Bayesian Generalized Sparse Symmetric Tensor-on-Vector Regression

Sharmistha Guha[‡] Rajarshi Guhaniyogi[§]

May 29, 2020

Abstract

Motivated by brain connectome datasets acquired using diffusion weighted magnetic resonance imaging (DWI), this article proposes a novel generalized Bayesian linear modeling framework with a symmetric tensor response and scalar predictors. The symmetric tensor coefficients corresponding to the scalar predictors are embedded with two features: low-rankness and group sparsity within the low-rank structure. Besides offering computational efficiency and parsimony, these two features enable identification of important "tensor nodes" and "tensor cells" significantly associated with the predictors, with characterization of uncertainty. The proposed framework is empirically investigated under various simulation settings and with a real brain connectome dataset. Theoretically, we establish that the posterior predictive density from the proposed model is "close" to the true data generating density, the closeness being measured by the Hellinger distance between these two densities, which scales at a rate very close to the finite dimensional optimal rate of $n^{-1/2}$, depending on how the number of tensor

[‡]Sharmistha Guha, Postdoctoral Associate, Department of Statistical Science, Duke University, 206 Old Chemistry Building (E-mail: sharmistha.guha@duke.edu).

[§]Rajarshi Guhaniyogi, Assistant Professor, Department of Statistics, SOE2, UC Santa Cruz, 1156 High Street, Santa Cruz, CA 95064 (E-mail: rguhaniy@ucsc.edu).

nodes grow with the sample size. The theoretical results with proofs are provided in the supplementary material which is available online.

Key Words: Brain connectome, Binary network, Low rank decomposition, Posterior convergence, Symmetric tensor, Spike and slab prior.

1 Introduction

In recent times, multidimensional arrays or tensors, which are higher order extensions of two dimensional matrices, are being encountered in datasets emerging from different disciplines. Similar to the rows and columns of a matrix, the dimensions or axes of a tensor are known as tensor *modes*. A tensor is known to be symmetric if interchanging the modes results in the same tensor. Therefore, a symmetric matrix is a special case of a symmetric tensor in two dimensions. The indices of any tensor mode in a symmetric tensor are often referred to as *tensor nodes*. This article is motivated by a variety of brain related data applications, where comprehensive maps of neural connections in the brain, also known as *brain connectomes*, are available for multiple subjects. These brain connectomes are often expressed in the form of symmetric tensors. Our focus is mainly on datasets in which a sample of symmetric tensors is available from multiple subjects of interest, along with a few subject specific observable attributes, often referred to as *phenotypes*. In these applications, there is a symmetric tensor corresponding to every subject, and the tensor nodes are labeled and shared across all subject specific tensors through a common map.

To provide a concrete example, we consider a dataset that contains brain network information along with a measure of creativity (e.g., composite creativity index (CCI)), age and sex for multiple subjects. Brain network information for each subject is encoded within a symmetric matrix of dimension 68×68, with the (k, l)th cell consisting of the number of neuron connections between the k-th and l-th regions of interest (ROI). Each mode of this symmetric matrix (when viewed as a 2-D symmetric tensor) consists of 68 nodes, each corre-

sponding to a specific ROI in the human brain. The goal of such scientific studies is two fold. First, it is important to build a predictive model to assess changes in the symmetric tensor (e.g., the brain network) as the phenotypic predictors vary. Second, it is of interest to identify nodes and cells in the symmetric tensor significantly impacted by CCI. In the context of brain connectome applications, this goal boils down to making inference on brain regions of interest (ROIs) and their inter-connections significantly associated with CCI. Although this specific example involves a symmetric tensor response of order 2, there are other examples from fields such as international trade, involving higher order symmetric tensor response on vector regressions. We discuss one such example pertinent to our modeling framework in Section 6.

In developing a modeling approach to our problem of interest, one can possibly proceed to vectorize the symmetric tensor response and regress it on the predictors, leading to a high dimensional vector regression problem. This approach can take advantage of the recent developments in high dimensional multivariate reduced rank sparse regression literature, consisting of both penalized optimization (Yuan *et al.*, 2007; Rothman *et al.*, 2010) and Bayesian shrinkage (Goh *et al.*, 2017). However, this approach treats the cells of the symmetric tensor coefficients as if they were fully exchangeable, ignoring the fact that coefficients that involve common tensor nodes can be expected to be correlated a priori. Ignoring this correlation can lead to poor predictive performance and potentially impact model selection. Moreover, this architecture does not allow identification of important tensor nodes.

Instead, we develop a generalized symmetric tensor response regression model with a symmetric tensor response and scalar predictors that simultaneously embeds symmetric low-rankness and tensor node-wise sparsity in the symmetric tensor coefficient corresponding to each predictor. Both structures are introduced to achieve several inferential goals simultaneously. The low-rank structure is primarily assumed to capture the interactions between different pairs of tensor nodes, while the node-wise sparsity offers inference on tensor nodes

and cells significantly associated with a predictor. Both structures jointly accomplish parsimony, as well as yield well-calibrated uncertainties for the model parameters. As the mode of inference, a Bayesian framework is adopted, since there is a strong need for uncertainty quantification on the inference related to identifying significant nodes and cells, especially in presence of moderate sample size, with the number of tensor cells far exceeding the number of observations.

The proposed framework shares some commonalities with, but is distinct from, the recent developments in high dimensional regression with multidimensional arrays (tensors) and other object oriented data. For example, recent literature that builds regression models with a scalar response and tensor predictors (Zhou *et al.*, 2013; Guhaniyogi *et al.*, 2017) does not incorporate the symmetry constraint in the tensor. There is more related work about regression frameworks with a scalar response and undirected network predictors, expressed in the form of symmetric matrices (Guha and Rodriguez, 2018; Durante *et al.*, 2018). All these articles treat the tensor as a predictor, whereas we treat it as a response. This difference in the modeling approach leads to a different focus and interpretation. In a way, their difference is analogous to that between multi-response regression and multi-predictor regression in the classical vector-valued regression context.

There have been some recent efforts to build regression models with a tensor response, mostly in the frequentist literature. For example, the formulation proposed in Rabusseau and Kadri (2016) constructs a regression model with a tensor response exploiting a low-rank structure, but does not embed sparsity to identify important tensor nodes and cells. Li and Zhang (2017) propose an envelope-based tensor response model exploiting a generalized sparsity principle, designed to identify linear combinations of the response irrelevant to the regression. In the same vein, Guhaniyogi and Spencer (2018) formulate a Bayesian tensor response regression approach that is built upon a multiway stick breaking shrinkage prior on the tensor coefficients. Spencer *et al.* (2019) further extend the approach by Guhaniyogi

and Spencer (2018) to jointly identify activated brain regions due to a task and connectivity between different brain regions. While Li and Zhang (2017), Guhaniyogi and Spencer (2018) and Spencer *et al.* (2019) are able to identify important tensor cells, they do not allow detection of tensor nodes influenced by a predictor. Moreover, these approaches have not been extended to accommodate scenarios other than a tensor response with continuous cell entries and do not directly incorporate the symmetry constraint in the tensor coefficient corresponding to a predictor. More recently, Sun and Li (2017) have devised a new class of models, referred to as STORE, that impose element-wise sparsity in tensor coefficients. Instead of enforcing element wise sparsity, our proposal develops a tensor node-wise sparsity structure that is in tandem with our inferential goal of detecting tensor nodes influenced by a predictor. Additionally, uncertainty quantification may be somewhat challenging in a frequentist high dimensional regression approach (such as STORE) because standard bootstrap techniques are not consistent for Lasso-type methods (Kyung *et al.*, 2010). Chatterjee and Lahiri (2011) have proposed modifications of the bootstrap producing well-calibrated confidence intervals in the context of standard Lasso regression, but it is not clear whether they extend to the kind of penalties discussed in STORE. In contrast, our Bayesian formulation naturally yields characterization of uncertainty in parameters. Our framework is distinct from the recent Bayesian approach by Roy *et al.* (2017) for modeling brain structural connectomes. While Roy *et al.* (2017) adopt B-splines to model the tensor coefficients, our approach uses low-rank factorization with group sparsity that may offer computational benefits in presence of larger tensors. Additionally, Roy *et al.* (2017) is not designed to identify influential tensor nodes, which is one of the prime inferential objectives in this article.

An important contribution of this article is proving the near optimal contraction rate for the predictive density of the generalized symmetric tensor response model. Theory of posterior contraction for ordinary high dimensional regression models has lately gained traction, with several articles establishing posterior contraction properties, either for various

point-mass priors in the many normal-means models (Castillo *et al.*, 2012; Belitser and Nurushev, 2015; Martin *et al.*, 2017), or for classes of continuous shrinkage priors (Song and Liang, 2017; Wei and Ghosal, 2017). In contrast, the study of posterior contraction properties for generalized linear models involving tensor objects in the Bayesian paradigm has been given far less attention. This article lays down sufficient conditions on the number of tensor nodes, ranks and magnitudes of the true tensor coefficients as a function of the sample size to obtain a *near optimal* convergence rate for the posterior predictive density of the proposed generalized symmetric tensor response model. Our theoretical exposition offers novelty over a few recent articles (Guhaniyogi, 2017; Guhaniyogi and Spencer, 2018) in a number of aspects. The theoretical results derived in these articles mainly pertain to a variety of *multiway shrinkage priors* (Guhaniyogi *et al.*, 2017; Guhaniyogi and Spencer, 2018) in the linear regression context. In contrast, our results are derived for a different class of sparsity inducing priors in the context of a generalized linear model, and importantly, take into account the symmetric nature of the tensor response. Although the existing literature on regression, either with a tensor response or with tensor predictors, offers conditions for posterior consistency, we derive stronger results on the rate of contraction for the posterior predictive density of the proposed model. The details are presented in the supplementary material.

The rest of the article evolves as follows. Section 2 proposes the regression framework with scalar predictors and the symmetric tensor response, and introduces a novel prior distribution on the predictor coefficients to enable identification of nodes and cells of the tensor response related to predictors. Section 3 discusses posterior computation for the proposed model. Empirical investigations with various simulation studies are presented in Section 4, while Section 5 analyzes a real brain connectome dataset. We provide results on significant regions of interest (ROI) and edges. Finally, Section 6 concludes the article with an eye towards future work. Theoretical results on the convergence rate of the posterior

predictive distribution of the proposed model are discussed in the supplementary material. The supplementary material also includes simulation studies for our proposed framework with a two-dimensional binary network response.

2 Model and Framework

This section develops the model and the prior structure for the parameters. We begin by introducing a few notations.

2.1 Notations

A tensor $\boldsymbol{B} \in \otimes_{j=1}^{D} \mathbb{R}^{p_j}$, referred to as a D-way tensor, is a multidimensional array whose $(j_1, ..., j_D)$th *cell* is denoted by $B_{(j_1,...,j_D)}$, $1 \leq j_1 \leq p_1$,...,$1 \leq j_D \leq p_D$. When $D = 2$, a tensor corresponds to a matrix. A tensor $\boldsymbol{B}$ is known to be a symmetric tensor with 0 diagonal entries if $B_{(j_1,...,j_D)} = B_{(g(j_1),...,g(j_D))}$, for any permutation $g(\cdot)$ of $\{j_1, ..., j_D\}$, and $B_{(j_1,...,j_D)} = 0$, if any two of the indices j_l and $j_{l'}$ are equal. The definition of symmetric tensors ensures $p_1 = \cdots = p_D = p$. The indices $\mathcal{N} = \{1, 2, ..., p\}$ for a symmetric tensor $\boldsymbol{B}$ are referred to as *tensor nodes*, analogous to row and column indices for a symmetric matrix. A D-way outer product between vectors $\boldsymbol{b}_k = (b_{k,1}, \ldots, b_{k,p_k})$, $1 \leq k \leq D$, is a $p_1 \times \cdots \times p_D$ tensor denoted by $\boldsymbol{B} = \boldsymbol{b}_1 \circ \boldsymbol{b}_2 \circ \cdots \circ \boldsymbol{b}_D$, with the entry in the $(j_1, .., j_D)$th *cell* given by $B_{(j_1,...,j_D)} = \prod_{k=1}^{D} b_{k,j_k}$. This is also referred to as a rank-1 tensor. Rank-1 decomposition of a symmetric tensor ensures $\boldsymbol{b}_1 = \cdots = \boldsymbol{b}_D$. A rank-R tensor is a sum of R rank-1 tensors. Finally, $||\cdot||$ and $||\cdot||_\infty$ are used to denote the L_2 and L_∞ norms, respectively, for both vectors and higher order tensors.

2.2 Model and prior specification

For $i = 1, ..., n$, let $\boldsymbol{Y}_i = ((y_{i,(j_1,...,j_D)}))_{j_1,...,j_D=1}^{p} \in \mathcal{Y} \subseteq \mathbb{R}^{p \times \cdots \times p}$ denote the symmetric tensor response with 0 diagonal entries, $\boldsymbol{x}_i = (x_{i1}, ..., x_{im})'$ be m predictors of interest and $\boldsymbol{z}_i = (z_{i1}, ..., z_{il})'$ be l auxiliary predictors corresponding to the ith individual. The 0 diagonal

entries of the response tensor are motivated by the brain connectome application described subsequently (Section 5), where a diagonal entry of the symmetric network adjacency matrix corresponds to the number of neuron connections of an ROI with itself, which is customarily taken to be 0. However, the modeling framework as well as the theoretical development can be extended trivially if a '0' in a diagonal entry has to be replaced by some other number using another convention. We assume that the relationship between $\boldsymbol{x}_i$ and the response varies in every tensor cell. In contrast, an auxiliary predictor explains the response in every tensor cell identically. In the context of brain connectome applications, the m predictors of interest may correspond to different phenotypic variables (e.g., composite creativity index) and the l auxiliary predictors consist of demographic variables, such as the age or sex of an individual. Let $\mathcal{J} = \{\boldsymbol{j} = (j_1, ..., j_D) : 1 \leq j_1 < \cdots < j_D \leq p\}$ be a set of indices. Since $\boldsymbol{Y}_i$ is symmetric with dummy diagonal entries, it suffices to build a probabilistic generative mechanism for $y_{i,\boldsymbol{j}}$ ($\boldsymbol{j} \in \mathcal{J}$), where $y_{i,\boldsymbol{j}}$ can either be continuous, binary or categorical. We propose to use a set of conditionally independent generalized linear models $E(y_{i,\boldsymbol{j}}) = \omega_{i,\boldsymbol{j}}$,

$$\omega_{i,\boldsymbol{j}} = H^{-1}\left(\beta_0 + B_{1,\boldsymbol{j}} x_{i1} + \cdots + B_{m,\boldsymbol{j}} x_{im} + \beta_1 z_{i1} + \cdots + \beta_l z_{il}\right), \boldsymbol{j} \in \mathcal{J}, \tag{1}$$

where $B_{1,\boldsymbol{j}}$,...,$B_{m,\boldsymbol{j}}$ are the $\boldsymbol{j} = (j_1, ..., j_D)$th cells of the $p \times \cdots \times p$ symmetric coefficient tensors $\boldsymbol{B}_1, ..., \boldsymbol{B}_m$ with 0 diagonal entries, respectively, and $H(\cdot)$ is an appropriate link function for the outcome of interest. When $y_{i,\boldsymbol{j}}$ corresponds to a normal linear model with the identity link function, (1) becomes

$$\omega_{i,\boldsymbol{j}} = \beta_0 + B_{1,\boldsymbol{j}} x_{i1} + \cdots + B_{m,\boldsymbol{j}} x_{im} + \beta_1 z_{i1} + \cdots + \beta_l z_{il} + \epsilon_{i,\boldsymbol{j}}. \tag{2}$$

The idiosyncratic errors $\epsilon_{i,\boldsymbol{j}}$ follow i.i.d $N(0, \sigma^2)$, and $\beta_0, \beta_1, ..., \beta_l \in \mathbb{R}$ are the intercept and coefficients corresponding to variables $z_{i1}, ..., z_{il}$, respectively. The model formulation implies a similar effect of any of the auxiliary variables $(z_{i1}, ..., z_{il})$ on all cells of the response tensor.

In contrast, $B_{h,\boldsymbol{j}}$, $h = 1, ..., m$; $\boldsymbol{j} \in \mathcal{J}$ corresponds to the coefficient determining the effect of the hth predictor of interest on the $\boldsymbol{j} = (j_1, ..., j_D)$th cell of the symmetric tensor response.

To account for associations of tensor nodes and cells with predictors $x_1, ..., x_m$, as well as for efficient estimation of the high dimensional symmetric tensors $\boldsymbol{B}_1, .., \boldsymbol{B}_m$, we introduce a low-rank structure of $\boldsymbol{B}_h$ as

$$B_{h,\boldsymbol{j}} = \sum_{r=1}^{R} \lambda_{h,r} u_{h,j_1}^{(r)} \cdots u_{h,j_D}^{(r)}, \;\; h = 1, ..., m; \;\; \boldsymbol{j} \in \mathcal{J}, \tag{3}$$

where $\boldsymbol{u}_h^{(r)} = (u_{h,1}^{(r)}, ..., u_{h,p}^{(r)})^T \in \mathbb{R}^p$ are latent vectors of dimensions $p \times 1$ and $\lambda_{h,r} \in \{0, 1\}$ is the binary inclusion variable determining if the rth summand in (3) is relevant in model fitting. Equation (3) has been largely motivated by the symmetric low-rank parallel factor or CP/PARAFAC decomposition of tensors, which are higher order analogues of the factor decomposition of matrices. In particular, define symmetric tensors $\boldsymbol{\Gamma}_h$, $h = 1, ..., m$, admitting a symmetric rank-R CP/PARAFAC decomposition (Kolda and Bader, 2009) of the form, $\boldsymbol{\Gamma}_h = \sum_{r=1}^{R} \lambda_{h,r} \boldsymbol{u}_h^{(r)} \circ \cdots \circ \boldsymbol{u}_h^{(r)}$. Notably, (3) implies that $B_{h,\boldsymbol{j}} = \Gamma_{h,\boldsymbol{j}}$ when $\boldsymbol{j} \in \mathcal{J}$. By the symmetry constraint on $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$, $B_{h,\boldsymbol{j}} = \Gamma_{h,\boldsymbol{j}}$ whenever two indices in $\boldsymbol{j}$ are not equal. The assumed low-rank structure on $\boldsymbol{B}_1, ..., \boldsymbol{B}_m$ offers parsimony by reducing the number of estimable parameters from mp^D to mRp, typically with $R \ll p$. When $D = 2$, the formulation assumes further simplification. To elaborate, denote $\tilde{\boldsymbol{u}}_{h,1} = (u_{h,1}^{(1)}, ..., u_{h,1}^{(R)})^T$,...,$\tilde{\boldsymbol{u}}_{h,p} = (u_{h,p}^{(1)}, ..., u_{h,p}^{(R)})^T$ as R dimensional latent variables corresponding to the tensor nodes and $\boldsymbol{\Lambda}_h = diag(\lambda_{h,1}, .., \lambda_{h,R})$. The $\boldsymbol{j} = (j_1, j_2)$th entry of $\boldsymbol{B}_h$ from (3) then simplifies as $B_{h,\boldsymbol{j}} = \tilde{\boldsymbol{u}}_{h,j_1}^T \boldsymbol{\Lambda}_h \tilde{\boldsymbol{u}}_{h,j_2}$, $\boldsymbol{j} \in \mathcal{J}$, which represents a bilinear interaction between the latent variables $\tilde{\boldsymbol{u}}_{h,j_1}$ and $\tilde{\boldsymbol{u}}_{h,j_2}$ (i.e., corresponding to the j_1th and j_2th nodes, $j_1 < j_2$). This kind of bilinear structure is commonly used to model social and biological networks because of its ability to capture *transitive effects* discussed in the literature (Hoff, 2005, 2008).

With general tensor coefficients $\boldsymbol{B}_1, ..., \boldsymbol{B}_m$, the hth predictor of interest is considered to have no impact on the kth tensor node if $\tilde{\boldsymbol{u}}_{h,k} = \boldsymbol{0}$, $k \in \mathcal{N}$. The $\boldsymbol{j}$th cell is considered

unrelated to the hth predictor of interest if $B_{h,\boldsymbol{j}} = 0$. Since $B_{h,\boldsymbol{j}} = 0$ if $\tilde{\boldsymbol{u}}_{h,j_s} = \mathbf{0}$ for some j_s, the proposed formulation assumes that the contribution of the hth predictor to the $\boldsymbol{j}$th tensor cell is insignificant if the j_sth node is unrelated to the hth predictor, for some j_s.

In order to directly infer on tensor nodes related to the hth predictor of interest, a *spike-and-slab* (Ishwaran and Rao, 2005) mixture distribution prior is assigned on the latent factor $\tilde{\boldsymbol{u}}_{h,k}$ as below

$$\tilde{\boldsymbol{u}}_{h,k} \sim \begin{cases} N(\mathbf{0}, \boldsymbol{M}_h), & \text{if } \eta_{h,k} = 1 \\ \delta_{\mathbf{0}}, & \text{if } \eta_{h,k} = 0 \end{cases}, \quad \eta_{h,k} \sim Ber(\xi_h), \;\; \boldsymbol{M}_h \sim IW(\boldsymbol{S}, \nu), \;\; \xi_h \sim U(0,1) \quad (4)$$

where $\delta_{\mathbf{0}}$ is the Dirac-delta function at $\mathbf{0}$ and $\boldsymbol{M}_h$ is a covariance matrix of order $R \times R$. $IW(\boldsymbol{S}, \nu)$ denotes an Inverse-Wishart distribution with an $R \times R$ positive definite scale matrix $\boldsymbol{S}$ and degrees of freedom ν. The parameter ξ_h corresponds to the probability of the nonzero mixture component and $\eta_{h,k}$ is a binary indicator set to 0 if $\tilde{\boldsymbol{u}}_{h,k} = \delta_{\mathbf{0}}$. Thus, the posterior distributions of $\eta_{h,k}$'s serve as tools to identify nodes related to a predictor.

In order to learn $\lambda_{h,r}$ from (3), we assign a hierarchical prior $\lambda_{h,r} \sim Ber(\upsilon_{h,r})$, $\upsilon_{h,r} \sim Beta(1, r^{\zeta})$, $\zeta > 1$, that imparts increasing shrinkage on $\lambda_{h,r}$ as r grows, to avoid over-fitting. $\hat{R} = \sum_{r=1}^{R} \lambda_{h,r}$ estimates the dimensions of $\tilde{\boldsymbol{u}}_{h,k}$ needed for effective modeling, and is referred to as the *effective dimensionality* of the latent variables. The coefficients $\beta_0, \beta_1, ..., \beta_l$ are a priori assigned a $N(a_\beta, b_\beta)$ distribution. For a tensor response with continuous cell entries, the prior specification is completed by specifying an $IG(a_\sigma, b_\sigma)$ prior on σ^2.

3 Posterior Computation

Although summaries of the posterior distribution cannot be computed in closed form, full conditional distributions for all the parameters are available and mostly correspond to standard families (available in the supplementary material). Thus, posterior computation can proceed through Gibbs sampling implementation. In what follows, we present various

simulations to assess performance of model (1) with continuous and binary cell entries in $\boldsymbol{Y}_i$. For computation with binary cell entries, we consider $H(\cdot)$ as the logit link, and invoke the popular data augmentation technique of Polson *et al.* (2013). The posterior computation does not involve inverting more than an $R \times R$ matrix in each iteration, and hence computation turns out to be rapid. The MCMC sampler is run for 15000 iterations, with the first 5000 discarded as burn-in. All posterior inference is based on post burn-in samples with thinning size 2. The next section reports the effective sample size for 5000 post burn-in iterations averaged over all cells of $\boldsymbol{B}$. All simulation scenarios show an average effective sample size over 3000, indicating fairly uncorrelated post burn-in 5000 MCMC samples.

We have implemented our code in `R` (without using any C++, Fortran or Python interface) on a cluster computing environment with three interactive analysis servers, 56 cores each with the Dell PE R820: 4x Intel Xeon Sandy Bridge E5-4640 processor, 16GB RAM and 1TB SATA hard drive. Different replications of the model are implemented under a parallel architecture by making use of the packages *doparallel* and *foreach* within `R`. Due to the parallel implementation of replications in different cores, computation time for multiple replications does not increase compared to the computation time for one replication. The computation times of running 15000 MCMC iterations with $p = 30$ and $p = 60$ tensor nodes (in Section 4) are given by 162 minutes and 298 minutes, respectively.

S (suitably thinned) post burn-in MCMC samples $\eta_{h,k}^{(1)}, ..., \eta_{h,k}^{(S)}$ of the binary indicator $\eta_{h,k}$ are used to empirically assess if the kth tensor node is significantly associated with the hth predictor of interest. In particular, node k is recognized to be related to the hth predictor if $(1/S)\sum_{s=1}^{S} \eta_{h,k}^{(s)} > t$, $0 < t < 1$. The ensuing simulation section computes the True Positive Rates (TPR) and False Positive Rates (FPR) for various choices of t. For the real data section, we use $t = 0.5$ to decide which nodes are related to a specific predictor.

To assess the accuracy of identifying important cells of the response tensor related to the hth predictor, we compute post burn-in MCMC samples of $B_{h,\boldsymbol{j}}$ following (3) for all $\boldsymbol{j} \in \mathcal{J}$.

The proportion of post burn-in MCMC samples where $B_{h,\boldsymbol{j}} = 0$ is computed, and the $\boldsymbol{j}$th cell is recognized to be related to the hth predictor if this is less than 0.5. The True Positive Rates (TPR) and False Positive Rates (FPR) for different simulation settings are computed.

4 Simulation Studies

We consider two simulation examples to illustrate our approach of the symmetric generalized tensor model (SGTM) and compare with competing methods. The first simulation example illustrates the proposed approach with the tensor response generated from (1) having continuous cell entries (Section 4.1), whereas the second simulation example (in the supplementary material) considers the tensor response with binary cell entries simulated from (1), setting $H(\cdot)$ as the logit link function. For the sake of simplicity, we fix $m = 1$ and $l = 2$ in both simulation examples, which also holds for the brain connectome application in Section 5.

For simulations, the predictor of interest x_i and the auxiliary covariates z_{i1} and z_{i2} are drawn iid from $N(0, 1)$. The true intercept β_0^*, true coefficients β_1^* and β_2^* are set as 0.2, 0.4 and -0.1, respectively, for data simulation. To simulate the true symmetric tensor coefficient $\boldsymbol{B}^*$, we begin by drawing p tensor node specific latent variables $\tilde{\boldsymbol{u}}_k^* = (u_k^{*(1)}, ..., u_k^{*(R^*)})^T$, $k = 1, ..., p$, each of dimension R^*, from a mixture distribution given by $\tilde{\boldsymbol{u}}_k^* \sim \pi_1^* N_{R^*}(0.8\mathbf{1}, 0.25\boldsymbol{I}) + (1 - \pi_1^*)\delta_{\mathbf{0}}$. The probability of a tensor node being not related to x_i, referred to as the *node sparsity* parameter, is denoted by $(1 - \pi_1^*)$. In simulations, the entries of $B_{\boldsymbol{j}}^*$, $\boldsymbol{j} \in \mathcal{J}$ are constructed under two different scenarios.

Scenario 1: The first scenario constructs the cell coefficients as multi-linear interactions between the corresponding node specific latent variables, $B_{\boldsymbol{j}}^* = \sum_{r=1}^{R^*} u_{j_1}^{*(r)} \cdots u_{j_D}^{*(r)}$.

Scenario 2: The second scenario assumes mis-specification between the fitted and simulated models and simulates $B_{\boldsymbol{j}}^*$ from a mixture distribution $B_{\boldsymbol{j}}^* \sim \pi_2^* N(0, 1) + (1 - \pi_2^*)\delta_0$ when $\tilde{\boldsymbol{u}}_{j_1}^*$,...,$\tilde{\boldsymbol{u}}_{j_D}^*$ are all nonzero; $B_{\boldsymbol{j}}^* = 0$ otherwise. The quantity $(1 - \pi_2^*)$ is referred to as the *cell*

Cases	R	R^*	n	p	π_1^*
1	4	2	70	30	0.4
2	3	2	70	60	0.6
3	5	2	100	30	0.5
4	5	3	100	60	0.7

Table 1: Simulation cases for Scenario 1. The node sparsity parameter is given by $(1 - \pi_1^*)$.

Cases	R	R^*	n	p	π_1^*	π_2^*
1	4	2	70	30	0.4	0.5
2	3	2	70	60	0.6	0.5
3	5	2	100	30	0.5	0.5
4	5	3	100	60	0.7	0.5
5	4	2	70	30	0.4	0.7
6	4	2	70	30	0.6	0.5
7	4	2	70	30	0.6	0.7

Table 2: Simulation cases for Scenario 2. The node and cell sparsity parameters are $(1 - \pi_1^*)$ and $(1 - \pi_2^*)$, respectively.

sparsity parameter.

In Scenario 1, we choose four (n, p, R^*, π_1^*) combinations (shown in Table 1), whereas seven different combinations of $(n, p, R^*, \pi_1^*, \pi_2^*)$ are considered in Scenario 2 (shown in Table 2). The tensor coefficient $\boldsymbol{B}^*$ assumes a low-rank decomposition in Scenario 1 similar to the fitted model, though Scenario 1 includes cases with a mismatch between the true and fitted dimensions of node specific latent variables. Scenario 2 constructs a sparse true coefficient tensor $\boldsymbol{B}^*$ that does not have a low-rank structure, allowing investigation with model mis-specification. It also introduces the additional notion of cell sparsity and allows us to investigate model performance with varying node and cell sparsity.

Choice of prior hyperparameters: Note that the choice of a $U(0, 1)$ prior distribution for ξ is to ensure a uniform distribution on the number of active nodes, and conditional on the size of the model (i.e., the number of active nodes), a uniform distribution on all possible models of that size. For model fitting, the hyper-parameters are fixed at $\boldsymbol{S} = \boldsymbol{I}_{R\times R}$, $\nu = 10$,

$a_\sigma = b_\sigma = 1$ and $a_\beta = 0$, $b_\beta = 1$. The choice of $\nu = 10$ and $\boldsymbol{S} = \boldsymbol{I}_{R\times R}$ ensure that the prior distribution of $\boldsymbol{M}$ is concentrated around a scaled identity matrix. To provide a justification for this choice when $D = 2$, notice that the cell coefficient for $D = 2$, given by $B_{\boldsymbol{j}} = \tilde{\boldsymbol{u}}_{j_1}^T \boldsymbol{\Lambda} \tilde{\boldsymbol{u}}_{j_2}$, is invariant to rotations of the latent positions $\tilde{\boldsymbol{u}}_k$s, and so we would like the prior on $\tilde{\boldsymbol{u}}_k$s to also be invariant under rotations. This requires that we center $\boldsymbol{M}$ around a matrix that is proportional to the identity matrix. Finally, $a_\sigma = b_\sigma = 1$ ensures an infinite prior mean of σ^2, implying somewhat less prior information on σ^2. Upon perturbing the hyper-parameters moderately, we find practically indistinguishable inference.

Performance metrics: To infer on the performance of SGTM in terms of identifying tensor nodes significantly associated with x_i, we present TPR and FPR values with different choices of the cut-off t for all simulation cases. To assess the performance of SGTM in identifying tensor cells significantly associated with x_i, we present TPR and FPR values in all simulations for a cut-off of 0.5, as discussed in Section 3. The accuracy of estimating $\boldsymbol{B}^*$ is measured by the scaled mean squared error (MSE) defined as $||\boldsymbol{B}^* - \hat{\boldsymbol{B}}||^2/||\boldsymbol{B}^*||^2$, where $\hat{\boldsymbol{B}}$ corresponds to a suitable point estimate of $\boldsymbol{B}$, for e.g., the posterior mean of $\boldsymbol{B}$ for SGTM. The length and coverage of posterior 95% credible intervals for each $B_{\boldsymbol{j}}\,, \boldsymbol{j} \in \mathcal{J}$ are available empirically from the post burn-in MCMC samples of $\boldsymbol{B}$ to assess the quantification of uncertainty by the proposed approach. Finally, posterior distributions of the effective dimensionality $\hat{R}$ under different simulation cases are also reported to infer on the dimension of the node specific latent variables required for model fitting. All results presented are averaged over 50 simulation replicates.

Competitors: In Section 4.1, we compare our approach to ordinary least squares (LS), which proposes a cell by cell regression of the response on the predictors. Although a naive approach, LS is included due to its widespread use in neuro-imaging applications. Additionally, we employ the *envelope method* (ENV) (Li and Zhang, 2017) and *Higher-Order Low-Rank Regression* (HOLRR) (Rabusseau and Kadri, 2016) as competitors. While

ENV proposes a regression framework with a general tensor response and scalar predictors, HOLRR provides a framework for higher order regression with a tensor response and scalar predictors. Comparing with these three methods will help assess possible gains in inference in our method for taking into account the symmetry in the tensor response. Unfortunately, HOLRR and ENV have not been designed for binary tensors yet, and hence are omitted as competitors for binary symmetric tensor response regression presented in the supplementary material, where we simply compare our approach with a cell by cell logistic regression model and refer to it as LS with a slight abuse of notation. Since none of these competitors are designed to identify influential tensor nodes, performance comparisons are mainly based on MSE, coverage and length of 95% CIs. The 95% confidence intervals of frequentist competitors are constructed using a bootstrap approximation. We are unable to compare our approach with the most relevant competitor STORE (Sun and Li, 2017) due to unavailability of open source codes.

4.1 2D Symmetric Continuous Tensor Response Example

We compute the effective sample size of thinned 5000 post burn-in iterates of $\boldsymbol{B}$ averaged over all cells for each replication. This quantity is computed over all replications and their average over all replications are reported. In particular, the values of the average effective sample size of $\boldsymbol{B}$, averaged over all replications, are given by 4822, 4915, 4986 and 4746 in the four cases in Scenario 1, and 3337, 3618, 3677, 3246, 3898, 3512 and 3702 in the seven cases in Scenario 2 respectively, indicating fairly uncorrelated samples to draw inference on. Tables 3 and 4 present scaled MSE, coverage and length of 95% CI for all four competitors under Scenarios 1 and 2, respectively. The four cases under Scenario 1 show much lower MSE for SGTM than its competitors, ENV turning out to be the next best performer with a larger p/n ratio. Notably, these two competitors (SGTM and ENV) account for the structure of the tensor response, though ENV does not incorporate the symmetry constraint in the response tensor. This is perhaps responsible for ENV performing marginally inferior to LS when the

sample size is larger. All four competitors under Scenario 1 demonstrate over-coverage, with SGTM producing substantially narrower credible intervals. The over-coverage in SGTM is perhaps due to mild over-fitting which is observed frequently in high dimensional Bayesian models in presence of small sample size, with data simulated from the true model (see for e.g., Guhaniyogi *et al.*, 2017). Moreover, increasing the number of nodes p for a fixed n results in narrower credible intervals for all competitors.

The true coefficient $\boldsymbol{B}^*$ does not assume any low-rank decomposition in Scenario 2. Consequently, theoretical guarantee on optimal performance of SGTM is no longer valid (see supplementary material). Nevertheless, SGTM appears to mildly outperform all competitors in cases with smaller p and higher node sparsity (Cases 1 and 3), as seen in Table 4. Perhaps the low-rank decomposition of $\boldsymbol{B}$ offered by SGTM is sufficient to estimate $\boldsymbol{B}^*$ in presence of a smaller p/n ratio and higher node sparsity. However, with a larger p, as in Cases 2 and 4, LS outperforms SGTM. Since ENV and HOLRR are constructed on variants of low-rank and/or sparsity principles as well, they also generally lose edge over LS in terms of MSE in these cases. Comparing MSE of SGTM between Cases 1 and 6, we find that decreasing node sparsity has an adverse effect on MSE. Similarly, comparing Cases 1 and 5 reveals a marginally adverse effect of decreasing cell sparsity on MSE of SGTM. It is generally found that the effect of node sparsity is more profound than the effect of cell sparsity on the performance. In terms of uncertainty characterization, under Cases 1 and 3, with smaller p and higher sparsity, SGTM, ENV and HOLRR demonstrate close to nominal coverage, while more challenging Cases 2 and 4 show under-coverage of all these three competitors. Again, with decreasing node and edge sparsity, the coverage of SGTM, ENV and HOLRR are found to drop around $80 - 85\%$, with SGTM having sufficiently narrower credible intervals than ENV and HOLRR. LS offers over-coverage with much wider 95% credible intervals in all cases.

Under both Scenarios 1 and 2, SGTM yields the posterior probability of a node being

Case		Competitors			
		SGTM	LS	HOLRR	ENV
1	MSE$\times 10^3$	$\mathbf{2.3}_{0.9}$	$50_{9.4}$	$48_{14.2}$	$37_{14.9}$
	Avg. Cov	$0.98_{0.00}$	$0.98_{0.00}$	$0.98_{0.00}$	$0.98_{0.00}$
	Avg. Length	$0.08_{0.00}$	$1.81_{0.00}$	$1.65_{0.00}$	$1.59_{0.00}$
2	MSE$\times 10^3$	$\mathbf{0.7}_{0.8}$	$18_{4.9}$	$11_{4.2}$	$8.7_{3.1}$
	Avg. Cov	$0.99_{0.00}$	$0.99_{0.00}$	$0.99_{0.00}$	$0.99_{0.00}$
	Avg. Length	$0.10_{0.00}$	$2.94_{0.00}$	$2.75_{0.00}$	$2.82_{0.00}$
3	MSE$\times 10^3$	$\mathbf{1.5}_{0.7}$	$23_{7.84}$	$33_{7.29}$	$29_{8.22}$
	Avg. Cov	$0.99_{0.00}$	$0.99_{0.00}$	$0.99_{0.00}$	$0.99_{0.00}$
	Avg. Length	$0.10_{0.01}$	$2.49_{0.00}$	$2.41_{0.00}$	$2.36_{0.00}$
4	MSE$\times 10^3$	$\mathbf{0.3}_{0.2}$	$4.4_{1.3}$	$7.3_{2.56}$	$6.25_{2.28}$
	Avg. Cov	$0.99_{0.00}$	$1.00_{0.00}$	$0.99_{0.00}$	$1.00_{0.00}$
	Avg. Length	$0.16_{0.01}$	$4.55_{0.00}$	$4.40_{0.00}$	$4.38_{0.00}$

Table 3: Mean Squared Error (MSE), average coverage and average length of 95% credible interval for SGTM, LS, HOLRR and ENV are presented for cases under simulation Scenario 1 for the continuous case. The lowest MSE in each case is boldfaced. Results are averaged over 50 replications and the standard deviations over 50 replications are reported in the subscript of every number.

related to the predictor of interest to be very close to 1 or 0 for all reasonable values of cut-off t, depending on whether a tensor node is related or not to the predictor of interest in the truth, respectively. As a consequence, TPR and FPR values (see Figure 1) turn out to be close to 1 and 0, respectively, for all the simulation cases, indicating a close to perfect active node detection. A close investigation of Figure 1 also reveals marginally better performance in terms of node identification with decreasing node or edge sparsity in Scenario 2.

Table 5 presents TPR and FPR values for identifying cells significantly related to the predictor of interest. While Scenario 1 yields excellent performance, the performance tends to be moderate in Scenario 2 under model mis-specification. In particular, under Scenario 2, with decreasing node and/or edge sparsity, both TPR and FPR increase significantly. We also observe significantly higher TPR and FPR upon increasing the number of tensor nodes p.

Figure 2 presents posterior probabilities of the effective dimensionality $\hat{R}$ in all simulation cases under Scenarios 1 and 2. Under Scenario 1, $\hat{R}$ yields the highest posterior probability

Case		Competitors			
		SGTM	LS	HOLRR	ENV
1	MSE	$\mathbf{0.13}_{0.04}$	$0.23_{0.06}$	$0.29_{0.08}$	$0.22_{0.04}$
	Avg. Cov	$0.97_{0.01}$	$0.97_{0.01}$	$0.91_{0.02}$	$0.93_{0.01}$
	Avg. Length	$0.14_{0.02}$	$0.94_{0.00}$	$0.74_{0.00}$	$0.71_{0.00}$
2	MSE	$0.33_{0.03}$	$0.25_{0.04}$	$\mathbf{0.20}_{0.02}$	$0.28_{0.02}$
	Avg. Cov	$0.82_{0.03}$	$0.98_{0.00}$	$0.84_{0.00}$	$0.84_{0.00}$
	Avg. Length	$0.10_{0.01}$	$1.48_{0.00}$	$0.85_{0.00}$	$0.82_{0.00}$
3	MSE	$\mathbf{0.12}_{0.02}$	$0.17_{0.04}$	$0.28_{0.05}$	$0.20_{0.03}$
	Avg. Cov	$0.95_{0.01}$	$0.98_{0.00}$	$0.93_{0.01}$	$0.91_{0.01}$
	Avg. Length	$0.22_{0.02}$	$1.14_{0.00}$	$0.92_{0.00}$	$0.88_{0.00}$
4	MSE	$0.28_{0.04}$	$\mathbf{0.12}_{0.00}$	$0.25_{0.03}$	$0.29_{0.03}$
	Avg. Cov	$0.79_{0.04}$	$0.99_{0.00}$	$0.83_{0.01}$	$0.82_{0.01}$
	Avg. Length	$0.18_{0.02}$	$1.70_{0.00}$	$1.00_{0.00}$	$0.92_{0.00}$
5	MSE	$\mathbf{0.15}_{0.02}$	$0.21_{0.04}$	$0.31_{0.08}$	$0.20_{0.04}$
	Avg. Cov	$0.96_{0.00}$	$0.98_{0.01}$	$0.92_{0.01}$	$0.92_{0.01}$
	Avg. Length	$0.16_{0.02}$	$1.11_{0.00}$	$0.84_{0.00}$	$0.80_{0.00}$
6	MSE	$0.24_{0.04}$	$\mathbf{0.10}_{0.01}$	$0.36_{0.04}$	$0.35_{0.02}$
	Avg. Cov	$0.87_{0.02}$	$0.99_{0.00}$	$0.92_{0.01}$	$0.91_{0.01}$
	Avg. Length	$0.23_{0.01}$	$1.46_{0.00}$	$1.09_{0.00}$	$1.07_{0.00}$
7	MSE	$0.26_{0.03}$	$\mathbf{0.12}_{0.01}$	$0.36_{0.03}$	$0.41_{0.03}$
	Avg. Cov	$0.85_{0.01}$	$0.99_{0.00}$	$0.90_{0.01}$	$0.90_{0.01}$
	Avg. Length	$0.26_{0.04}$	$1.72_{0.00}$	$1.30_{0.00}$	$1.25_{0.00}$

Table 4: Mean Squared Error (MSE), average coverage and average length of 95% credible interval for SGTM, LS, HOLRR and ENV are presented for cases under simulation Scenario 2 for the continuous case. The lowest MSE in each case in boldfaced. Results are averaged over 50 replications and standard deviations over 50 replications are reported in the subscript of every number.

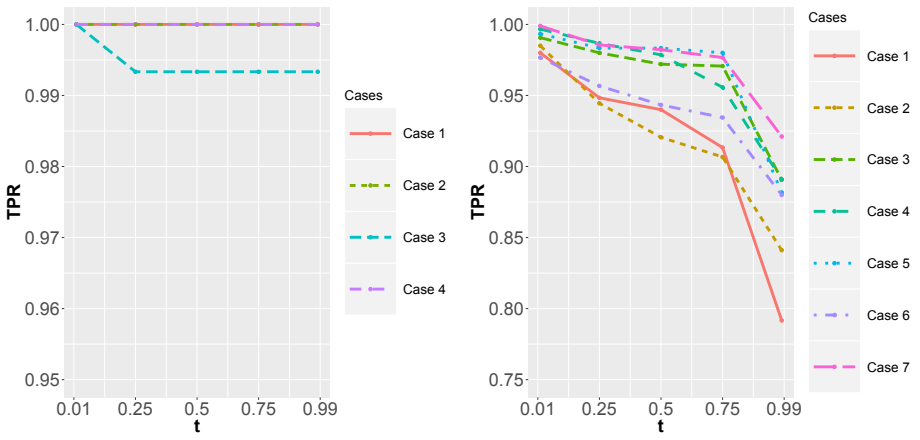
corresponding to the true dimension R^* of the latent variables in all cases. We expect this observation, since under Scenario 1, the true coefficient $\boldsymbol{B}^*$ assumes a low-rank structure with R^* less than the fitted dimension R. The variable selection architecture on the λ_r's are able to recover the dimension of the true symmetric tensor. In contrast, $\boldsymbol{B}^*$ is full rank under Scenario 2. Hence, SGTM, with the tensor coefficient $\boldsymbol{B}$ having a low-rank structure, while estimating $\boldsymbol{B}^*$ by $\boldsymbol{B}$, consumes all available ranks, resulting in the posterior mode of $\hat{R}$ being R in all simulations.

Scenario	Operating Characteristic		
		TPR	FPR
1	Case 1	0.98	0.00
	Case 2	0.99	0.00
	Case 3	0.97	0.00
	Case 4	1.00	0.00
2	Case 1	0.42	0.00
	Case 2	0.65	0.13
	Case 3	0.52	0.07
	Case 4	0.78	0.23
	Case 5	0.48	0.02
	Case 6	0.68	0.12
	Case 7	0.64	0.08

Table 5: True Positive Rates (TPR) and False Positive Rates (FPR) in identifying cells which are significantly related to the predictor under all cases in Section 4.1.

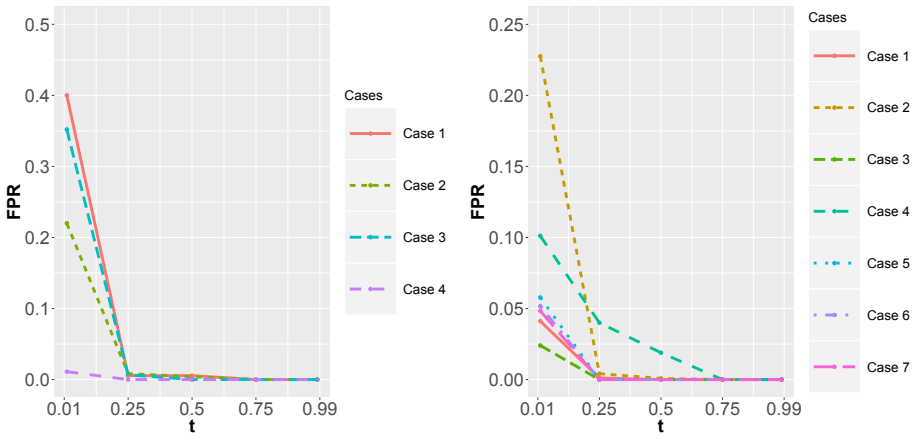
5 Brain Connectome Application

This section illustrates the inferential ability of SGTM for symmetric tensor responses with continuous-valued cell entries in the context of a diffusion tensor magnetic resonance imaging (DTI) dataset. DTI is an imaging procedure that allows the measurement of restricted diffusion of water in brain tissues to construct neural tract images. In the context of DTI, the human brain is divided into 68 cortical regions of interest (ROIs), with 34 regions each in the left and the right hemispheres, respectively, using the Desikan brain atlas (Desikan *et al.*, 2006). Using DTI, a *brain network* for each subject is constructed as a symmetric matrix with row and column indices corresponding to different ROIs, and entries corresponding to the estimated number of 'fibers' connecting pairs of brain regions. We standardize each entry of the network response matrix by centering and scaling over all the subjects. The centered and scaled network response matrices have cell entries in $\mathbb{R}$ which allows us to assume normality of the error distributions. For each subject, the dataset also has information on a measure of *creativity*, known as the Composite Creativity Index (CCI). The CCI measure, proposed by Jung *et al.* (2010), is formulated by linking measures of divergent thinking and creative achievement to cortical thickness of young (23.7 ± 4.2 years),



(a) TPR: Scenario 1

(b) TPR: Scenario 2



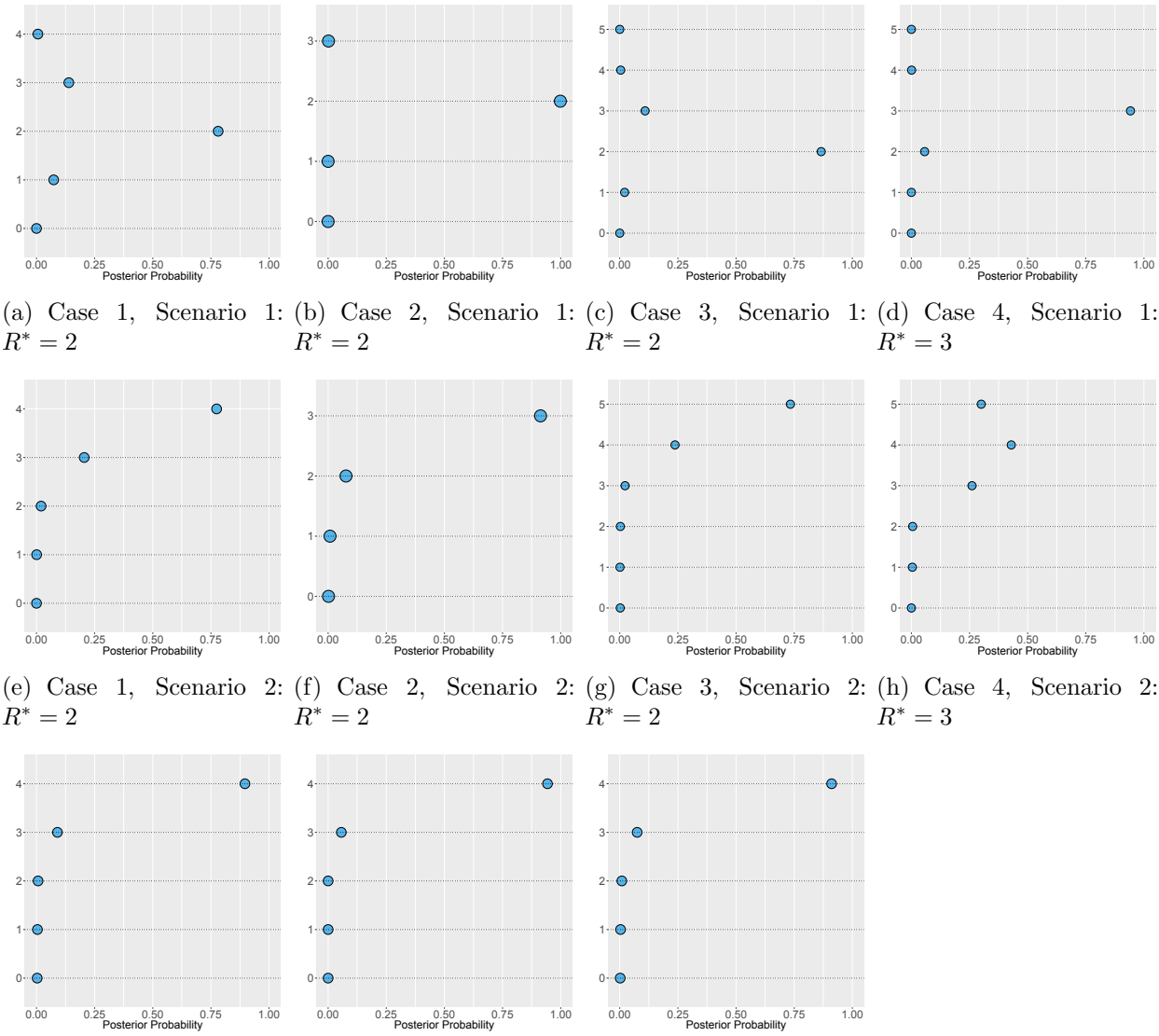
(c) FPR: Scenario 1

(d) FPR: Scenario 2

Figure 1: Figures in the first row show TPR values in detecting tensor nodes significantly associated with the predictor of interest, under each scenario (Section 4.1). Figures in the second row show FPR values in detecting tensor nodes significantly associated with the predictor of interest, under each scenario (Section 4.1).

healthy subjects. Three independent judges grade the creative products of a subject from which the CCI is derived. The DTI dataset we consider consists of brain network information along with CCI for $n = 79$ individuals. As mentioned before, both the symmetric brain network tensor and CCI values are standardized over individuals. Age (standardized) and sex (binary) are also available as additional covariates.

Our primary scientific goal in this context is to comprehend the relationship between brain connectivity patterns and creativity, as measured by the CCI. Principally, we would like to



(a) Case 1, Scenario 1: $R^* = 2$ (b) Case 2, Scenario 1: $R^* = 2$ (c) Case 3, Scenario 1: $R^* = 2$ (d) Case 4, Scenario 1: $R^* = 3$

(e) Case 1, Scenario 2: $R^* = 2$ (f) Case 2, Scenario 2: $R^* = 2$ (g) Case 3, Scenario 2: $R^* = 2$ (h) Case 4, Scenario 2: $R^* = 3$

(i) Case 5, Scenario 2: $R^* = 2$ (j) Case 6, Scenario 2: $R^* = 2$ (k) Case 7, Scenario 2: $R^* = 2$

Figure 2: Posterior distribution of the effective dimensionality in various cases under Scenarios 1 and 2 for SGTM in the continuous case. The first row shows the four cases in Scenario 1, while the second and third rows show the seven cases in Scenario 2.

identify nodes in the brain (or brain *regions of interest* - ROIs) significantly related to the CCI, controlling for confounding covariates (age and gender). Thus model (2) with $R = 5$ is analyzed with a standardized brain network matrix as the symmetric tensor response, CCI as the predictor of interest, and age and sex as auxiliary predictors. Identical prior distributions are used as in the simulation studies.

5.1 Findings from SGTM

Figure 3 shows the estimated posterior probability of each ROI being actively related to CCI. The kth ROI is identified as *active* if $P(\eta_k = 1|Data)$ exceeds 0.5. This criteria, when applied to the real data discussed above, identifies 34 ROIs out of 68 as *active*, which are mentioned in Table 6 . Of these 34 ROIs, 18 belong to the left portion of the brain (or the left hemisphere) and 16 belong to the right hemisphere.

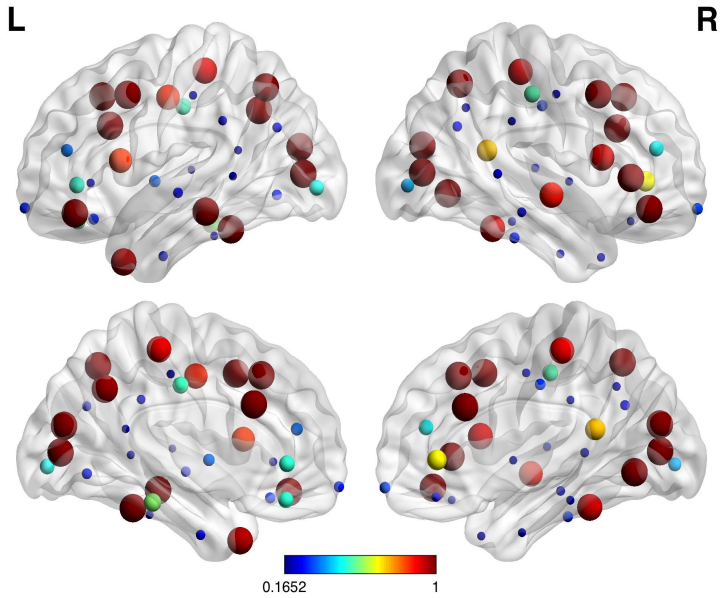


Figure 3: Left and right hemispheres of the human brain (lateral and medial views) with all 68 regions of interest (ROIs). The color and size of the ROIs vary according to the value of the posterior probabilities of them being actively related to the composite creativity index (CCI).

Among the active ROIs detected by our method, a sizable number is part of the *frontal* (13) and *temporal* (6) regions in both hemispheres. The frontal cortex has been scientifically linked with spontaneity, memory, language, initiation, divergent thinking, problem solving ability, motor function, judgement, risk taking, impulse control and social behavior. *De novo* artistic expression has also been found to be associated with the temporal and frontal regions (Razumnikova, 2007, Miller and Milner, 1985).

Left Hemisphere Lobes					
Temporal	**Cingulate**	**Frontal**	**Occipital**	**Parietal**	**Insula**
fusiform	rostral-anteriorcingulate	caudal-middlefrontal	cuneus	precuneus	
middle frontal gyrus	caudal-anteriorcingulate	medial-orbitofrontal	pericalcarine	superior-parietal	
parahippocampal	posterior cingulate	paracentral			
temporal pole		pars-triangularis			
		precentral			
		superior-frontal gyrus			
		pars-orbitalis			
Right Hemisphere Lobes					
Temporal	**Cingulate**	**Frontal**	**Occipital**	**Parietal**	**Insula**
fusiform	caudal-anteriorcingulate	caudal-middlefrontal	cuneus	postcentral	
superior-temporal	isthmus-cingulate	paracentral	pericalcarine	superior-parietal	
	rostal-anteriorcingulate	pars-opercularis	lingual		
		pars-orbitalis			
		pars-triangularis			
		superior-frontal			

Table 6: Brain regions (ROIs) associated with the composite creativity index as detected by SGTM. The significant ROIs have been listed according to their memberships in 12 *lobes* (anatomical portions) of the brain.

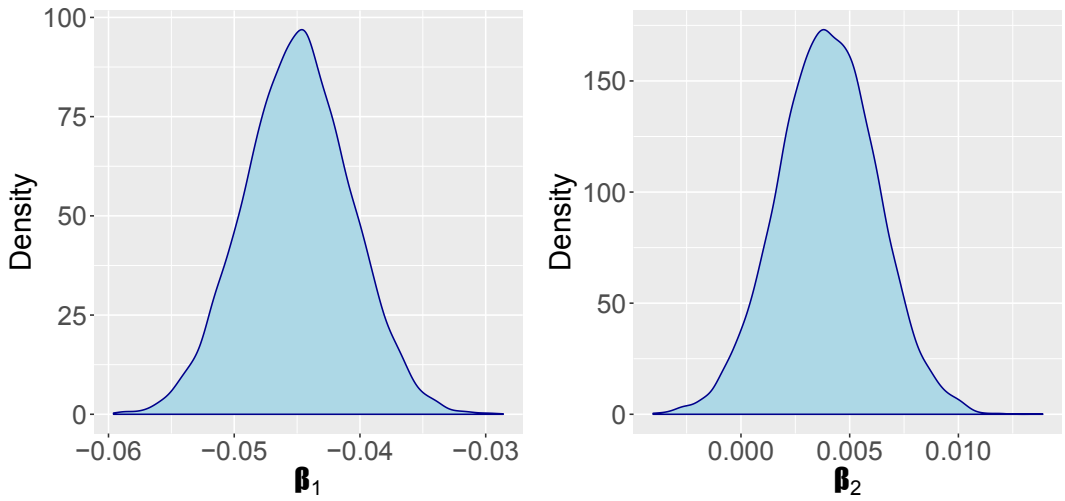
	SGTM	LS	HOLRR	ENV
MSPE	0.72	0.79	0.78	0.76
Avg. Coverage	0.96	0.99	0.98	0.99
Avg. Length	3.47	4.96	4.53	4.04

Table 7: MSPE, coverage and length of 95% predictive intervals corresponding to all competitors for the real data. Coverage and length of 95% predictive intervals are averaged over all samples.

Figure 4 shows the estimated posterior densities of the sex and age covariates, where sex appears to be significantly related to the brain connectome. Age appears not to be significant, perhaps due to the fact that all subjects in the brain connectome data are within a narrow age range of about $19-28$ years. The posterior mean of the effective dimensionality $\hat{R}$ turns out to be 2.

The predictive performance of SGTM, LS, HOLRR and ENV are also compared and presented in Table 7, which shows marginally superior performance by SGTM over its competitors in terms of MSPE, with ENV being the second best performer. All competitors show coverage more than nominal, though SGTM yields the narrowest predictive interval among all competitors.

Sensitivity to the choice of R. Finally, to check sensitivity to the choice of R on the



(a) Posterior density of β_1

(b) Posterior density of β_2

Figure 4: Figure on the left shows the posterior density of β_1, the coefficient for the sex covariate. The figure on the right shows the posterior density of β_2, the coefficient for the age covariate.

	$R = 5$	$R = 7$	$R = 9$
MSPE	0.72	0.73	0.73
Mean effect. dimensionality	2.00	2.19	2.28

Table 8: MSPE and mean effective dimensionality of SGTM under different choices of R.

performance of SGTM, we moderately perturb R and run the data analysis for SGTM with $R = 7$ and $R = 9$, and report the posterior mean of the effective dimensionality, along with MSPE. Table 8 shows moderate increase in the mean effective dimensionality with choices of larger R, though the shrinkage effect pulls the mean of $\hat{R}$ close to 2 under different choices of R. The perturbation of R has a negligible effect on MSPE, as observed in Table 8.

6 Conclusion and Future Work

This article proposes a Bayesian generalized linear modeling framework with a symmetric tensor response and a vector covariate. Adopting a symmetric rank-R PARAFAC decomposition for the tensor coefficients corresponding to the scalar components of the vector

covariate, the proposed model is able to considerably reduce the number of free parameters to model. We employ a novel hierarchical mixture prior to enable identification of tensor nodes and cells significantly related to each predictor. A major contribution of this article pertains to detailing out theoretical conditions to achieve a near optimal posterior convergence rate for the predictive distribution of the proposed model. Simulation examples and the brain connectome data application demonstrate superior empirical performance of the proposed method with respect to the state-of-the-art competitors.

Although this article involves a symmetric tensor response of order two in the data application, there are some interesting applications and scientific questions that may be addressed using symmetric tensor regression with higher order tensors, for instance in the field of international trade. Consider the following motivating example. Economists studying international trade closely monitor multilateral trade (for e.g., trilateral trade agreements) between countries, along with a number of country specific economic indicators, e.g., the sum of the gross domestic products (GDP) and the sum of the total output of the manufacturing sectors of these countries (Chan and Kuo, 2005; Li and Zhou, 2009). These economic datasets often lead to multiple scientific questions relevant to world economic behavior. First, it is important to understand how trilateral trade varies with the sum of GDP or the total output of the manufacturing sectors of these countries over time. Another important question is to identify countries which are significant economic engines of the world, from the perspective of driving GDP growth and other important economic indicators. To formulate this problem statistically, consider p countries at time i, and looking at triplets of countries, data on trilateral trade between them can be thought of as a 3-D symmetric tensor, denoted by $\boldsymbol{y}_i = ((y_{i,\boldsymbol{j}}))$, with the $\boldsymbol{j} = (j_1, j_2, j_3)$th entry $y_{i,\boldsymbol{j}}$ recording the total trilateral trade between countries j_1, j_2 and j_3. The predictor variables of interest are x_{i1} = sum of the total GDP of the p countries at time i, and x_{i2} = sum of total manufacturing output of the p countries at time i. Monthly data on the symmetric tensor and the two variables of interest over the

relevant time period can be modeled using the proposed symmetric tensor regression model. Importantly, our model offers inference on the nodes of the symmetric tensor significantly related to the predictors, i.e., the countries which are major drivers of the world economy. In this context, it is generally believed that free trade agreements between countries can benefit the overall economic health of the world. Notably, there have been important trilateral free trade agreements between China-Japan-South Korea (Chiang, 2013), and U.S.A.-Canada-Mexico (referred to as the North Atlantic Free Trade Agreement (NAFTA)) (Brown *et al.*, 1995). It is instructive to statistically analyze if the countries benefited from the trilateral free trade agreements and which were the economic drivers of the world.

Our framework can be extended to address modeling questions mildly different from ours. For example, if instead of having dummy entries in the diagonal, our interest also lies in modeling nontrivial entries in the diagonal, we can simply model $\boldsymbol{B}_h = \sum_{r=1}^{R} \lambda_{h,r} \boldsymbol{u}_h^{(r)} \circ \cdots \circ \boldsymbol{u}_h^{(r)}$, instead of modeling its upper triangular entries jointly. Moreover, there are motivating applications in which a higher-order tensor response has symmetry across some modes, but not others. We can extend our modeling framework in such cases by sharing components of the CP/PARAFAC decomposition along modes having symmetry. For example, a brain network can be inferred over multiple conditions, which would yield a three-way tensor for each individual, ROIs $\times$ ROIs $\times$ Conditions, with symmetry in the first two modes but not the third. In this case, we model $B_{h,\boldsymbol{j}} = \sum_{r=1}^{R} \lambda_{h,r} u_{h,j_1}^{(r)} u_{h,j_2}^{(r)} w_{h,j_3}^{(r)}$, i.e., the component of the CP factorization corresponding to the mode representing the conditions is modeled using a different vector $\boldsymbol{w}_h^{(r)} = (w_{h,1}^{(r)}, ..., w_{h,s}^{(r)})'$. Here s represents dimension of $\boldsymbol{B}_h$ along the mode 'condition.'

7 Acknowledgement

The first author is partially supported by funds from the Trinity College of Arts and Sciences at Duke University. The second author is partially supported by Office of Naval Research award no. BAA-N00014-18-2741 and National Science Foundation award no. DMS-

1854662.

Supplementary Material

Proofs, additional simulation results, and the R codes for implementing the results are available online.

References

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

Brown, D. K., Deardorff, A. V., and Stern, R. M. (1995). Estimates of a north american free trade agreement. In *Modeling North American Economic Integration*, pages 59–74. Springer.

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

Chan, S. and Kuo, C.-C. (2005). Trilateral trade relations among china, japan and south korea: Challenges and prospects of regional economic integration. *East Asia*, **22**(1), 33–50.

Chatterjee, A. and Lahiri, S. N. (2011). Bootstrapping Lasso Estimators. *Journal of the American Statistical Association*, **106**(494), 608–625.

Chiang, M.-H. (2013). The potential of china-japan-south korea free trade agreement. *East Asia*, **30**(3), 199–216.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., *et al.* (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, **31**(3), 968–980.

Durante, D., Dunson, D. B., *et al.* (2018). Bayesian inference and testing of group differences in brain networks. *Bayesian analysis*, **13**(1), 29–58.

Goh, G., Dey, D. K., and Chen, K. (2017). Bayesian sparse reduced rank multivariate regression. *Journal of multivariate analysis*, **157**, 14–28.

Guha, S. and Rodriguez, A. (2018). Bayesian Regression with Undirected Network Predictors with an Application to Brain Connectome data. *arXiv preprint arXiv:1803.10655*.

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

Guhaniyogi, R. and Spencer, D. (2018). Bayesian tensor response regression with an application to brain activation studies. *UCSC Tech Report: UCSC-SOE-18-15*.

Guhaniyogi, R., Qamar, S., and Dunson, D. B. (2017). Bayesian Tensor Regression. *Journal of Machine Learning Research*, **18**(79), 1–31.

Hoff, P. (2008). Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in neural information processing systems*, pages 657–664.

Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, **100**(469), 286–295.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Annals of Statistics*, **33**(2), 730–773.

Jung, R. E., Segall, J. M., Jeremy Bockholt, H., Flores, R. A., Smith, S. M., Chavez, R. S., and Haier, R. J. (2010). Neuroanatomy of creativity. *Human Brain Mapping*, **31**(3), 398–409.

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

Kyung, M., Gill, J., Ghosh, M., Casella, G., *et al.* (2010). Penalized Regression, Standard Errors, and Bayesian Lassos. *Bayesian Analysis*, **5**(2), 369–411.

Li, H. and Zhou, L.-Y. (2009). A trilateral trade model with the duopoly feature: Experimental simulation analysis of state welfare based on world oil market [j]. *Journal of Finance and Economics*, **7**.

Li, L. and Zhang, X. (2017). Parsimonious Tensor Response Regression. *Journal of the American Statistical Association*, **112**(519), 1131–1146.

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

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

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, **108**(504), 1339–1349.

Rabusseau, G. and Kadri, H. (2016). Higher-order low-rank regression. *arXiv preprint arXiv:1602.06863*.

Razumnikova, O. M. (2007). Creativity related cortex activity in the remote associates task. *Brain Research Bulletin*, **73**(1), 96–102.

Rothman, A. J., Levina, E., and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, **19**(4), 947–962.

Roy, A., Ghosal, S., Prescott, J., and Choudhury, K. R. (2017). Bayesian modeling of the structural connectome for studying alzheimer disease. *arXiv preprint arXiv:1710.04560*.

Song, Q. and Liang, F. (2017). Nearly optimal bayesian shrinkage for high dimensional regression. *arXiv preprint arXiv:1712.08964*.

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

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

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

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

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

Supplementary Material: Bayesian Generalized Sparse Symmetric Tensor-on-Vector Regression

Sharmistha Guha*‡ Rajarshi Guhaniyogi*§

May 16, 2020

The supplementary material consists of four sections. Section 1 presents results on posterior convergence for the proposed SGTM model. Section 2 presents proofs of the theoretical results mentioned in Section 1. Section 3 presents simulation results for SGTM along with its competitors when the response is a two dimensional binary symmetric tensor. The simulation results for the two dimensional continuous symmetric tensor response have been presented in Section 4.1 of the main article. Section 4 discusses convergence diagnostics of the posterior distribution for SGTM. Section 5 adds results for two simulation cases representing weak signal strength. Finally, Section 6 outlines the full conditional distributions of the parameters in SGTM for implementing MCMC.

1 Convergence Rate for Predictive Densities

This section presents posterior convergence properties of the proposed symmetric generalized tensor model (SGTM). We adopt the framework outlined in Jiang (2007), with some

‡Sharmistha Guha, Postdoctoral Associate, Department of Statistical Science, Duke University (E-mail: sg516@duke.edu).

§Rajarshi Guhaniyogi, Assistant Professor, Department of Statistics, SOE2, UC Santa Cruz, 1156 High Street, Santa Cruz, CA 95064 (E-mail: rguhaniy@ucsc.edu).

important differences. While Jiang (2007) studies the convergence rate of the posterior predictive distribution with a scalar response and vector predictors, we focus on a symmetric tensor response and vector predictors. The novel model development and the prior structure described in Section 2 of the main article present theoretical challenges which are unique and very different from Jiang (2007).

Let $f^*(\boldsymbol{Y}|\boldsymbol{x})$ be the true conditional density of $\boldsymbol{Y}$ given $\boldsymbol{x}$ and $f(\boldsymbol{Y}|\boldsymbol{x})$ be the random predictive density for which we obtain a posterior. Define an integrated Hellinger distance between f^* and f as $\mathcal{D}_H(f, f^*) = \sqrt{\int\int(\sqrt{f(\boldsymbol{Y}|\boldsymbol{x})} - \sqrt{f^*(\boldsymbol{Y}|\boldsymbol{x})})^2 \nu_{\boldsymbol{Y}}(d\boldsymbol{Y})\nu_{\boldsymbol{x}}(d\boldsymbol{x})}$, where $\nu_{\boldsymbol{x}}$ is the unknown probability measure for $\boldsymbol{x}$ and $\nu_{\boldsymbol{Y}}$ is the dominating measure for f and f^*. We focus on showing $E_{f^*}\Pi[\mathcal{D}_H(f, f^*) > \epsilon_n|\{\boldsymbol{Y}_i, \boldsymbol{x}_i\}_{i=1}^n] < \kappa_n$, for large n, for some sequences ϵ_n, κ_n converging to 0 as $n \to \infty$, where $\Pi(\mathcal{A}|\{\boldsymbol{Y}_i, \boldsymbol{x}_i\}_{i=1}^n)$ is the posterior probability of the set $\mathcal{A}$. The result implies that the posterior probability outside a shrinking neighborhood around the true predictive density f^* converges to 0 as $n \to \infty$. In particular, we seek to establish a convergence rate ϵ_n of order close to the parametric optimal rate of $n^{-1/2}$ upto a $\log(n)$ factor.

1.1 Framework and Main Results

In what follows, we assume $m = 1$ predictor of interest (hence get rid of the subscript h for all parameters) and no auxiliary predictor for simplifying calculations, though the results assume straightforward extension to cases where $m > 1$ and $l > 1$. Without loss of generality, the predictor x satisfies $|x_i| < 1$ for all i. Let p_n denote the number of nodes and R_n denote the rank of CP/PARAFAC decomposition for $\boldsymbol{B}$ in presence of sample size n. We assume that p_n and R_n are both non-decreasing functions of n, with $R_n < p_n$ for all large n. Hence, the number of elements in $\mathcal{J}$, given by $q_n = p_n(p_n - 1)...(p_n - D + 1)/D!$, naturally becomes a function of n. This paradigm attempts to capture the fact that q_n grows much faster than

n, and a higher rank CP decomposition of $\boldsymbol{B}$ can be estimated more precisely in presence of a larger sample size n. The true density and the predictive density of the fitted model are assumed to lie in the class of generalized linear models given by

$$f(\boldsymbol{Y}|x) = \prod_{\boldsymbol{j}\in\mathcal{J}} f_{\boldsymbol{j}}(Y_{\boldsymbol{j}}|x),\ f_{\boldsymbol{j}}(Y_{\boldsymbol{j}}|x) = \exp(a(\alpha_{\boldsymbol{j}})Y_{\boldsymbol{j}} + b(\alpha_{\boldsymbol{j}}) + c(Y_{\boldsymbol{j}})),\ \alpha_{\boldsymbol{j}} = xB_{\boldsymbol{j}}$$
$$f^*(\boldsymbol{Y}|x) = \prod_{\boldsymbol{j}\in\mathcal{J}} f^*_{\boldsymbol{j}}(Y_{\boldsymbol{j}}|x),\ f^*_{\boldsymbol{j}}(Y_{\boldsymbol{j}}|x) = \exp(a(\alpha^*_{\boldsymbol{j}})Y_{\boldsymbol{j}} + b(\alpha^*_{\boldsymbol{j}}) + c(Y_{\boldsymbol{j}})),\ \alpha^*_{\boldsymbol{j}} = xB^*_{\boldsymbol{j}}, \qquad (1)$$

where $a(w)$ and $b(w)$ are continuously differentiable functions, with $a(w)$ having a nonzero derivative. This parametrization includes some popular classes of densities, e.g., for the binary logit or probit link and a continuous response with i.i.d. normal errors having known variance. Similar to $\boldsymbol{B}$, the true tensor coefficient $\boldsymbol{B}^*$ (having the $\boldsymbol{j}$th cell as $B^*_{\boldsymbol{j}}$, $\boldsymbol{j} \in \mathcal{J}$) also assumes symmetric tensor decomposition with rank R_n, i.e., $B^*_{\boldsymbol{j}} = \sum_{r=1}^{R_n} u^{*(r)}_{j_1}...u^{*(r)}_{j_D}$, for $\boldsymbol{j} \in \mathcal{J}$. Although this is a somewhat restrictive assumption, it has been frequently employed in earlier theoretical literature on tensor regressions for simplifying calculations (Guhaniyogi *et al.*, 2017; Guhaniyogi and Spencer, 2018). Further, we assume $\boldsymbol{M} = \boldsymbol{I}$ for simplifying calculations and $\lambda_r = 1$ for all r, since no rank selection is required due to the assumption of the true rank being equal to the fitted rank. Finally, for two sequences c_n and d_n, $c_n \prec d_n$ signifies $c_n/d_n \to 0$ as $n \to \infty$. With these notations, we state the following theorem, the proof of which can be found in the next section (i.e., Section 2 of this supplementary material).

Theorem 1.1 *Define* $G(\Delta) = 1 + \Delta \sup_{|w|\leq\Delta} |a'(w)| \sup_{|w|\leq\Delta} |b'(w)/a'(w)|$, *where* $a'(w)$ *and* $b'(w)$ *are derivatives of functions* $a(w)$ *and* $b(w)$, *respectively. For a sequence* ϵ_n *satisfying* $0 < \epsilon_n < 1$ *and* $n\epsilon_n^2 \to \infty$ *and another sequence* C_n, *let the following conditions hold*

(i) $R_n p_n \log(p_n) \prec n\epsilon_n^2$

(ii) $R_n p_n \log(1/\epsilon_n^2) \prec n\epsilon_n^2$

(iii) $R_n p_n \log(G(R_n C_n^D)) \prec n\epsilon_n^2$,

(iv) $(1 - \Phi(C_n)) \leq e^{-4n\epsilon_n^2}$, *for all large* n

(v) $\lim_{n\to\infty} \sum_{k=1}^{p_n} ||\tilde{\boldsymbol{u}}_k^*|| < \infty$, *where* $\tilde{\boldsymbol{u}}_k^* = (u_k^{*(1)}, ..., u_k^{*(R_n)})^T$.

Then, $E_{f^*}\Pi\{\mathcal{D}_H(f, f^*) > 4\epsilon_n | \{\boldsymbol{Y}_i, x_i\}_{i=1}^n\} < 4e^{-n\epsilon_n^2}$, *for all large* n.

As mentioned in Jiang (2007), $G(\Delta)$ grows at most in the order of $|\Delta|^2$ for regression involving i.i.d. normal errors and for binary probit regression. Also $G(\Delta)$ grows at most linearly with $|\Delta|$ for binary logistic regression. For our theoretical exposition, we will focus on continuous and binary regression only. Theorem 1.1, together with the mentioned functional properties of $G(\Delta)$, leads to the following result on the convergence rate ϵ_n of the proposed model.

Corollary 1.2 *Assume that the link function* $H(\cdot)$ *in equation (1) of the main article is a logit link, or a probit link in a binary regression, or an identity link in a continuous regression with i.i.d. normal errors having a known variance. Let,* $\lim_{n\to\infty} \sum_{k=1}^{p_n} ||\tilde{\boldsymbol{u}}_k^*|| < \infty$, *where* $\tilde{\boldsymbol{u}}_k^* = (u_k^{*(1)}, ..., u_k^{*(R_n)})^T$. *Further assume* p_n *grows at a rate slower than* n^θ, $\theta < 1$ *(i.e.* $p_n \leq C_1^* n^\theta$*) and the tensor rank* R_n *grows at a much slower rate of* $(\log n)^{k_1}$ *for some* k_1 *(i.e.* $R_n \leq C_2^* (\log n)^{k_1}$*). Choose* C_n *such that* $n^{\mu_1} \leq C_n \leq n^{\mu_2}$, *for some* μ_1, μ_2 *satisfying* $\theta/2 < \mu_1 < \mu_2$. *Then, the convergence rate* ϵ_n *can be taken as* $\epsilon_n \sim n^{-(1-\theta)/2} (\log n)^{k_1/2+1}$.

It is evident that the convergence rate is a function of how the number of tensor nodes and the rank of the true tensor (same as the rank of the fitted tensor) grow with n. Intuitively, p_n should grow at a much faster rate than R_n, and both should be bounded by an appropriate function of n to achieve a good convergence rate. Finally, it is worth noting that the condition $\lim_{n\to\infty} \sum_{k=1}^{p_n} ||\tilde{\boldsymbol{u}}_k^*|| < \infty$ includes as a special case the scenario where only a fixed and finite

number of $||\tilde{\boldsymbol{u}}_k^*||$'s are nonzero, while also allowing perhaps a setup with many small $||\tilde{\boldsymbol{u}}_k^*||$'s, none of which are exactly zero.

2 Proofs of Theoretical Results in Section 1 of the Supplementary Material

Lemma 2.1 *Let* $\tilde{\boldsymbol{u}}_k^* = (u_k^{*(1)}, ..., u_k^{*(R_n)})^T$ *and* $\gamma_{\boldsymbol{j},n}$, $\boldsymbol{j} \in \mathcal{J}$ *be the only positive root of the equation*

$$x\prod_{s=2}^{D}(x+||\tilde{\boldsymbol{u}}_{j_s}^*||)+||\tilde{\boldsymbol{u}}_{j_1}^*||x\prod_{s=3}^{D}(x+||\tilde{\boldsymbol{u}}_{j_s}^*||)+\cdots+x\prod_{s=1}^{D-1}||\tilde{\boldsymbol{u}}_{j_s}^*||=\delta_n \tag{2}$$

Assume $\gamma_n = min_{\boldsymbol{j}\in\mathcal{J}}\,\gamma_{\boldsymbol{j},n}$. *Then,* $\Pi(||\boldsymbol{B}-\boldsymbol{B}^*||_\infty \le \delta_n) \ge \Pi(||\tilde{\boldsymbol{u}}_k - \tilde{\boldsymbol{u}}_k^*|| \le \gamma_n, k = 1, .., p_n)$.

Proof for $\boldsymbol{j} \in \mathcal{J}$,
$|B_{\boldsymbol{j}} - B_{\boldsymbol{j}}^*| = |\sum_{r=1}^{R_n} u_{j_1}^{(r)}\cdots u_{j_D}^{(r)} - \sum_{r=1}^{R_n} u_{j_1}^{*(r)}\cdots u_{j_D}^{*(r)}| = |\sum_{r=1}^{R_n}(u_{j_1}^{(r)} - u_{j_1}^{*(r)})\prod_{s=2}^{D} u_{j_s}^{(r)}| + \cdots + |\sum_{r=1}^{R_n}(u_{j_D}^{(r)} - u_{j_D}^{*(r)})\prod_{s=1}^{D-1} u_{j_s}^{*(r)}| \le ||\tilde{\boldsymbol{u}}_{j_1} - \tilde{\boldsymbol{u}}_{j_1}^*||\prod_{s=2}^{D}||\tilde{\boldsymbol{u}}_{j_s}|| + \cdots + ||\tilde{\boldsymbol{u}}_{j_D} - \tilde{\boldsymbol{u}}_{j_D}^*||\prod_{s=1}^{D-1}||\tilde{\boldsymbol{u}}_{j_s}^*|| \le ||\tilde{\boldsymbol{u}}_{j_1} - \tilde{\boldsymbol{u}}_{j_1}^*||\prod_{s=2}^{D}(||\tilde{\boldsymbol{u}}_{j_s} - \tilde{\boldsymbol{u}}_{j_s}^*|| + ||\tilde{\boldsymbol{u}}_{j_s}^*||) + \cdots + ||\tilde{\boldsymbol{u}}_{j_D} - \tilde{\boldsymbol{u}}_{j_D}^*||\prod_{s=1}^{D-1}||\tilde{\boldsymbol{u}}_{j_s}^*||$.

If $||\tilde{\boldsymbol{u}}_k - \tilde{\boldsymbol{u}}_k^*|| \le \gamma_n, k = 1, .., p_n$, the above inequality implies that $|B_{\boldsymbol{j}} - B_{\boldsymbol{j}}^*| \le \gamma_n\prod_{s=2}^{D}(\gamma_n + ||\tilde{\boldsymbol{u}}_{j_s}^*||) + \cdots + \gamma_n\prod_{s=1}^{D-1}||\tilde{\boldsymbol{u}}_{j_s}^*|| \le \delta_n$.
Thus $\Pi(||\boldsymbol{B}-\boldsymbol{B}^*||_\infty \le \delta_n) \ge \Pi(||\tilde{\boldsymbol{u}}_k - \tilde{\boldsymbol{u}}_k^*|| \le \gamma_n,\ k = 1, .., p_n)$.

Lemma 2.2 *With* γ_n *and* $\tilde{\boldsymbol{u}}_k^*$ *defined as in Lemma 2.1,*

$$\Pi(||\boldsymbol{B}-\boldsymbol{B}^*||_\infty \le \delta_n) \ge e^{-\sum_{k=1}^{p_n}||\tilde{\boldsymbol{u}}_k^*||^2/2}(1/\sqrt{2\pi})^{R_n p_n}\frac{R_n p_n}{R_n p_n + 1}\left(2\gamma_n/R_n\right)^{R_n p_n} e^{-p_n\gamma_n^2/R_n}.$$

Proof

$$\Pi(||\boldsymbol{B}-\boldsymbol{B}^*||_\infty \le \delta_n) \ge \Pi(||\tilde{\boldsymbol{u}}_k - \tilde{\boldsymbol{u}}_k^*|| \le \gamma_n,\ k=1,..,p_n) \ge E\left[\Pi(||\tilde{\boldsymbol{u}}_k - \tilde{\boldsymbol{u}}_k^*|| \le \gamma_n,\ k=1,..,p_n|\xi)\right]$$
$$\ge E\left[\prod_{k=1}^{p_n}\left\{e^{-||\tilde{\boldsymbol{u}}_k^*||^2/2}\Pi(||\tilde{\boldsymbol{u}}_k|| \le \gamma_n|\xi)\right\}\right] = e^{-\sum_{k=1}^{p_n}||\tilde{\boldsymbol{u}}_k^*||^2/2} E\left[\prod_{k=1}^{p_n}\Pi(||\tilde{\boldsymbol{u}}_k|| \le \gamma_n|\xi)\right], \qquad (3)$$

where the first inequality follows from Lemma 2.1 and the second inequality follows from the Anderson's Lemma. We will now make use of the fact that $\int_{-a}^{a} e^{-x^2/2}dx \ge e^{-a^2}2a$ to conclude

$$\Pi(||\tilde{\boldsymbol{u}}_k|| \le \gamma_n|\xi) \ge \prod_{r=1}^{R_n}\Pi(|u_k^{(r)}| \le \frac{\gamma_n}{R_n}|\xi) = \prod_{r=1}^{R_n}\left((1-\xi) + \left(\frac{\xi}{\sqrt{2\pi}}\right)\int_{-\gamma_n/R_n}^{\gamma_n/R_n} e^{-x^2/2}\right)$$
$$\ge \prod_{r=1}^{R_n}\left((1-\xi) + \left(\frac{\xi}{\sqrt{2\pi}}\right)e^{-\gamma_n^2/R_n^2}\frac{2\gamma_n}{R_n}\right) \ge \left[(1-\xi) + \frac{\xi}{\sqrt{2\pi}}e^{-\gamma_n^2/R_n^2}\frac{2\gamma_n}{R_n}\right]^{R_n}.$$

$$\prod_{k=1}^{p_n}\Pi(||\tilde{\boldsymbol{u}}_k|| \le \gamma_n) \ge E\left[(1-\xi) + \frac{\xi}{\sqrt{2\pi}}\exp\left(-\frac{\gamma_n^2}{R_n^2}\right)\frac{2\gamma_n}{R_n}\right]^{R_np_n}$$
$$= E\left[\sum_{h_1=0}^{R_np_n}\binom{R_np_n}{h_1}(1-\xi)^{h_1}\left(\frac{\xi}{\sqrt{2\pi}}\right)^{R_np_n-h_1}\left(\frac{2\gamma_n}{R_n}\right)^{R_np_n-h_1}\exp(-\frac{(R_np_n-h_1)\gamma_n^2}{R_n^2})\right]$$
$$\ge \left(\frac{1}{\sqrt{2\pi}}\right)^{R_np_n}\sum_{h_1=0}^{R_np_n}\binom{R_np_n}{h_1}Beta(R_np_n-h_1+1,h_1+1)\left(\frac{2\gamma_n}{R_n}\right)^{R_np_n-h_1}\exp\left(-\frac{(R_np_n-h_1)\gamma_n^2}{R_n^2}\right)$$
$$\ge \left(\frac{1}{\sqrt{2\pi}}\right)^{R_np_n}\sum_{h_1=0}^{R_np_n}\frac{(R_np_n)!}{h_1!(R_np_n-h_1)!}\frac{h_1!(R_np_n-h_1)!}{(R_np_n+1)!}\left(\frac{2\gamma_n}{R_n}\right)^{R_np_n-h_1}\exp\left(-\frac{(R_np_n-h_1)\gamma_n^2}{R_n^2}\right)$$
$$\ge \left(\frac{1}{\sqrt{2\pi}}\right)^{R_np_n}\frac{R_np_n}{R_np_n+1}\left(\frac{2\gamma_n}{R_n}\right)^{R_np_n}\exp\left(-\frac{p_n\gamma_n^2}{R_n}\right).$$

Thus, $\Pi(||\boldsymbol{B}-\boldsymbol{B}^*||_\infty \le \delta_n) \ge \exp\left(-\frac{\sum_{k=1}^{p_n}||\tilde{\boldsymbol{u}}_k^*||^2}{2}\right)\left(\frac{1}{\sqrt{2\pi}}\right)^{R_np_n}\frac{R_np_n}{R_np_n+1}\left(\frac{2\gamma_n}{R_n}\right)^{R_np_n}\exp\left(-\frac{p_n\gamma_n^2}{R_n}\right)$

Lemma 2.3 *Let x^* be a real positive root of the equation $P(x) = x^D + a_{D-1}x^{D-1} + \cdots + a_1x - a_0 = 0$ with $a_0, a_1, ..., a_{D-1} \ge 0$. Then $\frac{1}{x^*} \le 1 + \frac{a_1}{a_0}$.*

Proof Using a change of variable $x_1 = \frac{1}{x}$, we have $x_1^D - \frac{a_1}{a_0}x_1^{D-1} - \cdots - \frac{a_{D-1}}{a_0}x - \frac{1}{a_0} = 0$. Since this is a monic polynomial with $\frac{1}{x^*}$ as one of its positive real roots, by Lagrange-Maclaurin theorem $\frac{1}{x^*} \leq 1 + \frac{a_1}{a_0}$.

Proof of Theorem 1.1

Proof Define,

$$\mathcal{D}_0(f, f^*) = \int\int f^*(\boldsymbol{Y}|x) \log(f^*(\boldsymbol{Y}|x)/f(\boldsymbol{Y}|x))\nu_{\boldsymbol{Y}}(d\boldsymbol{Y})\nu_x(dx),$$
$$\mathcal{D}_t(f, f^*) = (1/t)\left\{\int\int f^*(\boldsymbol{Y}|x)(f^*(\boldsymbol{Y}|x)/f(\boldsymbol{Y}|x))^t\nu_{\boldsymbol{Y}}(d\boldsymbol{Y})\nu_x(dx)\right\}.$$

Let $\mathcal{P}_n$ be the sequence of sets of probability densities and $\mathcal{F}_n(\epsilon_n, \mathcal{P}_n)$ be the minimum number of Hellinger balls of radius ϵ_n needed to cover $\mathcal{P}_n$. Invoking Proposition 1 in Jiang (2007), it suffices to show that the following conditions hold for sufficiently large n to prove Theorem 1.1:
(a) $\log \mathcal{F}_n(\epsilon_n, \mathcal{P}_n) \leq n\epsilon_n^2$, (b) $\Pi(\mathcal{P}_n^c) \leq e^{-2n\epsilon_n^2}$, (c) $\Pi[f : \mathcal{D}_t(f, f^*) \leq \epsilon_n^2/4] \geq e^{-n\epsilon_n^2/4}$, with $t = 1$.

Proof of condition (b): Define $\mathcal{P}_n$ as the set of all densities f s.t. $|u_k^{(r)}| \leq C_n$, for all $k \in \mathcal{N}$ and $r = 1, ..., R_n$. Then for all large n,

$$\Pi(\mathcal{P}_n^c) = \Pi(\cup_{k=1}^{p_n} \cup_{r=1}^{R_n} \{|u_k^{(r)}| > C_n\}) \leq R_n p_n \Pi(|u_k^{(r)}| > C_n) = 2R_n p_n(1 - \Phi(C_n)) \leq e^{-2n\epsilon_n^2},$$

where the last inequality follows by assumptions (i) and (iv).

Proof of condition (a): Let us consider balls of the form $(u_k^{(r)} - \rho, u_k^{(r)} + \rho)_{k,r=1}^{p_n,R_n}$ with their centers $|u_k^{(r)}| \leq C_n$, i.e., the densities f defined through parameters $u_k^{(r)}$'s belong to $\mathcal{P}_n$.

There are at most $F(\rho) = (C_n/\rho + 1)^{R_n p_n}$ such balls needed to cover the parameter space $\{u_k^{(r)} : k = 1, .., p_n; r = 1, .., R_n, |u_k^{(r)}| \leq C_n\}$.

Let $\tilde{f}$ be any density in $\mathcal{P}_n$, with $\tilde{f}(\boldsymbol{Y}|x) = \prod_{\boldsymbol{j}\in\mathcal{J}} \tilde{f}_{\boldsymbol{j}}(Y_{\boldsymbol{j}}|x)$, $\tilde{f}_{\boldsymbol{j}}(Y_{\boldsymbol{j}}|x) = \exp(a(\tilde{\alpha}_{\boldsymbol{j}})Y_{\boldsymbol{j}} + b(\tilde{\alpha}_{\boldsymbol{j}}) + c(Y_{\boldsymbol{j}}))$, $\tilde{\alpha}_{\boldsymbol{j}} = x\tilde{B}_{\boldsymbol{j}}$, where $\tilde{B}_{\boldsymbol{j}} = \sum_{r=1}^{R_n} v_{j_1}^{(r)} ... v_{j_D}^{(r)}$, with $|v_k^{(r)}| \leq C_n$ for all $k \in \mathcal{N}$, $r = 1, .., R_n$. There exists a density $f \in \mathcal{P}_n$ represented by parameters $u_k^{(r)}$'s such that $v_k^{(r)} \in (u_k^{(r)} - \rho, u_k^{(r)} + \rho)$ for every r and k. Note that,

$$\mathcal{D}_H(f, \tilde{f}) \leq \left\{\mathcal{D}_0(f, \tilde{f})\right\}^{1/2} = \left\{\sum_{\boldsymbol{j}\in\mathcal{J}} \mathcal{D}_0(f_{\boldsymbol{j}}, \tilde{f}_{\boldsymbol{j}})\right\}^{1/2}.$$

One can apply Taylor expansion on $\mathcal{D}_0(f_{\boldsymbol{j}}, \tilde{f}_{\boldsymbol{j}})$ to show that $\mathcal{D}_0(f_{\boldsymbol{j}}, \tilde{f}_{\boldsymbol{j}}) \leq E_x(a'(\bar{\alpha}_{\boldsymbol{j}})(-b'(\alpha_{\boldsymbol{j}})/a'(\alpha_{\boldsymbol{j}})) + b'(\bar{\alpha}_{\boldsymbol{j}}))(\alpha_{\boldsymbol{j}} - \tilde{\alpha}_{\boldsymbol{j}})$, where $\bar{\alpha}_{\boldsymbol{j}}$ is an intermediate point between $\alpha_{\boldsymbol{j}}$ and $\tilde{\alpha}_{\boldsymbol{j}}$. Now note that,

$$\begin{aligned}
&|\alpha_{\boldsymbol{j}} - \tilde{\alpha}_{\boldsymbol{j}}| \leq |B_{\boldsymbol{j}} - \tilde{B}_{\boldsymbol{j}}| = |\sum_{r=1}^{R_n} u_{j_1}^{(r)} ... u_{j_D}^{(r)} - \sum_{r=1}^{R_n} v_{j_1}^{(r)} ... v_{j_D}^{(r)}| \\
&\leq \sum_{r=1}^{R_n} \left\{|u_{j_1}^{(r)} - v_{j_1}^{(r)}| \prod_{l=2}^{D} |u_{j_l}^{(r)}| + |v_{j_1}^{(r)}||u_{j_2}^{(r)} - v_{j_2}^{(r)}| \prod_{l=3}^{D} |u_{j_l}^{(r)}| + \cdots + \prod_{l=1}^{D-1} |v_{j_l}^{(r)}||u_{j_2}^{(r)} - v_{j_2}^{(r)}|\right\} \\
&\leq R_n \rho C_n^{D-1}.
\end{aligned}$$

Similar calculations lead to $|\alpha_{\boldsymbol{j}}|$, $|\tilde{\alpha}_{\boldsymbol{j}}|$ (and therefore $|\bar{\alpha}_{\boldsymbol{j}}|$) being bounded by $R_n C_n^D$. Hence,

$$\mathcal{D}_H(f, \tilde{f}) \leq \left\{\sum_{\boldsymbol{j}\in\mathcal{J}} \mathcal{D}_0(f_{\boldsymbol{j}}, \tilde{f}_{\boldsymbol{j}})\right\}^{1/2} \leq \{2q_n \sup_{|w|\leq R_n C_n^D} |a'(w)| \sup_{|w|\leq R_n C_n^D} |b'(w)/a'(w)| \rho R_n C_n^{D-1}\}^{1/2}.$$

Choosing $\rho = \epsilon_n^2/(2q_n \sup_{|w|\leq R_n C_n^D} |a'(w)| \sup_{|w|\leq R_n C_n^D} |b'(w)/a'(w)| R_n C_n^{D-1})$, one gets $\mathcal{D}_H(f, \tilde{f}) \leq$

ϵ_n. Hence

$$\log \mathcal{F}_n(\epsilon_n, \mathcal{P}_n) \leq \log F(\rho) \leq R_n p_n \log\left(1 + 2q_n \sup_{|w| \leq R_n C_n^D} |a'(w)| \sup_{|w| \leq R_n C_n^D} |b'(w)/a'(w)| R_n C_n^D/\epsilon_n^2\right)$$

$$\leq R_n p_n \log(2q_n/\epsilon_n^2) + R_n p_n \log(G(R_n C_n^D))$$

$$\leq D R_n p_n \log(2p_n) + R_n p_n \log(1/\epsilon_n^2) + R_n p_n \log(G(R_n C_n^D))$$

$$\leq n\epsilon_n^2, \text{ for large } n, \text{ by assumptions (i)-(iii).}$$

Proof of Condition (c): Take $t = 1$. By mean value theorem, $\exists\, \boldsymbol{\kappa}$ s.t. $\mathcal{D}_t(f, f^*) = E_x\left\{\boldsymbol{g}(\boldsymbol{\kappa})^T(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)\right\}$, where $\boldsymbol{g}$ represents the continuous derivative function of f in the neighborhood of f^*. Let $\delta_n = \epsilon_n^2/q_n$. If for each $\boldsymbol{j} \in \mathcal{J}$, $B_{\boldsymbol{j}} \in (B_{\boldsymbol{j}}^* - \delta_n, B_{\boldsymbol{j}}^* + \delta_n)$, then $||\boldsymbol{\alpha} - \boldsymbol{\alpha}^*|| \leq \sum_{\boldsymbol{j} \in \mathcal{J}} |\alpha_{\boldsymbol{j}} - \alpha_{\boldsymbol{j}}^*| \leq \sum_{\boldsymbol{j} \in \mathcal{J}} |B_{\boldsymbol{j}} - B_{\boldsymbol{j}}^*| \leq q_n \delta_n \leq \epsilon_n^2$, for large n. Again, $||\boldsymbol{\kappa}|| \leq ||\boldsymbol{\alpha} - \boldsymbol{\alpha}^*|| + ||\boldsymbol{\alpha}^*|| \leq q_n\delta_n + m_n = \epsilon_n^2 + m_n$, where $m_n = ||\boldsymbol{B}^*|| \leq \sum_{\boldsymbol{j} \in \mathcal{J}} ||\tilde{\boldsymbol{u}}_{j_1}^*|| \cdots ||\tilde{\boldsymbol{u}}_{j_D}^*|| \leq (\sum_{k=1}^{p_n} ||\tilde{\boldsymbol{u}}_k^*||)^D$, which is bounded by assumption (v), for sufficiently large n. Hence $||\boldsymbol{g}(\boldsymbol{\kappa})||$ is bounded for sufficiently large n. Thus, $\mathcal{D}_t(f, f^*) = E_x\left\{\boldsymbol{g}(\boldsymbol{\kappa})^T(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)\right\} \leq \Delta_2 q_n \delta_n \leq \epsilon_n^2/4$ for large n, for some constant Δ_2.

This implies that $\Pi(\{f : \mathcal{D}_t(f, f^*) \leq \epsilon_n^2/4\}) \geq \Pi(\{\boldsymbol{B} : B_{\boldsymbol{j}} \in (B_{\boldsymbol{j}}^* - \delta_n, B_{\boldsymbol{j}}^* + \delta_n), \forall\, \boldsymbol{j} \in \mathcal{J}\})$. By Lemma 2.2, $-\log \Pi(\{\boldsymbol{B} : B_{\boldsymbol{j}} \in (B_{\boldsymbol{j}}^* - \delta_n, B_{\boldsymbol{j}}^* + \delta_n), \forall\, \boldsymbol{j} \in \mathcal{J}\}) = -\log \Pi(||\boldsymbol{B} - \boldsymbol{B}^*||_\infty \leq \delta_n) \leq \sum_{k=1}^{p_n} ||\tilde{\boldsymbol{u}}_k^*||^2/2 + (R_n p_n/2)\log(2\pi) + \log(1 + (1/(R_n p_n))) + R_n p_n \log(R_n) + R_n p_n \log(1/\gamma_n) + p_n\gamma_n^2/R_n$.

Since $||\tilde{\boldsymbol{u}}_k^*|| \geq 0$, $\sum_{k=1}^{p_n} ||\tilde{\boldsymbol{u}}_k^*||^2 \leq (\sum_{k=1}^{p_n} ||\tilde{\boldsymbol{u}}_k^*||)^2$ is bounded for large n, by assumption (v). By assumption (i), $R_n p_n \log(R_n) \prec n\epsilon_n^2$ (hence $R_n p_n \prec n\epsilon_n^2$). Using Lagrange-Maclaurin bound on the positive root of a monic polynomial of degree D, we have $\gamma_n \leq 1 + \delta_n^{1/D}$, implying $p_n\gamma_n^2/R_n \prec n\epsilon_n^2$, for all large n, by assumption (i). Using Lemma 2.1 and 2.3, $1/\gamma_n \leq (\sum_{k=1}^{p_n} ||\tilde{\boldsymbol{u}}_k^*||)^D/\delta_n + 1$. If $m_0 = \lim_{n\to\infty} \sum_{k=1}^{p_n} ||\tilde{\boldsymbol{u}}_k^*||$, then $R_n p_n \log(1/\gamma_n) \leq R_n p_n \log(m_0^D/\delta_n) = D R_n p_n \log(m_0) + R_n p_n \log(q_n) + R_n p_n \log(1/\epsilon_n^2) \leq D R_n p_n \log(m_0) +$

$DR_n p_n \log(p_n) + R_n p_n \log(1/\epsilon_n^2) \prec n\epsilon_n^2$, by assumptions (i) and (ii).

All the aforementioned calculations yield $-\log \Pi(||\boldsymbol{B} - \boldsymbol{B}^*||_\infty \leq \delta_n) \leq n\epsilon_n^2/4$, for all large n, which implies $\Pi(\{f : \mathcal{D}_t(f, f^*) \leq \epsilon_n^2/4\}) \geq e^{-n\epsilon_n^2/4}$ for all large n. This concludes the proof.

3 2D Symmetric Binary Tensor Response Example

This section presents a special case in which the response is an undirected network with p nodes, expressed in the form of a symmetric $p \times p$ matrix having entries 0 or 1. A value of 1 in the (j_1, j_2)th entry signifies an edge between the j_1th and the j_2th nodes. We run SGTM along with LS in the four cases under Simulation 1 and 7 cases under Simulation 2 described in the main article. Referring to Table 1, SGTM yields a vastly superior performance to LS in terms of point estimation and uncertainties under Scenario 1, where theoretical guarantee exists for SGTM. However, the difference in performance becomes less stark under Scenario 2. Under Cases 1 and 3 with a smaller p and higher node and cell sparsities, SGTM outperforms LS in terms of point estimation. Under Cases 2 and 4 with a higher p/n ratio and lower node sparsity, SGTM has a marginal edge over LS. When the p/n ratio is small, SGTM continues to enjoy superior point estimation of $\boldsymbol{B}$ over LS, even after decreasing node sparsity (comparing Cases 1 & 6) or cell sparsity (comparing Cases 1 & 5). However, the gap between the performances of SGTM and LS narrows when either node or cell sparsity decreases. In fact, in presence of both low node sparsity and low cell sparsity (Case 7), LS becomes competitive with SGTM.

The uncertainty quantified by SGTM is more precise than LS, with SGTM delivering much narrower credible intervals with a similar coverage in Scenario 1. In Cases 2 & 4 under Scenario 2, both competitors suffer from under-coverage, perhaps due to the larger p/n ratio. In all other cases, SGTM yields either nominal or close to the nominal coverage. LS yields

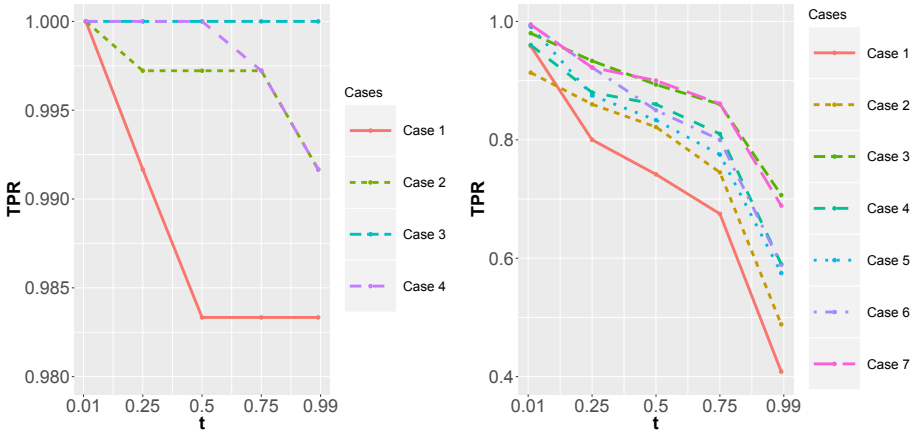
Scenario 1						
	MSE		Avg. Coverage of 95% CI		Avg. length of 95% CI	
Cases	SGTM	LS	SGTM	LS	SGTM	LS
1	$\mathbf{0.02}_{0.01}$	$0.44_{0.18}$	$0.98_{0.00}$	$0.98_{0.00}$	$0.15_{0.01}$	$2.57_{0.00}$
2	$\mathbf{0.01}_{0.00}$	$0.24_{0.07}$	$0.99_{0.00}$	$0.99_{0.00}$	$0.18_{0.02}$	$3.37_{0.00}$
3	$\mathbf{0.01}_{0.00}$	$0.18_{0.10}$	$0.97_{0.01}$	$0.99_{0.00}$	$0.18_{0.02}$	$3.16_{0.00}$
4	$\mathbf{0.01}_{0.00}$	$0.14_{0.05}$	$0.99_{0.00}$	$1.00_{0.00}$	$0.23_{0.02}$	$3.28_{0.00}$
Scenario 2						
	MSE		Avg. Coverage of 95% CI		Avg. length of 95% CI	
Cases	SGTM	LS	SGTM	LS	SGTM	LS
1	$\mathbf{0.40}_{0.10}$	$1.16_{0.56}$	$0.96_{0.00}$	$0.97_{0.01}$	$0.15_{0.02}$	$1.47_{0.00}$
2	$\mathbf{0.27}_{0.06}$	$0.33_{0.08}$	$0.68_{0.02}$	$0.73_{0.03}$	$0.82_{0.01}$	$1.09_{0.00}$
3	$\mathbf{0.34}_{0.06}$	$0.52_{0.07}$	$0.94_{0.01}$	$0.97_{0.00}$	$0.20_{0.03}$	$1.54_{0.00}$
4	$\mathbf{0.26}_{0.03}$	$0.28_{0.03}$	$0.67_{0.01}$	$0.70_{0.02}$	$0.79_{0.02}$	$1.18_{0.00}$
5	$\mathbf{0.36}_{0.09}$	$0.90_{0.26}$	$0.95_{0.00}$	$0.97_{0.00}$	$0.15_{0.02}$	$1.70_{0.00}$
6	$\mathbf{0.44}_{0.07}$	$0.62_{0.08}$	$0.90_{0.00}$	$0.98_{0.02}$	$0.28_{0.03}$	$1.90_{0.00}$
7	$\mathbf{0.45}_{0.08}$	$0.48_{0.08}$	$0.88_{0.01}$	$0.98_{0.00}$	$0.28_{0.04}$	$2.10_{0.00}$

Table 1: Mean Squared Error (MSE), average coverage and average length of 95% credible intervals for SGTM and LS are presented for various simulation scenarios for the binary regression case. The lowest MSE in each case in boldfaced. All results are averaged over 50 replications.

over coverage with considerably wider 95% CIs compared to SGTM.

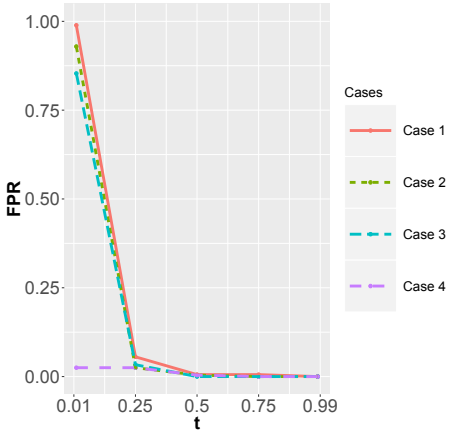
In terms of identifying nodes significantly associated with the predictor of interest, we observe TPR values close to 1 and FPR values close to 0 for all reasonable cut-offs under Scenario 1 (see Figure 1). As expected, TPR deteriorates mildly for all cases under Scenario 2, whereas FPR values are close to 0 for all reasonable cut-offs under Scenario 2. We observe higher TPR with decreasing node and/or cell sparsity. In terms of detecting cells (edges in the context of network response in this section) significantly related to the predictor of interest, SGTM performs very well under all cases in Scenario 1, as observed in Table 2. Under Scenario 2, both TPR and FPR increase with an increase in the p/n ratio. Decreasing node and/or cell sparsity also result in higher TPR for SGTM in terms of identifying significant cells.

As shown in Figure 2, the posterior mode of effective dimensionality $\hat{R}$ turns out to be

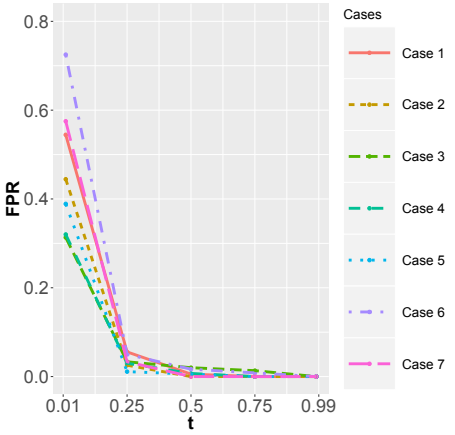


(a) TPR: Scenario 1

(b) TPR: Scenario 2



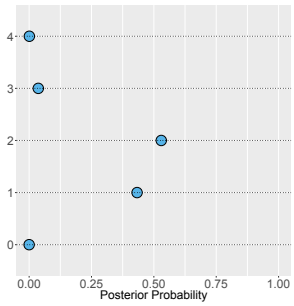
(c) FPR: Scenario 1



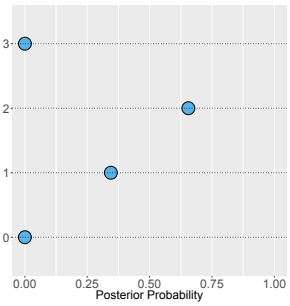
(d) FPR: Scenario 2

Figure 1: Figures in the first row show TPR in detecting tensor nodes significantly associated with the predictor of interest, under each scenario. Figures in the second row show FPR in detecting tensor nodes significantly associated with the predictor of interest, under each scenario.

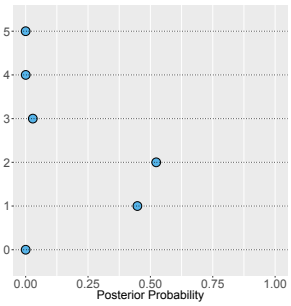
R^* in all cases under Scenario 1. As explained in Section 4.1 of the main article, $\boldsymbol{B}^*$ is a full-rank matrix in all cases under Scenario 2. Thus, to estimate $\boldsymbol{B}^*$, all available ranks of $\boldsymbol{B}$ are consumed by SGTM. This explains the posterior mode being R in all cases under Scenario 2.



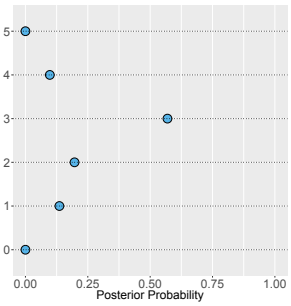
(a) Case 1, Scenario 1: $R^* = 2$



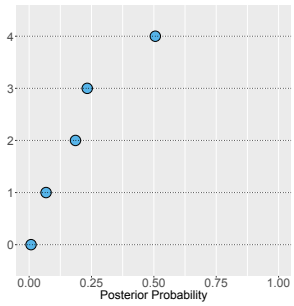
(b) Case 2, Scenario 1: $R^* = 2$



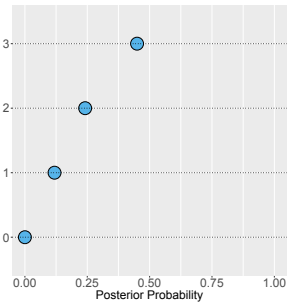
(c) Case 3, Scenario 1: $R^* = 2$



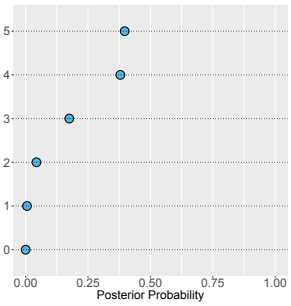
(d) Case 4, Scenario 1: $R^* = 3$



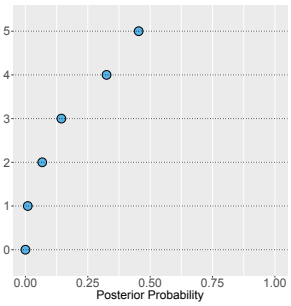
(e) Case 1, Scenario 2: $R^* = 2$



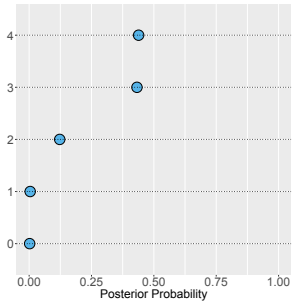
(f) Case 2, Scenario 2: $R^* = 2$



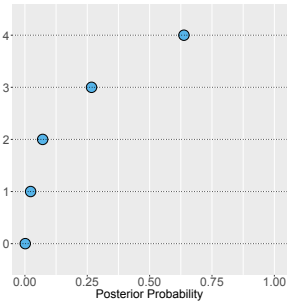
(g) Case 3, Scenario 2: $R^* = 2$



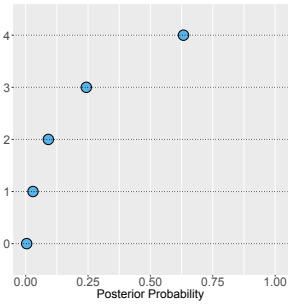
(h) Case 4, Scenario 2: $R^* = 3$



(i) Case 5, Scenario 2: $R^* = 2$



(j) Case 6, Scenario 2: $R^* = 2$



(k) Case 7, Scenario 2: $R^* = 2$

Figure 2: Posterior distribution of the effective dimensionality in various cases under Scenarios 1 and 2 for SGTM in the binary case. The first row shows the four cases in Scenario 1, while the second and third rows show the seven cases in Scenario 2.

Scenario	Operating Characteristics		
		TPR	FPR
1	Case 1	0.94	0.00
	Case 2	0.96	0.00
	Case 3	1.00	0.00
	Case 4	0.98	0.00
2	Case 1	0.52	0.00
	Case 2	0.62	0.01
	Case 3	0.74	0.03
	Case 4	0.84	0.14
	Case 5	0.53	0.00
	Case 6	0.66	0.06
	Case 7	0.62	0.04

Table 2: True Positive Rates (TPR) and False Positive Rates (FPR) in identifying edges (cells) which are significantly related to the predictor under all cases.

4 Checking Convergence of the MCMC Chain for the Posterior Distribution

Although the proposed model is a high dimensional Bayesian model, the mixing is quite good for our proposed model and convergence happens rapidly. In fact, a lot of the earlier literature on Bayesian tensor regression models (Guhaniyogi *et al.*, 2017; Spencer *et al.*, 2019) show rapid convergence as a feature. Some of these earlier work have even reported posterior inference with only a few hundred burn-ins (Guhaniyogi *et al.*, 2017). In order to provide evidence for the same with regard to our proposed approach, the article uses a burn-in of 5000 iterations, with an additional 10000 MCMC samples drawn for inference using a thinning of size 2 (please see the first paragraph of Section 3, page 11, in the article for details). We report the average effective sample size for all the coefficients in $\boldsymbol{B}$ for the 5000 post burn-in samples in the main article. They confirm fairly uncorrelated post burn-in samples in all simulation cases. Please see the first paragraph, Section 4.1 on page 15 of the main article, where they are all reported.

We also provide the trace plots of the negative log-posterior densities for the simulated examples (please refer to Figure 3 in this document) starting from the 500-th iteration. All plots indicate negative log-likelihood stabilizing within a range after a few iterations.

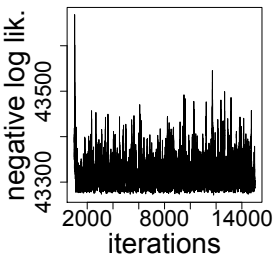
5 Simulation Studies with Weak Signal Strength

Note that the True Positive Rate (TPR) and False Positive Rate (FPR) associated with identifying the truly influential nodes are quite high in all simulation cases in the main article. To assess the effect of signal strength on the identification of influential nodes, we pursue two more simulation cases under *Scenario 1* with considerably weak signal strength. To be more precise, we consider Case 1 under Scenario 1 (described in Section 4 of the main draft) in simulation study and simulate node specific latent variables $\tilde{\boldsymbol{u}}_k^*$ under the two scenarios described below:

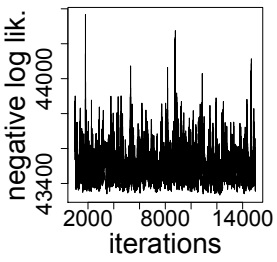
(a) $\pi_1^* N_{R^*}(0.3\mathbf{1}, 0.01\boldsymbol{I}) + (1-\pi_1^*)\delta_{\mathbf{0}}$,

(b) $\pi_1^* N_{R^*}(0.1\mathbf{1}, 0.01\boldsymbol{I}) + (1-\pi_1^*)\delta_{\mathbf{0}}$

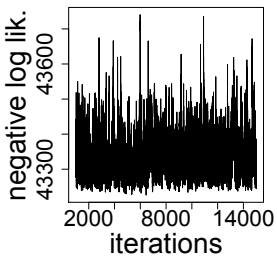
In both cases (a) and (b), the signal strength is substantially low. In fact, case (b) shows much weaker signal strength than case (a). Figure 4 presents the ROC curves related to identifying influential tensor nodes for cases (a) and (b). As expected, with a much reduced signal strength the node identification suffers considerably compared to the cases presented in the main paper. The MSE of estimating $\boldsymbol{B}^*$ in these two cases are given by 0.05 and 0.89 which are also substantially larger than the corresponding number in Case 1 under Scenario 1 in the main paper.



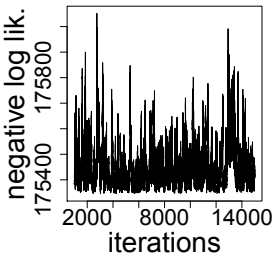
(a) Simulation 1: case 1



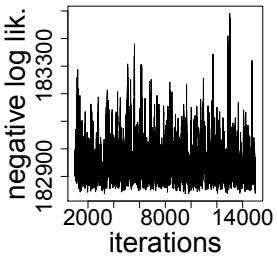
(b) Simulation 2: case 1



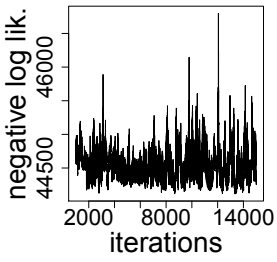
(c) Simulation 2: case 5



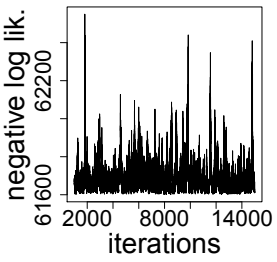
(d) Simulation 1: case 2



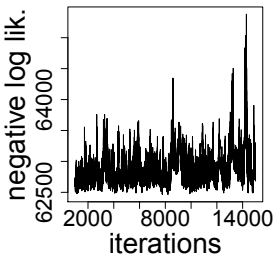
(e) Simulation 2: case 2



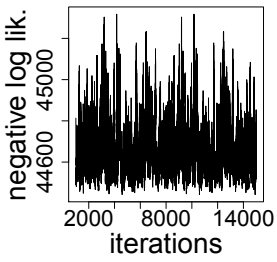
(f) Simulation 2: case 6



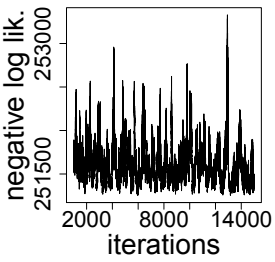
(g) Simulation 1: case 3



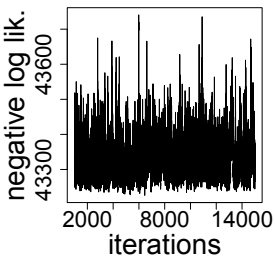
(h) Simulation 2: case 3



(i) Simulation 2: case 7

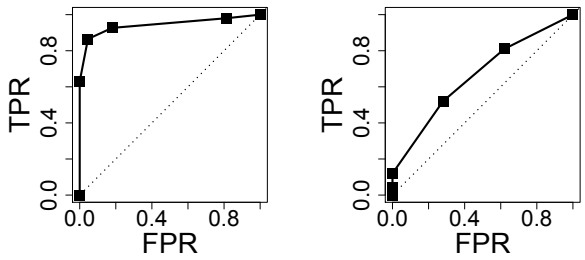


(j) Simulation 1: case 4



(k) Simulation 2: case 4

Figure 3: Figures show negative log-likelihood of the posterior for the proposed model.



(a) Case (a) (b) Case (b)

Figure 4: Figures show ROC curves in identifying the influential tensor nodes.

6 Posterior Full Conditionals

6.1 Full conditionals for symmetric continuous tensor response

In this section, we report the posterior full conditionals of the parameters in the model with a symmetric D-way tensor response (of dimension $p\times\cdots\times p$) with continuous cell entries and m predictors of interest $(x_{i1},...,x_{im})$, along with l auxiliary predictors $(z_{i1},\cdots,z_{il})$ corresponding to the ith individual. Let $\boldsymbol{y}_i=(y_{i,\boldsymbol{j}}:\boldsymbol{j}\in\mathcal{J})^T$, and $\boldsymbol{W}_h=(B_{h,\boldsymbol{j}}:\boldsymbol{j}\in\mathcal{J})^T$, $h=1,...,m$. Further assume $\mathcal{J}_k=\{\boldsymbol{j}\in\mathcal{J}:j_s=k,\text{for some s}\}$. The full conditionals are in closed form and hence allow a Gibbs sampling procedure to sample posteriors. They can be listed as

- $\beta_0|-\sim N\left[\frac{\sum_{i=1}^n \mathbf{1}^T(\boldsymbol{y}_i-\sum_{h=1}^m \boldsymbol{W}_h x_{ih}+\mathbf{1}\sum_{h_2=1}^l \beta_{h_2}z_{ih_2})/\sigma^2}{(nq)/\sigma^2+1},\frac{1}{(nq)/\sigma^2+1}\right]$
- $\beta_{h_1}|-\sim N\left(\frac{\sum_{i=1}^n z_{ih_1}^2\mathbf{1}^T(y_i-\sum_{h=1}^m \boldsymbol{W}_h x_{ih}+\mathbf{1}\sum_{h_2=1,h_2\neq h_1}^l \beta_{h_2}z_{ih_2})/\sigma^2+a_\beta/b_\beta}{q\sum_{i=1}^n z_{ih_1}^2/\sigma^2+1/b_\beta},\frac{1}{q\sum_{i=1}^n z_{ih_1}^2/\sigma^2+1/b_\beta}\right), h_1=1,...,l.$
- $\sigma^2|-\sim IG(a_\sigma+(nq)/2, b_\sigma+\sum_{i=1}^n||\boldsymbol{y}_i-\sum_{h=1}^m \boldsymbol{W}_h x_{ih}+\mathbf{1}\sum_{h_2=1}^l \beta_{h_2}z_{ih_2}||^2/2)$
- $\boldsymbol{M}_h|-\sim IW\left[(\boldsymbol{S}+\sum_{k:\tilde{\boldsymbol{u}}_{h,k}\neq\mathbf{0}}\tilde{\boldsymbol{u}}_{h,k}\tilde{\boldsymbol{u}}_{h,k}^T),(\nu+\{\#k:\tilde{\boldsymbol{u}}_{h,k}\neq\mathbf{0}\})\right]$
- $\upsilon_{h,r}|-\sim Beta[(1+\lambda_{h,r}),(r^\zeta+1-\lambda_{h,r})]$

- $\lambda_{h,r}|- \sim Ber(p_{h,r})$, where $p_{h,r} = \frac{\upsilon_{h,r} J(\mathbf{\Lambda}_h)_{(\lambda_{h,r}=1)}}{\upsilon_{h,r} J(\mathbf{\Lambda}_h)_{(\lambda_{h,r}=1)} + (1-\upsilon_{h,r}) J(\mathbf{\Lambda}_h)_{(\lambda_{h,r}=0)}}$ and $J(\mathbf{\Lambda}_h) = \prod_{i=1}^{n} N(\boldsymbol{y}_i | \beta_0 \mathbf{1} + \sum_{h=1}^{m} \boldsymbol{W}_h x_{ih} + \mathbf{1} \sum_{h_2=1}^{l} \beta_{h_2} z_{ih_2}, \sigma^2 I)$. $J(\mathbf{\Lambda}_h)_{(\lambda_{h,r}=1)}$ denotes $J(\mathbf{\Lambda}_h)$ evaluated at $\lambda_{h,r} = 1$.

- $\tilde{\boldsymbol{u}}_{h,k}|- \sim w_{\boldsymbol{u}_{h,k}} \delta_0(\tilde{\boldsymbol{u}}_{h,k}) + (1 - w_{u_{h,k}}) N(\tilde{\boldsymbol{u}}_{h,k} | m_{\boldsymbol{u}_{h,k}}, \mathbf{\Sigma}_{\boldsymbol{u}_{h,k}})$, where $\boldsymbol{U}_{h,\mathcal{J}_k} = [\boldsymbol{U}_{1,h,\mathcal{J}_k}^T : \cdots : \boldsymbol{U}_{n,h,\mathcal{J}_k}^T]^T$, $\boldsymbol{U}_{i,h,\mathcal{J}_k}^T$ has rows $(x_{ih}\lambda_{h,1} \prod_{s=1, j_s \neq k}^{D} u_{h,j_s}^{(1)}, ..., x_{ih}\lambda_{h,R} \prod_{s=1, j_s \neq k}^{D} u_{h,j_s}^{(R)})$. Further assume $\tilde{y}_{i,j}^h = y_{i,j} - \beta_0 - \sum_{h_1=1}^{l} \beta_{h_1} z_{ih_1} - \sum_{h_2=1, h_2 \neq h}^{m} B_{h_2,\boldsymbol{j}} x_{ih_2}$, $\tilde{y}_{i,\mathcal{J}_k}^h$ is a vector of collections of $\tilde{y}_{i,\boldsymbol{j}}^h$ over $\boldsymbol{j} \in \mathcal{J}_k$ and $\tilde{y}_{\mathcal{J}_k}^h$ is a vector consisting of $\tilde{y}_{i,\mathcal{J}_k}^h$ over $i = 1, ..., n$. Also,

$$\mathbf{\Sigma}_{\boldsymbol{u}_k} = \left(\boldsymbol{U}_{h,\mathcal{J}_k}^T \boldsymbol{U}_{h,\mathcal{J}_k} / \sigma^2 + \boldsymbol{M}^{-1}\right)^{-1}, \ \boldsymbol{m}_{\boldsymbol{u}_k} = \mathbf{\Sigma}_{\boldsymbol{u}_k} \boldsymbol{U}_{h,\mathcal{J}_k}^T \tilde{\boldsymbol{y}}_{\mathcal{J}_k}^h / \sigma^2$$
$$w_{u_k} = \frac{(1-\xi_h) N(\tilde{\boldsymbol{y}}_{\mathcal{J}_k}^h | 0, \sigma^2 I)}{(1-\xi_h) N(\tilde{\boldsymbol{y}}_{\mathcal{J}_k}^h | 0, \sigma^2 I) + \pi N(\tilde{y}_{\mathcal{J}_k}^h | 0, \sigma^2 \boldsymbol{I} + \boldsymbol{U}_{h,\mathcal{J}_k} \boldsymbol{M} \boldsymbol{U}_{h,\mathcal{J}_k}^T)}$$

- $\eta_{h,k}|- \sim Ber(1 - w_{u_{h,k}})$

- $\xi_h|- \sim Beta(\sum_{k=1}^{p} \eta_{h,k} + 1, \sum_{k=1}^{p} (1 - \eta_{h,k}) + 1)$.

6.2 Full conditionals for symmetric binary tensor response

In this section, we report the posterior full conditionals of the parameters in the model with a binary symmetric D-way tensor response (of dimension $p \times \cdots \times p$) and m predictors of interest $(x_{i1}, ..., x_{im})$, along with l auxiliary predictors $(z_{i1}, \cdots, z_{il})$ corresponding to the ith individual. In this case, the model can be written as

$$p(y_{i,j} = 1) = e^{\beta_0 + \sum_{h=1}^{m} B_{h,j} x_{ih} + \sum_{h_2=1}^{l} \beta_{h_2} z_{ih_2}} / (1 + e^{\beta_0 + \sum_{h=1}^{m} B_{h,j} x_{ih} + \sum_{h_2=1}^{l} \beta_{h_2} z_{ih_2}}); \ i = 1, ..., n; \ j \in \mathcal{J}.$$

Let $\psi_{i,j} = \beta_0 + \sum_{h=1}^{m} B_{h,j} x_{ih} + \sum_{h_2=1}^{l} \beta_{h_2} z_{ih_2}$. Then

$$p(y_{i,j}) = e^{\psi_{i,j} y_{i,j}} / (1 + e^{\psi_{i,j}});\ i = 1, ..., n;\ j \in \mathcal{J}. \tag{4}$$

Using the result in Polson *et al.*, 2013, the data augmented representation of the distribution of $y_{i,j}$ given in (4) follows as below

$$p(y_{i,j}|\nu_{i,j}) = \frac{1}{2} \exp((y_{i,j} - 0.5)\psi_{i,j}) \exp(-\nu_{i,j}\psi_{i,j}^2/2);\ \ \nu_{i,j} \sim \text{Polya - Gamma}(1, 0),$$

which implies

$$p(y_{i,j}) = \int_0^\infty \frac{1}{2} \exp((y_{i,j} - 0.5)\psi_{i,j}) \exp(-\nu_{i,j}\psi_{i,j}^2/2) p(\nu_{i,j}) d\nu_{i,j}$$

Also, let $c_i = \left(\frac{y_{i,j} - 0.5}{\nu_{i,j}} : j \in \mathcal{J}\right)^T$, $\Omega_i = diag(\nu_{i,j} : j \in \mathcal{J})$ and $W_h = (B_{h,j} : j \in \mathcal{J})^T$, $h = 1, ..., m$.

The full conditionals are in closed form and hence allow a Gibbs sampling procedure to sample posteriors. They can be listed as

- $\beta_0|- \sim N\left[\frac{\sum_{i=1}^{n} 1^T \Omega_i (c_i - \sum_{h=1}^{m} W_h x_{ih} - 1 \sum_{h_2=1}^{l} \beta_{h_2} z_{ih_2})}{\sum_{i=1}^{n} 1^T \Omega_i 1 + 1}, \frac{1}{\sum_{i=1}^{n} 1^T \Omega_i 1 + 1}\right]$
- $\beta_{h_1}|- \sim N\left(\frac{\sum_{i=1}^{n} z_{ih_1}^2 1^T \Omega_i (c_i - \sum_{h=1}^{m} W_h x_{ih} - 1 \sum_{h_2=1, h_2 \neq h_1}^{l} \beta_{h_2} z_{ih_2}) + a_\beta / b_\beta}{\sum_{i=1}^{n} z_{ih_1}^2 1^T \Omega_i 1 + 1/b_\beta}, \frac{1}{\sum_{i=1}^{n} z_{ih_1}^2 1^T \Omega_i 1 + 1/b_\beta}\right), h_1 = 1, ..., l.$
- $\nu_{i,j}|- \sim \text{Polya - Gamma}(1, \beta_0 + \sum_{h=1}^{m} B_{h,j} x_{ih} + \sum_{h_2=1}^{l} \beta_{h_2} z_{ih_2})$
- $M_h|- \sim IW\left[(S + \sum_{k:\tilde{u}_{h,k} \neq 0} \tilde{u}_{h,k} \tilde{u}'_{h,k}), (\nu + \{\#k : \tilde{u}_{h,k} \neq 0\})\right]$
- $\upsilon_{h,r}|- \sim Beta[(1 + \lambda_{h,r}), (r^\zeta + 1 - \lambda_{h,r})]$

- $\tilde{u}_{h,k}|- \sim w_{u_{h,k}}\,\delta_0(\tilde{u}_{h,k}) + (1 - w_{u_{h,k}})\,N(\tilde{u}_{h,k}|m_{u_{h,k}}, \Sigma_{u_{h,k}})$, where $U_{h,\mathcal{J}_k} = [U^T_{i,h,\mathcal{J}_k} : \cdots : U^T_{i,h,\mathcal{J}_k}]^T$, $U^T_{i,h,\mathcal{J}_k}$ has rows $(x_{ih}\lambda_{h,1}\prod_{s=1,j_s\neq k}^{D} u^{(1)}_{h,j_s}, ..., x_{ih}\lambda_{h,R}\prod_{s=1,j_s\neq k}^{D} u^{(R)}_{h,j_s})$. Further assume $\tilde{c}^h_{i,j} = c_{i,j} - \beta_0 - \sum_{h_1=1}^{l}\beta_{h_1}z_{ih_1} - \sum_{h_2=1,h_2\neq h}^{m} B_{h_2}x_{ih_2}$, $\tilde{c}^h_{i,\mathcal{J}_k}$ is a vector of collections of $\tilde{c}^h_{i,j}$ over $j \in \mathcal{J}_k$ and $\tilde{c}^h_{\mathcal{J}_k}$ is a vector consisting of $\tilde{c}^h_{i,\mathcal{J}_k}$ over $i = 1, ..., n$. $\Omega_{\mathcal{J}_k} = diag(\Omega_{1,\mathcal{J}_k}, ..., \Omega_{n,\mathcal{J}_k})$. Also,

$$\Sigma_{u_k} = \left(\sum_{i=1}^{n} U^T_{i,h,\mathcal{J}_k}\Omega_{i,\mathcal{J}_k}U_{i,h,\mathcal{J}_k} + M^{-1}\right)^{-1}, \;\; m_{u_k} = \Sigma_{u_k}\sum_{i=1}^{n} U^T_{i,h,\mathcal{J}_k}\Omega_{i,\mathcal{J}_k}\tilde{c}^h_{i,\mathcal{J}_k}$$
$$w_{u_k} = \frac{(1-\xi_h)N(\tilde{c}^h_{\mathcal{J}_k}|0,\Omega_{\mathcal{J}_k})}{(1-\xi_h)N(\tilde{c}^h_{\mathcal{J}_k}|0,\Omega_{\mathcal{J}_k}) + \pi N(\tilde{c}^h_{\mathcal{J}_k}|0,\Omega_{\mathcal{J}_k} + U_{h,\mathcal{J}_k}MU^T_{h,\mathcal{J}_k})}$$

- $\eta_{h,k}|- \sim Ber(1 - w_{u_{h,k}})$
- $\xi_h|- \sim Beta(\sum_{k=1}^{p}\eta_{h,k} + 1, \sum_{k=1}^{p}(1-\eta_{h,k}) + 1)$.

References

Guhaniyogi, R. and Spencer, D. (2018). Bayesian tensor response regression with an application to brain activation studies. *UCSC Tech Report: UCSC-SOE-18-15*.

Guhaniyogi, R., Qamar, S., and Dunson, D. B. (2017). Bayesian Tensor Regression. *Journal of Machine Learning Research*, **18**(79), 1–31.

Jiang, W. (2007). Bayesian variable selection for high dimensional generalized linear models: convergence rates of the fitted densities. *The Annals of Statistics*, **35**(4), 1487–1511.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, **108**(504), 1339–1349.

Spencer, D., Guhaniyogi, R., and Prado, R. (2019). Bayesian mixed effect sparse tensor

response regression model with joint estimation of activation and connectivity. *arXiv preprint arXiv:1904.00148*.