}} \exp\left(-\frac{(|\beta^{0(r)}_{j,n,l}|^2 + \kappa_n^2/p_{j,n})}{\phi_r\tau} z_{jr,l} - \frac{\lambda_{jr}^2}{2z_{jr,l}} \right) dz_{jr,l} \right\} \\
&\quad = \sqrt{\left(\frac{2\pi}{\lambda^2_{jr}}\right)} \exp\left(-\lambda_{jr} \frac{\sqrt{2\left(|\beta^{0(r)}_{j,n,l}|^2 + \kappa_n^2/p_{j,n}\right)}}{\sqrt{\phi_r\tau}} \right).
\end{aligned}$$

(19) can be written as

$$\begin{aligned}
&\pi_n(||\boldsymbol{\beta}^{(r)}_{j,n} - \boldsymbol{\beta}^{0(r)}_{j,n}|| \leq \kappa_n | \lambda_{jr}, \{\phi_r\}_{r=1}^{R-1}, \tau) \\
&\geq \left(\frac{2\kappa_n\lambda^2_{jr}}{2\sqrt{2p_{j,n}\pi\phi_r\tau}} \right)^{p_{j,n}} \prod_{l=1}^{p_{j,n}} \left[\sqrt{\left(\frac{2\pi}{\lambda^2_{jr}}\right)} \exp\left(-\lambda_{jr} \frac{\sqrt{2\left(|\beta^{0(r)}_{j,n,l}|^2 + \kappa_n^2/p_{j,n}\right)}}{\sqrt{\phi_r\tau}} \right) \right] \\
&= \left(\frac{2\kappa_n\lambda_{jr}}{2\sqrt{p_{j,n}\phi_r\tau}} \right)^{p_{j,n}} \exp\left(-\lambda_{jr} \frac{\sum_{l=1}^{p_{j,n}} \sqrt{2\left(|\beta^{0(r)}_{j,n,l}|^2 + \kappa_n^2/p_{j,n}\right)}}{\sqrt{\phi_r\tau}} \right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\pi_n(||\boldsymbol{\beta}^{(r)}_{j,n} - \boldsymbol{\beta}^{0(r)}_{j,n}|| \leq \kappa_n | \{\phi_r\}_{r=1}^{R-1}, \tau) \\
&\geq \left(\frac{2\kappa_n}{2\sqrt{p_{j,n}\phi_r\tau}} \right)^{p_{j,n}} \frac{b_{\lambda,r}^{a_{\lambda,r}}}{\Gamma(a_{\lambda,r})} \int_{\lambda_{jr}} \lambda_{jr}^{p_{j,n}+a_{\lambda,r}-1} \exp\left(-\lambda_{jr} \left[\frac{\sum_{l=1}^{p_{j,n}} \sqrt{2\left(|\beta^{0(r)}_{j,n,l}|^2 + \kappa_n^2/p_{j,n}\right)}}{\sqrt{\phi_r\tau}} + b_{\lambda,r} \right] \right) d\lambda_{jr} \\
&= \left(\frac{2\kappa_n}{2\sqrt{p_{j,n}\phi_r\tau}} \right)^{p_{j,n}} \frac{b_{\lambda,r}^{a_{\lambda,r}}}{\Gamma(a_{\lambda,r})} \frac{\Gamma(p_{j,n} + a_{\lambda,r})}{\left[\frac{\sum_{l=1}^{p_{j,n}} \sqrt{2\left(|\beta^{0(r)}_{j,n,l}|^2 + \kappa_n^2/p_{j,n}\right)}}{\sqrt{\phi_r\tau}} + b_{\lambda,r} \right]^{p_{j,n}+a_{\lambda,r}}} \\
&= \left(\frac{2\kappa_n}{2b_{\lambda,r}\sqrt{p_{j,n}\phi_r\tau}} \right)^{p_{j,n}} \frac{1}{\Gamma(a_{\lambda,r})} \frac{\Gamma(p_{j,n} + a_{\lambda,r})}{\left[\frac{\sum_{l=1}^{p_{j,n}} \sqrt{2\left(|\beta^{0(r)}_{j,n,l}|^2 + \kappa_n^2/p_{j,n}\right)}}{b_{\lambda,r}\sqrt{\phi_r\tau}} + 1 \right]^{p_{j,n}+a_{\lambda,r}}}.
\end{aligned}$$

The final expression as in the above yields

$$\pi_n(||\boldsymbol{\beta}_{j,n}^{(r)}-\boldsymbol{\beta}_{j,n}^{0(r)}||\leq\kappa_n, j=1,...,D, r=1,...,R|\{\phi_r\}_{r=1}^{R-1},\tau)$$
$$\geq\mathbb{E}\left\{\prod_{j=1}^{D}\prod_{r=1}^{R}\left[\left(\frac{2\kappa_n}{2b_{\lambda,r}\sqrt{p_{j,n}\phi_r\tau}}\right)^{p_{j,n}}\frac{1}{\Gamma(a_{\lambda,r})}\lambda_{j,r}^{p_{j,n}+a_{\lambda,r}-1}\frac{\Gamma(p_{j,n}+a_{\lambda,r})}{\left[\frac{\sum_{l=1}^{p_{j,n}}\sqrt{2\left(|\beta_{j,n,l}^{0(r)}|^2+\kappa_n^2/p_{j,n}\right)}}{b_{\lambda,r}\sqrt{\phi_r\tau}}+1\right]^{p_{j,n}+a_{\lambda,r}}}\right]\right\}.$$

We will now use the fact that for $\phi_r\leq 1$,

$$\frac{1}{\left[\frac{\sum_{l=1}^{p_{j,n}}\sqrt{2\left(|\beta_{j,n,l}^{0(r)}|^2+\kappa_n^2/p_{j,n}\right)}}{b_{\lambda,r}\sqrt{\phi_r\tau}}+1\right]^{p_{j,n}+a_{\lambda,r}}}\geq\frac{1}{\left[\frac{\sum_{l=1}^{p_{j,n}}\sqrt{2\left(|\beta_{j,n,l}^{0(r)}|^2+\kappa_n^2/p_{j,n}\right)}}{b_{\lambda,r}\sqrt{\phi_r\tau}}+\frac{1}{\sqrt{\tau\phi_r}}\right]^{p_{j,n}+a_{\lambda,r}}}I_{\tau\in[0,1]}.$$

This inequality is critical to provide a lower bound on $\pi_n(||\boldsymbol{\beta}_{j,n}^{(r)}-\boldsymbol{\beta}_{j,n}^{0(r)}||\leq\kappa_n, j=1,...,D, r=1,...,R)$ as following

$$\pi_n(||\boldsymbol{\beta}_{j,n}^{(r)}-\boldsymbol{\beta}_{j,n}^{0(r)}||\leq\kappa_n, j=1,...,D, r=1,...,R)$$
$$\geq\frac{\lambda_2^{\lambda_1}\Gamma(Ra)}{\Gamma(\lambda_1)\Gamma(a)^R}\prod_{j=1}^{D}\prod_{r=1}^{R}\left[\left(\frac{\kappa_n}{\sqrt{p_{j,n}}b_{\lambda,r}}\right)^{p_{j,n}}\frac{\Gamma(p_{j,n}+a_{\lambda,r})}{\Gamma(a_{\lambda,r})}\right]\int_{\tau}\tau^{\lambda_1-R\sum_{j=1}^{D}\frac{p_{j,n}}{2}-1}\exp(-\lambda_2\tau)$$
$$\int_{\boldsymbol{\phi}\in\mathcal{S}^{R-1}}\frac{\prod_{r=1}^{R}\phi_r^{a-1}}{\prod_{r=1}^{R}\phi_r^{\sum_{j=1}^{D}\frac{p_{j,n}}{2}}}\prod_{j=1}^{D}\prod_{r=1}^{R}\frac{1}{\left[\frac{\sum_{l=1}^{p_{j,n}}\sqrt{2\left(|\beta_{j,n,l}^{0(r)}|^2+\kappa_n^2/p_{j,n}\right)}}{b_{\lambda,r}\sqrt{\phi_r\tau}}+1\right]^{p_{j,n}+a_{\lambda,r}}}d\boldsymbol{\phi}d\tau$$
$$\geq\frac{\lambda_2^{\lambda_1}\Gamma(Ra)}{\Gamma(\lambda_1)\Gamma(a)^R}\prod_{j=1}^{D}\prod_{r=1}^{R}\left[\left(\frac{\kappa_n}{\sqrt{p_{j,n}}b_{\lambda,r}}\right)^{p_{j,n}}\frac{\Gamma(p_{j,n}+a_{\lambda,r})}{\Gamma(a_{\lambda,r})}\right]\prod_{j=1}^{D}\prod_{r=1}^{R}\frac{1}{\left[\frac{\sum_{l=1}^{p_{j,n}}\sqrt{2\left(|\beta_{j,n,l}^{0(r)}|^2+\kappa_n^2/p_{j,n}\right)}}{b_{\lambda,r}}+1\right]^{p_{j,n}+a_{\lambda,r}}}$$
$$\left(\int_{\tau=0}^{1}\tau^{\lambda_1+\sum_{r=1}^{R}a_{\lambda,r}\frac{D}{2}-1}\exp(-\tau\lambda_2)d\tau\right)\int_{\boldsymbol{\phi}\in\mathcal{S}^{R-1}}\prod_{r=1}^{R}\phi_r^{a+a_{\lambda,r}\frac{D}{2}-1}d\boldsymbol{\phi}$$
$$=\frac{\lambda_2^{\lambda_1}\Gamma(Ra)}{\Gamma(\lambda_1)\Gamma(a)^R}\prod_{j=1}^{D}\prod_{r=1}^{R}\left[\left(\frac{\kappa_n}{\sqrt{p_{j,n}}b_{\lambda,r}}\right)^{p_{j,n}}\frac{\Gamma(p_{j,n}+a_{\lambda,r})}{\Gamma(a_{\lambda,r})}\right]\prod_{j=1}^{D}\prod_{r=1}^{R}\frac{1}{\left[\frac{\sum_{l=1}^{p_{j,n}}\sqrt{2\left(|\beta_{j,n,l}^{0(r)}|^2+\kappa_n^2/p_{j,n}\right)}}{b_{\lambda,r}}+1\right]^{p_{j,n}+a_{\lambda,r}}}$$
$$\times\frac{\exp(-\lambda_2)}{(\lambda_1+\sum_{r=1}^{R}a_{\lambda,r}\frac{D}{2})}\frac{\prod_{r=1}^{R}\left[\Gamma(a+a_{\lambda,r}\frac{D}{2})\right]}{\Gamma(Ra+\frac{D}{2}\sum_{r=1}^{R}a_{\lambda,r})}.$$

Denote $C_6 = \frac{\lambda_2^{\lambda_1}\Gamma(Ra)}{\Gamma(\lambda_1)[\Gamma(a)]^R}\frac{\exp(-\lambda_2)}{\left(\lambda_1+\sum_{r=1}^R a_{\lambda,r}\frac{D}{2}\right)}\frac{\prod_{r=1}^R\left[\Gamma(a+a_{\lambda,r}\frac{D}{2})\right]}{\Gamma(Ra+\sum_{r=1}^R a_{\lambda,r}\frac{D}{2})}$. Then the above expression gives us

$$\begin{aligned}
&-\log\left(||\boldsymbol{B}_n-\boldsymbol{B}_n^0||_2<\frac{2\eta}{3M_n\zeta_n}\right)\\
&\le -\log(C_6)+\sum_{j=1}^D\sum_{r=1}^R p_{j,n}\left[-\log(\kappa_n)+\frac{1}{2}\log(p_{j,n})+\log(b_{\lambda,r})+\log(\Gamma(a_{\lambda,r})\right]\\
&\quad-\sum_{r=1}^R\sum_{j=1}^D\log(\Gamma(p_{n,j}+a_{\lambda,r}))+\sum_{j=1}^D\sum_{r=1}^R(p_{j,n}+a_{\lambda,r})\log\left[\frac{\sum_{l=1}^{p_{j,n}}\sqrt{2\left(|\beta_{j,n,l}^{0(r)}|^2+\kappa_n^2/p_{j,n}\right)}}{b_{\lambda,r}}+1\right].
\end{aligned} \tag{20}$$

Using (16) and assumption (b), it is easy to see that $\frac{1}{\kappa_n}<G_5n^{\rho_2+\frac{\rho_3+1}{2}}\prod_{j=1}^D p_{j,n}$ for a constant $G_5>0$ for all large n. Therefore, $\sum_{j=1}^D\sum_{r=1}^R p_{j,n}\left[\log\left(\frac{1}{\kappa_n}\right)+\frac{1}{2}\log(p_{j,n})+\log(b_{\lambda,r})+\log(\Gamma(a_{\lambda,r})\right]=o(n)$. Also, $\sum_{r=1}^R\sum_{j=1}^D\log(\Gamma(p_{j,n}+a_{\lambda,r}))]\le\sum_{j=1}^D(p_{j,n}+a_{\lambda,r})\log(p_{j,n}+a_{\lambda,r})=o(n)$, by assumption (c). Finally, $\sum_{j=1}^D\sum_{r=1}^R(p_{j,n}+a_{\lambda,r})\log\left[\frac{\sum_{l=1}^{p_{j,n}}\sqrt{2\left(|\beta_{j,n,l}^{0(r)}|^2+\kappa_n^2/p_{j,n}\right)}}{b_{\lambda,r}}+1\right]=o(n)$, by assumptions (b) and (c). Thus, $-\log\left(\pi_n(\boldsymbol{B}_n:||\boldsymbol{B}_n-\boldsymbol{B}_n^0||_2<\frac{2\eta}{3M_n\zeta_n})\right)<dn$ for all $d>0$, for all large n. This proves the result. ■

Appendix B

This Section contains additional Lemmas relevant to the article.

Lemma 7 *Suppose $\boldsymbol{T}=T_1\circ\cdots\circ T_D$ and $\boldsymbol{F}=F_1\circ\cdots\circ F_D$ are two rank one tensors of same dimension. Then*

$$\boldsymbol{T}-\boldsymbol{F}=(T_1-F_1)\circ\cdots\circ(T_D-F_D)+\sum_{l=1}^{D-1}\sum_{\mathcal{I}_1\cup\mathcal{I}_2=1:D,|\mathcal{I}_1|=l,|\mathcal{I}_2|=D-l}\gamma_1\circ\cdots\circ\gamma_D,$$

where $\gamma_j=F_j$ if $j\in\mathcal{I}_2$; $=T_j-F_j$ if $j\in\mathcal{I}_1$.

Proof We will show it by induction. If $D=2$ then,

$$\begin{aligned}
\boldsymbol{T}-\boldsymbol{F}&=T_1\circ T_2-F_1\circ F_2=(T_1-F_1)\circ T_2+F_1\circ T_2-F_1\circ F_2\\
&=(T_1-F_1)\circ(T_2-F_2)+(T_1-F_1)\circ F_2+F_1\circ(T_2-F_2).
\end{aligned}$$

Assume the result to hold for $D-1$. For D,

$$
\begin{aligned}
&T_1 \circ \cdots \circ T_D - F_1 \circ \cdots \circ F_D \\
&= (T_1 - F_1) \circ T_2 \circ \cdots \circ T_D + F_1 \circ [T_2 \circ \cdots \circ T_D - F_2 \circ \cdots \circ F_D] \\
&= (T_1 - F_1) \circ [(T_2 - F_2) \circ \cdots \circ (T_D - F_D) + F_2 \circ \cdots \circ F_D + \\
&\sum_{l=1}^{D-2} \sum_{\mathcal{I}_1 \cup \mathcal{I}_2, |\mathcal{I}_1| = l, |\mathcal{I}_2| = D-1-l} \gamma_2 \circ \cdots \circ \gamma_D] + \\
&F_1 \circ [(T_2 - F_2) \circ \cdots \circ (T_D - F_D) + \sum_{l=1}^{D-2} \sum_{\mathcal{I}_1 \cup \mathcal{I}_2, |\mathcal{I}_1| = l, |\mathcal{I}_2| = D-1-l} \gamma_2 \circ \cdots \circ \gamma_D] \\
&= (T_1 - F_1) \circ \cdots \circ (T_D - F_D) + \sum_{l=1}^{D-1} \sum_{\mathcal{I}_1 \cup \mathcal{I}_2, |\mathcal{I}_1| = l, |\mathcal{I}_2| = D-l} \gamma_1 \circ \cdots \circ \gamma_D].
\end{aligned}
$$

Hence proved. ■

Lemma 8 *Let x^* be a real root of the polynomial $P(x) = a_k x^k + a_{k-1} x^{k-1} + \cdots + a_1 x - a_0$. Then $1/|x^*| < 1 + \max_{i=1,\ldots,k} \left| \frac{a_i}{a_0} \right|$.*

Proof Consider the polynomial $P_1(\zeta) = \zeta^k - \left(\frac{a_1}{a_0}\right) \zeta^{k-1} - \cdots - \left(\frac{a_k}{a_0}\right)$. By making a change of variable with $\zeta = \frac{1}{x}$, we obtain

$$
\begin{aligned}
P_1\left(\frac{1}{x}\right) &= \frac{1}{x^k} - \left(\frac{a_1}{a_0}\right) \frac{1}{x^{k-1}} - \cdots - \left(\frac{a_k}{a_0}\right) \\
&= -\frac{a_k x^k + \cdots + a_1 x - a_0}{a_0 x^k}.
\end{aligned}
$$

Note that $P_1\left(\frac{1}{x}\right) = 0$ is solved by $x = x^*$. Therefore, $P_1(\zeta) = 0$ is solved by $\zeta = \frac{1}{x^*}$. The result follows by using Cauchy bound on the roots of a polynomial. ■

Lemma 9 *Suppose $T_1, \ldots, T_m$ are independent random variables with T_j having density f_j supported in $(0, \infty)$. Let $\phi_j = \frac{T_j}{\sum_{l=1}^m T_m}$. Then the joint density of $(\phi_1, \ldots, \phi_{m-1})$ has a joint density supported on the simplex $\mathcal{S}^{m-1}$ and is given by*

$$
f(\phi_1, \ldots, \phi_{m-1}) = \int_{t=0}^{\infty} t^{m-1} \prod_{l=1}^{m} f_j(\phi_j t) dt,
$$

where $\phi_m = 1 - \sum_{l=1}^{m-1} \phi_l$.

Proof This result is well known in the theory of normalized random measures (Kruijer et al., 2010). ■

References

Artin Armagan, David B Dunson, and Jaeyong Lee. Generalized double pareto shrinkage. *Statistica Sinica*, 23(1):119–143, 2013a.

Artin Armagan, David B Dunson, Jaeyong Lee, Waheed U Bajwa, and Nate Strawn. Posterior consistency in linear models under shrinkage priors. *Biometrika*, 100(4):1011–1018, 2013b.

Anirban Bhattacharya, Debdeep Pati, Natesh S Pillai, and David B Dunson. Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110 (512):1479–1490, 2015.

Brian S Caffo, Ciprian M Crainiceanu, Guillermo Verduzco, Suresh Joel, Stewart H Mostofsky, Susan Spear Bassett, and James J Pekar. Two-stage decompositions for the analysis of functional connectivity for fmri with application to alzheimer's disease risk. *NeuroImage*, 51(3):1140–1149, 2010.

David Gerard and Peter Hoff. Adaptive higher-order spectral estimators. *arXiv preprint arXiv:1505.02114*, 2015.

Jeff Goldsmith, Lei Huang, and Ciprian M Crainiceanu. Smooth scalar-on-image regression via spatial Bayesian variable selection. *Journal of Computational and Graphical Statistics*, 23(1):46–64, 2014.

Chris Hinrichs, Vikas Singh, Lopamudra Mukherjee, Guofan Xu, Moo K Chung, Sterling C Johnson, and Alzheimer's Disease Neuroimaging Initiative. Spatially augmented lpboosting for ad classification with evaluations on the adni dataset. *Neuroimage*, 48(1):138–149, 2009.

Peter D Hoff et al. Multilinear tensor regression for longitudinal relational data. *The Annals of Applied Statistics*, 9(3):1169–1193, 2015.

Hung Hung and Chen-Chien Wang. Matrix variate logistic regression model with application to eeg data. *Biostatistics*, 14(1):189–202, 2013.

Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

Willem Kruijer, Judith Rousseau, and Aad Van Der Vaart. Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257, 2010.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Nicole Lazar. *The Statistical Analysis of Functional MRI Data*. Springer Science & Business Media, 2008.

Martin A Lindquist. The statistical analysis of fmri data. *Statistical Science*, 23(4):439–464, 2008.

Nicholas G Polson and James G Scott. Local shrinkage rules, lévy processes and regularized regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):287–311, 2012.

Peihua Qiu. Jump surface estimation, edge detection, and image restoration. *Journal of the American Statistical Association*, 102(478):745–756, 2007.

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

Srikanth Ryali, Kaustubh Supekar, Daniel A Abrams, and Vinod Menon. Sparse logistic regression for whole-brain classification of fmri data. *NeuroImage*, 51(2):752–764, 2010.

Taiji Suzuki. Convergence rate of Bayesian tensor estimatior and its minimax optimality. In *Proceedings of the 32nd International Conference on Machine Learning (Lille, 2015)*, pages 1273–1282, 2015.

Xuejing Wang, Bin Nan, Ji Zhu, and Robert Koeppe. Regularized 3d functional regression for brain image data via Haar wavelets. *The annals of applied statistics*, 8(2):1045, 2014.

Yun Yang and David B Dunson. Minimax optimal Bayesian aggregation. *arXiv preprint arXiv:1403.1345*, 2014.

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

Jing Zhou, Anirban Bhattacharya, Amy H Herring, and David B Dunson. Bayesian factorizations of big sparse tensors. *Journal of the American Statistical Association*, 110(512): 1562–1576, 2015.