



# High Dimensional Bayesian Regularization in Regressions Involving Symmetric Tensors

Rajarshi Guhaniyogi<sup>(✉)</sup>

Department of Statistics, SOE 2, University of California Santa Cruz,  
1156 High Street, Santa Cruz, CA 95064, USA  
rguhaniy@ucsc.edu

**Abstract.** This article develops a regression framework with a symmetric tensor response and vector predictors. The existing literature involving symmetric tensor response and vector predictors proceeds by vectorizing the tensor response to a multivariate vector, thus ignoring the structural information in the tensor. A few recent approaches have proposed novel regression frameworks exploiting the structure of the symmetric tensor and assume symmetric tensor coefficients corresponding to scalar predictors to be low-rank. Although low-rank constraint on coefficient tensors are computationally efficient, they might appear to be restrictive in some real data applications. Motivated by this, we propose a novel class of regularization or shrinkage priors for the symmetric tensor coefficients. Our modeling framework a-priori expresses a symmetric tensor coefficient as sum of low rank and sparse structures, with both these structures being suitably regularized using Bayesian regularization techniques. The proposed framework allows identification of tensor nodes significantly influenced by each scalar predictor. Our framework is implemented using an efficient Markov Chain Monte Carlo algorithm. Empirical results in simulation studies show competitive performance of the proposed approach over its competitors.

**Keywords:** Low-rank structure · Symmetric tensor predictor · Shrinkage prior · Spike and slab prior

## 1 Introduction

This article is motivated by a variety of applications, in which a sample of symmetric tensors is available along with a few scalar variables of interest. Analogous to rows and columns of a matrix, various axes of a tensor are known as tensor modes and the indices of a tensor mode are often referred to as “tensor nodes”. A tensor is known to be symmetric if interchanging modes results in the same tensor. Entries in a tensor are known as “tensor cells”. In our motivating applications, each sample point is represented by its own symmetric tensor, and the

---

Supported by Office of Naval Research and National Science Foundation.

tensor nodes are labeled and shared across all sample points through a registration process. The goals of such scientific applications are two fold. First, it is important to build a predictive model to assess the change of the symmetric tensor response as the predictors of interest vary. A more important scientific goal becomes identifying nodes of the symmetric tensor significantly impacted by each predictor.

Although there is a gamut of applications involving symmetric tensors, our work is specifically motivated by scientific applications pertaining to brain connectomics. In such applications the dataset contains brain network information along with a few observable phenotypes (e.g., IQ, presence or absence of any neuronal disorder, age) for multiple subjects. Brain network information for each subject is encoded within a symmetric matrix of dimension  $V \times V$ , with  $V$  as the number of regions of interest (ROI) a human brain is parceled into following a popular brain atlas. The  $(k, l)$ th cell of the matrix consists of the number of neuron connections between the  $k$ -th and  $l$ -th regions of interest (ROI). Thus, each mode of this symmetric matrix (when viewed as a 2-D symmetric tensor) consists of  $V$  nodes ( $V = 68$  when Desikan-Killany brain atlas is followed [1]), each corresponding to a specific ROI in the human brain. The most important scientific goal here boils down to making inference on brain regions of interest (ROIs) and their inter-connections significantly associated with each phenotypic predictor [9].

One approach is to vectorize the symmetric tensor and cast the modeling problem as a high dimensional multivariate reduced rank sparse regression framework with the vectorized tensor response and scalar predictors. There are adequate literature on frequentist penalized optimization [15], as well as on Bayesian shrinkage [2] which deal with model fitting and computational issues with high dimensional multivariate reduced rank regression models. Although computationally efficient, these approaches treat 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 may appear to be detrimental in terms of model performance. Additionally, such modeling framework does not directly lead to the identification of nodes significantly associated with each predictor.

We develop a symmetric tensor response regression model with a symmetric tensor response and scalar predictors. The symmetric tensor coefficients corresponding to each predictor in this regression is assigned a novel Bayesian shrinkage prior that combines ideas from low-rank parallel factor (PARAFAC) decomposition methods, spike-and-slab priors and Bayesian high dimensional regularization techniques to generate a model that respects the tensor structure of the response. These 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, the node-wise sparsity offers inference on various tensor nodes significantly associated with a predictor. The Bayesian regularization structure allows appropriate shrinkage of unimportant cell coefficients towards zero while minimally shrinking the important cell

coefficients. All structures jointly achieve parsimony and deliver accurate characterization of uncertainty for estimating parameters and identifying significant tensor nodes. The proposed approach finds excellent synergy with the recent literature on bilinear relational data models [4, 7], multiway regression models [6] and other object oriented regression models [3], where low-rank and/or regularization structures are imposed on parameters.

The proposed framework is similar to but distinct from the recent developments in high dimensional regressions with multidimensional arrays (tensors) and other object oriented data. For example, recent literature that builds regression models with a scalar response and tensor predictors [6] is less appealing in this context, since it does not incorporate the symmetry constraint in the tensor. In the same vein, [5] formulate a Bayesian tensor response regression approach that is built upon a novel multiway stick breaking shrinkage prior on the tensor coefficients. While [5] is able to identify important tensor cells, it does 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. Also, unlike these approaches, our approach does not assume a low-rank representation of tensor coefficients; hence allowing more flexible structure to analyze impact of the predictors on tensor cells and interaction between tensor nodes. A work closely related to our framework develops shrinkage priors in a regression framework with a scalar response and an undirected network predictor, expressed in the form of a symmetric matrix [3]. However, they 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. The symmetric tensor predictor regression focuses on understanding the change of a scalar response as the symmetric tensor predictor varies, while regression with symmetric tensor response aims to study the change of the symmetric tensor as the predictors vary.

Rest of the article flows as follows. In Sect. 2 the model and prior distributions on parameters are introduced. Section 2 also briefly discusses posterior computation, where as Sect. 3 presents simulation studies to validate our approach. Finally, Sect. 4 concludes the article with an eye to the future work.

## 2 Model Development and Posterior Computation

### 2.1 Model and Prior Distributions

For  $i = 1, \dots, n$ , let  $\mathbf{y}_i = ((y_{i,(k_1, \dots, k_D)}))_{k_1, \dots, k_D=1}^V \in \mathcal{Y} \subseteq \mathbb{R}^{V \times \dots \times V}$  denote the D-way symmetric tensor response with dummy diagonal entries and  $\mathbf{z}_i = (z_{i1}, \dots, z_{ip})'$  be  $p$  predictors of interest corresponding to the  $i$ th individual. The symmetric constraint in the tensor response implies  $y_{i,(k_1, \dots, k_D)} = y_{i,(P(k_1), \dots, P(k_D))}$ , with  $P(\cdot)$  being any permutation of  $\{k_1, \dots, k_D\}$ . Due to the diagonal entries being dummies in the symmetric tensor response, it is enough to build a probabilistic model for  $\mathbf{y}_{upper} = \{y_{i,\mathbf{k}} : \mathbf{k} = (k_1, \dots, k_D), 1 \leq k_1 < \dots < k_D \leq V\}$ .

For the sake of this article, we assume  $y_{i,\mathbf{k}} \in \mathbb{R}$  and propose

$$y_{i,\mathbf{k}} = \gamma_0 + \Gamma_{1,\mathbf{k}}z_{i1} + \dots + \Gamma_{p,\mathbf{k}}z_{ip} + \epsilon_{i,\mathbf{k}}, \quad \epsilon_{i,\mathbf{k}} \stackrel{iid}{\sim} N(0, \sigma^2), \quad (1)$$

where  $\Gamma_{1,\mathbf{k}}, \dots, \Gamma_{p,\mathbf{k}}$  are the  $\mathbf{k} = (k_1, \dots, k_D)$ th cells of the  $V \times \dots \times V$  symmetric coefficient tensors  $\mathbf{\Gamma}_1, \dots, \mathbf{\Gamma}_p$  with dummy diagonal entries, respectively. Here  $\Gamma_{s,\mathbf{k}}$ ,  $s = 1, \dots, p$ , is the coefficient corresponding to the  $s$ th predictor of interest on the  $\mathbf{k} = (k_1, \dots, k_D)$ th cell of the symmetric tensor response. The coefficient  $\gamma_0 \in \mathbb{R}$  is the intercept in the regression model (4).

To account for association between the tensor response  $\mathbf{y}$  and predictors  $z_1, \dots, z_p$ , we propose a *shrinkage prior* distribution on  $\Gamma_{s,\mathbf{k}}$ ,  $s = 1, \dots, p$ , represented as a location and scale mixture of normal distributions. In particular, the distribution of  $\Gamma_{s,\mathbf{k}}$  is assumed to be conditionally independent normal distribution with

$$\Gamma_{s,\mathbf{k}} \sim N\left(\sum_{r=1}^R \lambda_{s,r} \gamma_{s,k_1}^{(r)} \dots \gamma_{s,k_D}^{(r)}, \kappa_{s,\mathbf{k}} \sigma^2\right), \quad \mathbf{k} = (k_1, \dots, k_D), \quad 1 \leq k_1 < \dots < k_D \leq V. \quad (2)$$

(2) implies  $E[\Gamma_{s,\mathbf{k}}|\sigma^2, \kappa_{s,\mathbf{k}}] = \sum_{r=1}^R \lambda_{s,r} \gamma_{s,k_1}^{(r)} \dots \gamma_{s,k_D}^{(r)}$ , i.e., the prior distribution of  $\mathbf{\Gamma}_s$  is centered on a rank- $R$  PARAFAC [10] decomposed tensor. The PARAFAC or parallel factor decomposition is a multiway analogue to the two-dimensional factor modeling of matrices. In particular it provides a low-rank structure to the mean function of  $\mathbf{\Gamma}_s$ . Note that, in the tensor regression literature, it is a fairly common practice to assume a low-rank structure for  $\mathbf{\Gamma}_s$  directly [6]. In contrast, the prior distribution in (2) centers on a low-rank PARAFAC/CP representation [10], precluding any additional imposition of a low-rank structure on  $\mathbf{\Gamma}_s$  a priori. This allows more flexibility in the structure of the coefficients.  $\kappa_{s,\mathbf{k}}$  is the scale parameter corresponding to each  $\Gamma_{s,\mathbf{k}}$  controlling the local variability of each coefficient a priori and  $\lambda_{s,r} \in \{0, 1\}$ ,  $r = 1, \dots, R$ , are introduced to assess the effect of the  $r$ -th summand on the mean of  $\Gamma_{s,\mathbf{k}}$ . In particular,  $\lambda_{s,r} = 0$  implies that the  $r$ -th summand of the low-rank factorization is not informative to predict the response.

To develop a data dependent learning of nonzero  $\lambda_{s,r}$ 's, we propose  $\lambda_{s,r} \sim Ber(\theta_{s,r})$ ,  $\theta_{s,r} \sim Beta(1, r^c)$ ,  $c > 1$ , a priori. The hyper-parameters in the beta distribution are set so as to penalize the usage of large number of summands in the PARAFAC decomposition, which protects the model from over-fitting. Define  $\gamma_{s,v} = (\gamma_{s,v}^{(1)}, \dots, \gamma_{s,v}^{(R)})' \in \mathbb{R}^R$  as the tensor node specific vectors (for  $v = 1, \dots, V$ ) describing  $\mathbf{\Gamma}_s$ . In the course of identifying important tensor nodes significantly associated with the  $s$ th predictor, we note that the  $v$ -th node has minimal effect on the  $s$ th predictor if  $\gamma_{s,v} = \mathbf{0}$ . Thus, in order to directly infer on  $\gamma_{s,v}$ , a *spike-and-slab* mixture distribution prior [8] is assigned on  $\gamma_{s,v}$  as below

$$\gamma_{s,v} \sim \begin{cases} N(\mathbf{0}, \mathbf{H}_s), & \text{if } \xi_{s,v} = 1 \\ \delta_{\mathbf{0}}, & \text{if } \xi_{s,v} = 0 \end{cases} \quad (3)$$

where  $\delta_{\mathbf{0}}$  is the Dirac-delta function at  $\mathbf{0}$  and  $\mathbf{H}_s$  is a covariance matrix of order  $R \times R$ . The rest of the hierarchy is completed by setting  $\kappa_{s,\mathbf{k}} \sim Exp(\zeta^2/2)$ ,  $\xi_{s,v} \sim$

$Ber(\Delta_s)$ ,  $\mathbf{H}_s \sim IW(\mathbf{I}, \nu)$ ,  $\Delta_s \sim U(0, 1)$ , where IW stands for the inverse wishart distribution. Finally,  $\gamma_0$  and  $\sigma^2$  are assigned  $N(\mu_\gamma, \sigma_\gamma^2)$  and  $IG(a, b)$  priors respectively, IG corresponding to the inverse gamma density. The posterior distribution of  $\xi_{s,v}$  is monitored and the  $v$ th tensor node is related to the  $s$ th predictor if the posterior probability of  $\{\xi_{s,v} = 1\}$  turns out to be greater than 0.5.

A few remarks are in order. Note that, if  $\kappa_{s,\mathbf{k}} = 0$  for all  $\mathbf{k}$ , then  $\Gamma_s$  assumes an exact low-rank decomposition given by,  $\Gamma_{s,\mathbf{k}} = \sum_{r=1}^R \lambda_{s,r} \gamma_{s,k_1}^{(r)} \cdots \gamma_{s,k_D}^{(r)}$ . Also, if  $\gamma_{s,v} = \mathbf{0}$ , then a priori  $\Gamma_{s,\mathbf{k}}$  follows an ordinary Bayesian lasso shrinkage prior distribution [12] for all  $\mathbf{k}$  with some  $k_l = v$ . In general,  $\Gamma_{s,\mathbf{k}}$  a priori can be expressed as  $\Gamma_{s,\mathbf{k}} = \sum_{r=1}^R \lambda_{s,r} \gamma_{s,k_1}^{(r)} \cdots \gamma_{s,k_D}^{(r)} + \Gamma_{s,2,\mathbf{k}}$ , where  $\Gamma_{s,2,\mathbf{k}}$  following an ordinary Bayesian lasso shrinkage prior [12].

## 2.2 Posterior Computation

Although summaries of the posterior distribution cannot be computed in closed form, full conditional distributions for all the parameters are available and correspond, in most cases, to standard families. Thus, posterior computation can proceed through a Markov chain Monte Carlo algorithm. Details of the Markov chain Monte Carlo algorithm with the conditional posterior distributions are provided in the Appendix. We run the MCMC chain for 15000 iterations. With the first 5000 as burn-ins, the posterior inference is drawn on the  $L = 10000$  post burn-in draws suitably thinned. In order to identify whether the  $v$ -th tensor node is significantly related to the  $s$ th predictor, we rely on the post burn-in  $L$  samples  $\xi_{s,v}^{(1)}, \dots, \xi_{s,v}^{(L)}$  of  $\xi_{s,v}$ . Node  $v$  is said to be influential if  $\frac{1}{L} \sum_{l=1}^L \xi_{s,v}^{(l)} > 0.5$ . Here  $\frac{1}{L} \sum_{l=1}^L \xi_{s,v}^{(l)}$  corresponds to the empirical estimate of the posterior probability of  $\{\xi_{s,v} = 1\}$ . We also assess the point estimates on tensor cell coefficients  $\Gamma_{s,\mathbf{k}}$  and present uncertainty in the estimation procedure.

## 3 Simulation Studies

### 3.1 Simulation Settings

This article illustrates the performance of our proposed approach referred to as the symmetric tensor regression (STR) along with some of its competitors under various simulation scenarios. In fitting our model, we fix the hyper-parameters at  $a = 1, b = 1, \mu_\gamma = 0, \sigma_\gamma = 1, \nu = 10$  and  $\zeta = 1$ . 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 *Higher-Order Low-Rank Regression* (HOLRR) [14] as a competitor. HOLRR provides a framework for higher order regression with a tensor response and scalar predictors. A comparative assessment of these three methods will help evaluate possible gains in inference in our method for taking into account the symmetry in the tensor response.

For the sake of simplicity, we work with  $p = 1$  (hence get rid of the subscript  $s$  hereon), with the scalar variable of interest  $z_i$ 's are drawn iid from  $N(0, 1)$ . We also set  $D = 2$ . The response is simulated from the following model

$$y_{i,k} = \gamma_0^* + \Gamma_k^* z_i + \epsilon_{i,k}, \epsilon_{i,k} \stackrel{iid}{\sim} N(0, \sigma^{*2}), \tag{4}$$

where  $\gamma_0^*$  is the true intercept and  $\Gamma^* = ((\Gamma_k^*))_{k_1, k_2=1}^V$  is the true symmetric tensor coefficient. The true intercept  $\gamma_0^*$  is set to be 0.2. To simulate the true symmetric tensor coefficient  $\Gamma^*$ , we draw  $V$  tensor node specific latent variables  $\gamma_v^* = (\gamma_v^{*(1)}, \dots, \gamma_v^{*(R^*)})'$ ,  $v = 1, \dots, V$ , each of dimension  $R^*$ , from a mixture distribution given by  $\gamma_v^* \sim \pi_1^* N_{R^*}(0.6\mathbf{1}, 0.5\mathbf{I}) + (1 - \pi_1^*)\delta_{\mathbf{0}}$ . We then construct  $\Gamma^*$  under two different simulation scenarios, referred to as *Simulation 1* and *Simulation 2*, as described below.

Simulation 1

Simulation 1 constructs cell coefficients  $\Gamma_k^* = \sum_{r=1}^{R^*} \gamma_{k_1}^{*(r)} \dots \gamma_{k_D}^{*(r)}$ . Thus, the coefficient tensor assumes a symmetric rank- $R^*$  PARAFAC decomposition. Note that, if  $\gamma_v^* = \mathbf{0}$ , then the  $v$ th tensor node specific variable has no impact in describing the relationship between  $\mathbf{y}$  and  $z$ . Hence  $(1 - \pi_1^*)$  is the probability of a tensor node being not related to  $z_i$ . We refer to it as the *node sparsity* parameter. In particular, the node sparsity parameter indicates the proportion of nodes in the tensor response (among the total of  $V$  tensor nodes) which are not related to the predictor. Notably, this data generation mechanism in Simulation 1 is quite similar (although not identical) to our fitted model. Hence, the goal of this first simulation is to evaluate the ability of the model to recover the true data-generation mechanism. In particular, we consider four cases under Simulation 1 (see Table 1) by varying the fitted rank of PARAFAC ( $R$ ), true rank of  $\Gamma^*$  ( $R^*$ ), sample size ( $n$ ), no. of tensor nodes ( $V$ ) and the tensor node sparsity (defined before).

**Table 1.** First six columns present the cases under Simulation 1. The next seven columns present the cases under Simulation 2.

Simulation 1						Simulation 2						
Cases	$R$	$R^*$	$n$	$V$	$\pi_1^*$	Cases	$R$	$R^*$	$n$	$V$	$\pi_1^*$	$\pi_2^*$
1	4	2	70	30	0.4	1	4	2	70	30	0.4	0.5
2	3	2	70	60	0.6	2	3	2	70	60	0.6	0.5
3	5	2	100	30	0.5	3	4	2	70	30	0.4	0.7
4	5	3	100	60	0.7	4	4	2	70	30	0.6	0.7

*Simulation 2*

Under Simulation 2, we first simulate node specific latent variables, similar to Simulation 1. If either of  $\gamma_{k_1}^*, \dots, \gamma_{k_D}^*$  is  $\mathbf{0}$ , we set  $\Gamma_k^* = 0$ . Otherwise,  $\Gamma_k^*$  is simulated from a mixture distribution  $\pi_2^* N(0, 1) + (1 - \pi_2^*) \delta_0$ , where  $(1 - \pi_2^*)$  is referred to as the *cell sparsity* parameter and  $\delta_0$  refers to the Dirac delta function at  $\mathbf{0}$ . Unlike Simulation 1, Simulation 2 does not necessarily leads to a low-rank structure of  $\Gamma^*$ . Hence this simulation is ideal for investigating the performance under model mis-specification. Table 1 shows different cases under Simulation 2 where the model is investigated.

We compare the three competitors in terms of their accuracy of estimating  $\Gamma^*$ . The accuracy of estimating  $\Gamma^*$  is measured by the scaled mean squared error (MSE) defined as  $\|\Gamma^* - \hat{\Gamma}\|^2 / \|\Gamma^*\|^2$ , where  $\hat{\Gamma}$  corresponds to a suitable point estimate of  $\Gamma$ , e.g., the posterior mean of  $\Gamma$  for STR.  $\|\cdot\|$  refers to the Frobenius norm of a matrix. Additionally, we quantify the uncertainty offered by each these competitors through the coverage and length of 95% credible intervals of  $\Gamma_k$ , averaged over all  $k$ . Length and coverage of posterior 95% credible intervals for each  $\Gamma_k$  are available empirically from the post burn-in MCMC samples of  $\Gamma$  for our proposed approach. On the other hand, the 95% confidence intervals of frequentist competitors are constructed using a bootstrap approximation. To infer on the performance of STR in terms of identifying tensor nodes significantly associated with  $z_i$ , we present True Positive Rate (TPR) and False Positive Rate (FPR) with different choices of the cut-off  $t$  for all simulation cases. As mentioned earlier, such measures are not available for our competitors since they are not designed to detect tensor nodes related to the predictor of interest.

**3.2 Simulation Results**

Scaled MSE, coverage and length of 95% CI for all competitors are presented under Simulations 1 and 2 in Tables 2 and 3, respectively. With no model mis-specification under Simulation 1, STR is showing significantly better performance than LS and HOLRR under all four cases. The low-rank structure of HOLRR facilitates its superior performance over LS for larger  $V/n$  ratio. However, as  $V/n$  ratio increases, HOLRR loses its edge over LS. All three competitors show over-coverage under Simulation 1, with STR producing substantially narrower credible intervals. Moreover, for a fixed  $n$ , the credible intervals tend to be more narrow with increasing  $V$  for all competitors.

Even under model mis-specification in Simulation 2, STR outperforms all competitors in cases with smaller  $V$  and higher node sparsity (Cases 1 and 3), as seen in Table 3. However, with a larger  $V$  in case 2, LS and STR are competitive to each other. Since HOLRR is constructed on variants of sparsity within low-rank principle, it loses edge over LS in terms of MSE in these cases. Comparing MSE of STR between cases 3 and 4, we find that MSE increases as the node sparsity decreases. Similarly, comparing Cases 1 and 3 reveals adverse effect of decreasing cell sparsity on MSE of STR. It is generally found that the effect of node sparsity is more profound than the effect of cell sparsity on the performance.

**Table 2.** Mean Squared Error (MSE), average coverage and average length of 95% credible interval for STR, LS and HOLRR are presented for cases under Simulation 1, with the lowest MSE in each case is boldfaced.

Case	Competitors			
		SGTM	LS	HOLRR
1	MSE $\times 10^3$	<b>3.4</b>	46	42
	Avg. cov	0.99	0.99	0.99
	Avg. length	0.10	1.74	1.68
2	MSE $\times 10^3$	<b>1.3</b>	18	12
	Avg. cov	0.99	0.99	0.99
	Avg. length	0.08	2.79	2.63
3	MSE $\times 10^3$	<b>1.9</b>	26	35
	Avg. cov	0.99	0.99	0.99
	Avg. length	0.10	2.53	2.37
4	MSE $\times 10^3$	<b>0.8</b>	5.1	8.5
	Avg. cov	0.99	1.00	0.99
	Avg. length	0.21	4.82	4.74

**Table 3.** Mean Squared Error (MSE), average coverage and average length of 95% credible interval for STM, LS and HOLRR are presented for cases under Simulation 2, with the lowest MSE in each case in boldfaced.

Case	Competitors			
		SGTM	LS	HOLRR
1	MSE	<b>0.09</b>	0.21	0.30
	Avg. cov	0.96	0.97	0.89
	Avg. length	0.12	0.95	0.78
2	MSE	0.23	0.29	<b>0.21</b>
	Avg. cov	0.89	0.98	0.81
	Avg. length	0.14	1.53	0.78
3	MSE	<b>0.13</b>	0.22	0.30
	Avg. cov	0.96	0.98	0.92
	Avg. length	0.15	1.09	0.88
4	MSE	0.16	<b>0.15</b>	0.32
	Avg. cov	0.92	0.99	0.92
	Avg. length	0.26	1.69	1.29

Moving onto uncertainty characterization, STR shows close to nominal coverage along with its competitors in cases 1, 2 and 4 when  $V$  is small. With increasing  $V$ , coverage of STR and HOLRR drops below 0.90. Similar to Simulation 1, STR demonstrates sufficiently narrower credible intervals than HOLRR in all cases. LS offers over-coverage with much wider 95% credible intervals in all cases.



**Table 4.** True Positive Rates (TPR) and False Positive Rates (FPR) in identifying nodes which are significantly related to the predictor under all cases with cut-offs  $t = 0.1, 0.5, 0.9$ .

Simulation	Accuracy of tensor node identification						
		TPR ( $t = 0.1$ )	TPR ( $t = 0.5$ )	TPR ( $t = 0.9$ )	FPR ( $t = 0.1$ )	FPR ( $t = 0.5$ )	FPR ( $t = 0.9$ )
1	Case 1	1.00	1.00	1.00	0.00	0.00	0.00
	Case 2	1.00	1.00	1.00	0.00	0.00	0.00
	Case 3	1.00	1.00	1.00	0.00	0.00	0.00
	Case 4	1.00	1.00	1.00	0.00	0.00	0.00
2	Case 1	1.00	0.90	0.60	0.00	0.00	0.00
	Case 2	1.00	1.00	0.84	0.00	0.00	0.00
	Case 3	0.98	0.80	0.55	0.24	0.00	0.00
	Case 4	1.00	1.00	0.76	0.24	0.15	0.00

Since LS and HOLRR are not designed to detect nodes significantly related to the predictor, we focus our inference on STR for node detection. To this end, we choose three cut-off values  $t = 0.1, 0.5, 0.9$  and present TPR and FPR values for our approach. STR yields the posterior probability of a node being 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 Table 4) turn out to be close to 1 and 0, respectively, for all the simulation cases, indicating a close to perfect active node detection. The posterior distributions of  $\gamma_0$  also appear to be centered around the truth (not shown here).

## 4 Conclusion and Future Work

The overarching goal of this article is to propose a symmetric tensor regression framework with a symmetric tensor response and scalar predictors. The model is aimed at identifying tensor nodes significantly related to each scalar predictor. Unlike the existing approaches, the proposed framework does not assume any low-rank constraint on the symmetric tensor coefficients. Rather, we propose a tensor shrinkage prior which decomposes the symmetric tensor coefficients into low-rank and sparse components a priori. The low-rank component is further assigned a novel hierarchical mixture prior to enable identification of tensor nodes related to each predictor. The sparse component is equipped with Bayesian regularization or shrinkage priors to enable accurate estimation of tensor cell coefficients. Detailed simulation study with data generated under the true model and mis-specified model demonstrates superior performance of our approach compared to its competitors.

Although there is a considerable literature on theoretical understanding of Bayesian shrinkage priors in high dimensional regression, there is a limited literature on theoretical aspects of shrinkage prior on tensor coefficients. In future, we will focus on developing conditions for posterior consistency of the proposed approach under suitable conditions imposed on tensor shrinkage priors. It is also instructive to employ other shrinkage priors on  $\Gamma_{s,2,k}$  from the class of global-local shrinkage prior distributions [13] and provide a comparative understanding of their performances. Finally, we would like to extend our approach when each entry in  $y_{i,k}$  are categorical or count in nature. Some of these constitute our future work.

## Appendix

The full conditional distributions of parameters for implementing the MCMC is given by following.

1.  $\gamma_0 | - \sim N \left[ \frac{\sum_{i=1}^n \mathbf{1}^T (y_i - \sum_{s=1}^p \Gamma_s z_{is}) / \sigma^2}{(nq) / \sigma^2 + 1}, \frac{1}{(nq) / \sigma^2 + 1} \right]$
2.  $\kappa_{s,k} | - \sim RGIG \left[ \frac{1}{2}, \frac{(\Gamma_{s,k} - \sum_{r=1}^R \lambda_{s,r} \gamma_{s,k_1}^{(r)} \cdots \gamma_{s,k_D}^{(r)})^2}{\sigma^2}, \zeta^2 \right], s = 1, \dots, p; 1 \leq k_1 < \dots < k_D \leq V$
3. Let  $\mathbf{Z}$  be an  $n \times p$  matrix with the  $i$ th row as  $(z_{i1}, \dots, z_{ip})$ . Let  $\mathbf{y}_k = (y_{1,k}, \dots, y_{n,k})'$ ,  $\mathbf{D}_k = \text{diag}(\kappa_{1,k}, \dots, \kappa_{p,k})$ ,  $\mathbf{m}_k = (\sum_{r=1}^R \lambda_{s,r} \gamma_{s,k_1}^{(r)} \cdots \gamma_{s,k_D}^{(r)} : 1 \leq s \leq p)'$  and  $\boldsymbol{\gamma}_k = (\Gamma_{1,k}, \dots, \Gamma_{p,k})'$ . Let  $\boldsymbol{\Sigma}_\Gamma = \sigma^2 (\mathbf{Z}'\mathbf{Z} + \mathbf{D}_k^{-1})^{-1}$ ,  $\boldsymbol{\mu}_\Gamma = \boldsymbol{\Sigma}_\Gamma [\mathbf{Z}'(\mathbf{y}_k - \gamma_0 \mathbf{1}) + \mathbf{D}_k^{-1} \mathbf{m}_k] / \sigma^2$ . Then  $\boldsymbol{\gamma}_k | - \sim N(\boldsymbol{\mu}_\Gamma, \boldsymbol{\Sigma}_\Gamma)$ .
4.  $\sigma^2 | - \sim IG(a + (nq)/2 + (pq)/2, b + \sum_{i=1}^n \|\mathbf{y}_i - \sum_{s=1}^p \Gamma_s z_{is}\|^2 / 2 + \sum_{s=1}^p \sum_k (\Gamma_{s,k} - \sum_{r=1}^R \lambda_{s,r} \gamma_{s,k_1}^{(r)} \cdots \gamma_{s,k_D}^{(r)})^2)$
5. Let  $\mathcal{K}_v = \{\mathbf{k} : \exists l, s.t., k_l = v\}$   $\mathbf{v}_s = (\Gamma_{s,k} : \mathbf{k} \in \mathcal{K}_v)$ . Define  $\mathbf{J}$  as a matrix of the order  $\#\mathcal{K}_v \times R$  with a representative row  $(\prod_{k_l \neq v} \lambda_{s,1} \gamma_{s,k_l}^{(1)}, \dots, \prod_{k_l \neq v} \lambda_{s,R} \gamma_{s,k_l}^{(R)})$ . Define,  $w_{\gamma_{s,v}} = \frac{(1 - \Delta_s) N(v_s | \mathbf{0}, \sigma^2 \mathbf{I})}{(1 - \Delta_s) N(v_s | \mathbf{0}, \sigma^2 \mathbf{I}) + \Delta_s N(v_s | \mathbf{J} \mathbf{H}_s \mathbf{J}' + \sigma^2 \mathbf{I})}$ ,  $\mathbf{M}_s = \text{diag}(\kappa_{s,k} : \mathbf{k} \in \mathcal{K}_v)$   $\boldsymbol{\Sigma}_{\gamma_{s,v}} = (\mathbf{J}' \mathbf{M}_s^{-1} \mathbf{J} / \sigma^2 + \mathbf{H}_s^{-1})^{-1}$ ,  $\boldsymbol{\mu}_{\gamma_{s,v}} = \boldsymbol{\Sigma}_{\gamma_{s,v}} \mathbf{J}' \mathbf{M}_s^{-1} \boldsymbol{\gamma}_{s,v} / \sigma^2$ . Then  $\boldsymbol{\gamma}_{s,v} | - \sim w_{\gamma_{s,v}} \delta_{\mathbf{0}} + (1 - w_{\gamma_{s,v}}) N(\boldsymbol{\mu}_{\gamma_{s,v}}, \boldsymbol{\Sigma}_{\gamma_{s,v}})$
6.  $\mathbf{H}_s | - \sim IW(\mathbf{I} + \sum_{v: \gamma_{s,v} \neq \mathbf{0}} \boldsymbol{\gamma}_{s,v} \boldsymbol{\gamma}_{s,v}', \nu + \{\#v : \boldsymbol{\gamma}_{s,v} \neq \mathbf{0}\}), s = 1, \dots, p$
7.  $\theta_{s,r} | - \sim \text{Beta}(1 + \lambda_{s,r}, r^c + 1 - \lambda_{s,r}), s = 1, \dots, p; r = 1, \dots, R$
8.  $\Delta_s | - \sim \text{Beta}(1 + \sum_{v=1}^V \xi_{s,v}, 1 + \sum_{v=1}^V (1 - \xi_{s,v})), s = 1, \dots, p$
9.  $\lambda_{s,r} | - \sim \text{Ber}(p \lambda_{s,r})$ , where  $p \lambda_{s,r} = \frac{\theta_{s,r} N_{s,r}}{\theta_{s,r} N_{s,r} + (1 - \theta_{s,r}) D_{s,r}}$ , where  $N_{s,r} = \prod_{\mathbf{k}} N(y_{i,k} | \sum_{r'=1, r' \neq r}^R \lambda_{s,r'} \gamma_{s,k_1}^{(r')} \cdots \gamma_{s,k_D}^{(r')} + \gamma_{s,k_1}^{(r)} \cdots \gamma_{s,k_D}^{(r)}, \sigma^2)$ ,  $D_{s,r} = \prod_{\mathbf{k}} N(y_{i,k} | \sum_{r'=1, r' \neq r}^R \lambda_{s,r'} \gamma_{s,k_1}^{(r')} \cdots \gamma_{s,k_D}^{(r')}, \sigma^2)$  for  $r = 1, \dots, R$ .

## References

1. Desikan, R.S., et al.: An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**(3), 968–980 (2006)
2. Goh, G., Dey, D.K., Chen, K.: Bayesian sparse reduced rank multivariate regression. *J. Multivar. Anal.* **157**, 14–28 (2017)
3. Guha, S., Rodriguez, A.: Bayesian regression with undirected network predictors with an application to brain connectome data. arXiv preprint [arXiv:1803.10655](https://arxiv.org/abs/1803.10655) (2018)
4. Guhaniyogi, R., Rodriguez, A.: Joint modeling of longitudinal relational data and exogenous variables. *Bayesian Anal.* (2019, in press). <https://doi.org/10.1214/19-BA1160>
5. Guhaniyogi, R., Spencer, D.: Bayesian tensor response regression with an application to brain activation studies. UCSC Technical report: UCSC-SOE-18-15 (2019)
6. Guhaniyogi, R., Qamar, S., Dunson, D.B.: Bayesian tensor regression. *J. Mach. Learn. Res.* **18**(5), 1–31 (2017)
7. Hoff, P.D.: Bilinear mixed-effects models for dyadic data. *J. Am. Stat. Assoc.* **100**(469), 286–295 (2005)
8. Ishwaran, H., Rao, J.S.: Spike and slab variable selection: frequentist and Bayesian strategies. *Ann. Stat.* **33**(2), 730–773 (2005)
9. Jung, R.E., et al.: Neuroanatomy of creativity. *Hum. Brain Mapp.* **31**(3), 398–409 (2010)
10. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. *SIAM Rev.* **51**(3), 455–500 (2009)
11. Li, L., Zhang, X.: Parsimonious tensor response regression. *J. Am. Stat. Assoc.* **112**(519), 1131–1146 (2017)
12. Park, T., Casella, G.: The Bayesian lasso. *J. Am. Stat. Assoc.* **103**(482), 681–686 (2008)
13. Polson, N.G., Scott, J.G.: Shrink globally, act locally. Sparse Bayesian regularization and prediction. *Bayesian Stat.* **9**, 501–538 (2010)
14. Rabusseau, G., Kadri, H.: Higher-order low-rank regression. arXiv preprint [arXiv:1602.06863](https://arxiv.org/abs/1602.06863) (2016)
15. Rothman, A.J., Levina, E., Zhu, J.: Sparse multivariate regression with covariance estimation. *J. Comput. Graph. Stat.* **19**(4), 947–962 (2010)