BAYESIAN SUPERVISED CLUSTERING OF UNDIRECTED NETWORKS FOR UNDERSTANDING GROUP SPECIFIC IMPACTS OF THE BRAIN NETWORK ON HUMAN CREATIVITY

By Sharmistha Guha[1], and Rajarshi Guhaniyogi[2,*]

[1] *Department of Statistical Science, Duke University, sg516@duke.edu*

[2] *Department of Statistics, University of California Santa Cruz, * rguhaniy@ucsc.edu*

This article develops a flexible relationship between a measure of creative achievement, the Creative Achievement Questionnaire (CAQ), and the brain network of subjects from a brain connectome dataset obtained using a diffusion weighted magnetic resonance imaging (DWI) technique. Undirected brain networks are often visualized using symmetric adjacency matrices, with row and column indices of the matrix representing regions of interest (ROI), and a cell entry signifying the estimated number of fiber bundles connecting the corresponding row and column ROIs. Motivated by earlier studies on the differences in the relationship between brain connectivity networks and phenotypic traits for different groups of individuals, this article aims to cluster individuals according to the shared relationships of their brain networks and creativity. Additionally, scientific interest lies in identifying ROIs in the human brain significantly associated with creative achievement in each cluster of subjects. To address these questions, we propose a novel Bayesian mixture modeling framework with an undirected network response and scalar predictors. The symmetric matrix coefficients corresponding to the scalar predictors of interest in each mixture component are embedded with low-rankness and group sparsity within the low-rank structure. Being a principled Bayesian framework allows us to precisely characterize the uncertainty in detecting significant network nodes in each cluster. Empirical results in various simulation scenarios illustrate substantial inferential gains of the proposed framework in comparison with competitors. Analysis of the brain connectome data with the proposed model reveals interesting insights into the brain regions significantly related to creative achievement in each cluster of individuals.

1. Introduction.

In recent years, network data is regularly encountered in disciplines as diverse as neuroscience, genetics, finance and economics. Statistical models involving networks are particularly challenging, especially due to the need for flexible formulations to account for the topological structure of the network. This article is motivated by applications where undirected networks along with scalar variables are available for multiple subjects. More specifically, we focus on a brain connectome data obtained using a diffusion weighted magnetic resonance imaging (DWI) technique. Using data from DWI, a human brain can be segmented into different functional regions of interest (ROIs), simultaneously estimating the number of fiber bundles connecting any two regions. Fiber connections in a human brain can be viewed as constituting an undirected network expressed in the form of a symmetric matrix, with row and column indices of the matrix corresponding to the regions of interest (ROIs) and the (j_1, j_2)th cell representing the estimated number of fibre bundles connecting the j_1th and j_2th ROIs. Along with brain networks, information on a measure of creative achievement, as well as behavioral variables like age and sex, are available for each subject in the dataset of interest.

Keywords and phrases: Bayesian mixture modeling, Brain connectome data, Network clustering, Network node selection, Spike and slab prior

The dataset offers interesting opportunities to characterize the relationship between brain networks and brain related phenotypes for subjects included in the analysis. In fact, the number of fibers between ROIs varies across individuals, and perhaps features of the fibre connection networks relate to traits of the individuals, such as their creativity. Motivated by such neuro-scientific applications, this article develops Bayesian tools to establish a regression relationship between a network response and scalar predictors. Our modeling endeavor primarily aims at achieving the following inferential objectives simultaneously. First, we intend to cluster subjects into groups, with members in each group sharing the same relationship between the undirected network response and scalar covariates. An additional inferential interest lies in identifying nodes in the network significantly impacted by each predictor of interest in each cluster. In the context of the brain connectome application, the latter objective amounts to drawing inference on brain regions of interest (ROIs) significantly associated with creative achievement in each cluster.

Rather than focusing on multiple network observations collected over different individuals, an overwhelming literature in network data aims at understanding the topological structure of a single network. Some notable examples in this direction include exponential random graph models (Frank and Strauss, 1986), social space models (Hoff, Raftery and Handcock, 2002; Hoff, 2005, 2009) including random dot product graph (RDPG) models (Young and Scheinerman, 2007) and stochastic block models (Nowicki and Snijders, 2001). In the context of developing a regression/classification model with a network response, one possibility is to extract a few summary measures from the network to reshape the network object into a multivariate response (e.g., see Bullmore and Sporns, 2009 and references therein). Clearly, the success of this approach is highly dependent on the choice of summary measures. Furthermore, this kind of approach cannot identify the impact of specific nodes on the predictor, which is of clear interest in our setting. A more closely related article by Wang et al. (2017) exploits the relational nature of the network response to develop a Bayesian modeling framework with a network response and scalar predictors. However, it assumes an identical regression relationship between the network response and scalar predictors for every subject and is not designed to detect network nodes significantly related to a scalar predictor.

Viewing networks as symmetric tensors, our inferential problem can also be formulated under a regression framework with a symmetric tensor response and scalar predictors. To this end, there are recent efforts to build regression models with a tensor response and scalar predictors (Li and Zhang, 2017; Guhaniyogi, Qamar and Dunson, 2018) without enforcing any symmetry constraint on the tensor response, and hence are not directly applicable in our context. More recently, Sun and Li (2017) have devised a new class of models which are equipped to incorporate a symmetry constraint for the tensor response in the modeling framework. Their approach adds element-wise sparsity to the tensor coefficient for identifying tensor cells related to the predictors, but does not specifically aim at drawing inference on influential tensor nodes related to each predictor. A related approach to ours appears in Guha and Rodriguez (2018), where a regression framework with a scalar response and a network predictor is proposed. While Guha and Rodriguez (2018), as well as related methods in the literature (Durante et al., 2018; Relión et al., 2019), treat the network as a predictor, we treat it as the response. This difference in the modeling approach leads to a different focus and interpretation. Network predictor regression focuses on understanding the change in a biological outcome as the network image varies, while the network response regression aims to study the change in the network as the predictors such as the creativity levels, age and sex vary. In a sense, their difference is comparable to that between multi-response regression and multi-predictor regression in the classical vector-valued regression context. Also, our framework bypasses the need to invert any high dimensional matrix to draw Bayesian inference, thereby adding substantial computational gain over Guha and Rodriguez (2018). Such a computational advantage is crucial, especially in the analysis of networks with moderately large to a large number of nodes, when computation using the approach of Guha and Rodriguez (2018) may become quite prohibitive. Moreover, Guha and Rodriguez (2018) tacitly assume that the same set of network nodes influence the regression function in a similar manner for every individual. While this assumption may hold true for some applications, it may appear to be restrictive for a variety of neuro-scientific applications.

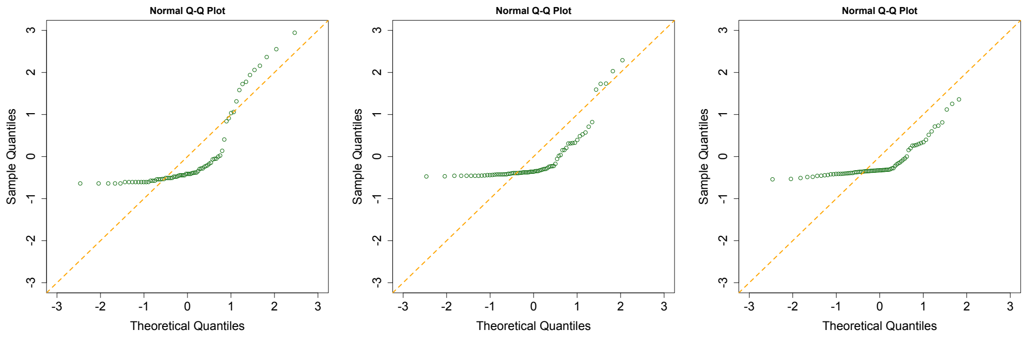
In fact, earlier literature in neuroscience provides evidence of differences in the relationship between brain connectivity networks with phenotypic traits for different groups of individuals (Saad et al., 2012; Meskaldji et al., 2013, 2015). However, flexible statistical methods for analyzing these differences have somewhat lagged behind the increasingly routine collection of such data. Existing literature has largely focused on scenarios where the undirected networks have a similar relationship with the scalar covariates for all individuals, in regression (Guha and Rodriguez, 2018) as well as classification problems (Durante et al., 2018; Relión et al., 2019), as opposed to addressing the general problem of developing a flexible relationship between the network response and the corresponding predictors which accounts for changes in different groups of individuals. Note that while the literature for network classification can be extended to ascertain group differences from a sample of symmetric networks, these methods pre-identify two groups having potentially different relationships between the network response and the scalar predictors prior to the analysis. Instead, it would be of scientific interest to formulate a modeling framework that is equipped to determine both the number and constitution of clusters from the data. To this end, one can invoke the literature on clustering of matrices or higher order tensor objects into multiple groups (Huang, Shen and Buja, 2009; Lee et al., 2010; Chi and Lange, 2015; Chi, Allen and Baraniuk, 2017; Li et al., 2014; Cao et al., 2013; Wu, Benson and Gleich, 2016; Sun and Li, 2017), though this literature is more pertinent to unsupervised clustering of networks, as opposed to our interest in the supervised clustering of undirected networks.

In this article, we propose a novel nonparametric Bayesian modeling approach to achieve the aforementioned inferential objectives simultaneously. To be more specific, a Dirichlet process (DP) mixture model is employed to the data, which leads to clustering of subjects into groups signifying differential relationships between the network response and scalar predictors. Further, the network valued coefficients corresponding to the predictors of interest in each mixture component are assigned a node-wise sparsity structure using a Bayesian spike-and-slab variable selection prior for identifying network nodes significantly associated with these predictors. The Bayesian framework helps in characterizing the uncertainty related to clustering as well as the uncertainty associated with identifying important network nodes in each group.

The rest of the article progresses as follows. Section 2 provides a brief description of the brain connectome data and the inferential objectives. Sections 3 and 4 describe the model development and posterior computation, respectively. Empirical investigations of the model with simulation studies and the brain connectome data analysis are presented in Sections 5 and 6, respectively. Finally, Section 7 concludes the paper with an eye towards future work.

2. Brain Connectome Dataset with the Creative Achievement Questionnaire (CAQ).

Our dataset of interest consists of brain connectome information of several subjects collected using a brain imaging technique called *Diffusion Weighted Magnetic Resonance Imaging* (DWI). It is openly available at `https://neurodata.io/mri`. Note that DWI is a magnetic resonance imaging technique that measures the restricted diffusion of water in tissues in order to produce neural tract images which are then pre-processed using the NDMG pre-processing pipeline (Kiar et al., 2016; Kiar, Gorgolewski and Kleissas, 2017; Kiar et al., 2017). In the context of DWI, the human brain is divided according to the Desikan atlas (Desikan et al., 2006) that identifies 34 cortical regions of interest (ROIs) in each of the left and



(a) QQ Plot: Cell 1 (b) QQ Plot: Cell 2 (c) QQ Plot: Cell 3

FIG 1. *QQ-plot of residuals corresponding to the linear regressions fitted on three representative cells (edges in the brain network) with $n = 73$ subjects of the CAQ dataset.*

right hemispheres of the human brain, implying 68 cortical ROIs in all. These 68 ROIs are contained in 6 *lobes* each in the left and the right hemispheres, namely the *temporal, frontal, occipital, parietal, insula* and *cingulate* lobes.

Using DWI, 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. Thus, for each subject, representing the brain network, is a symmetric matrix of dimension 68×68, with the (j_1, j_2)th off-diagonal entry being the estimated number of fibers connecting the j_1th and the j_2th brain ROIs and diagonal entries set to zero. For each subject, information on creativity as measured by the *Creative Achievement Questionnaire* (CAQ) is also available, which we treat as a predictor of interest. Creative achievement can be perceived as the aggregate of creative products of an individual during his/her lifetime (Carson, Peterson and Higgins, 2005). CAQ, in particular, is a self-reported measure of creative achievement that assesses achievement across ten domains of creativity. To obtain the CAQ, each subject is given a questionnaire to complete, which is then used to form a comprehensive measure of creative productivity across ten domains, including visual arts, music, creative writing, dance, drama, architecture, humor, scientific discovery, invention and culinary arts. As a measure of creativity, CAQ has been recognized in the literature to be both reliable and valid (Jung et al., 2010). Along with the brain network information and CAQ, age and sex are also available as additional covariates for $n = 73$ subjects in our dataset of interest. All subjects recruited in the study belong to the age group of 18-29 years.

The main objective of the data analysis lies in supervised clustering of brain networks from 73 subjects. The Bayesian mixture model of network objects proposed in this article achieves clustering of subjects into different groups, each group having a different regression relationship of the brain connectome with CAQ, age and sex. This model offers inference on influential network nodes related to CAQ in different clusters, allowing for the scientific understanding of the relationship between creativity and the brain connectome with characterization of uncertainty in different groups/clusters of subjects. In addition, the network mixture model automatically relaxes the normality assumption on the distribution of the network response matrix cells. This is deemed appropriate for our data application, since after fitting linear regression models independently for each cell of the network response matrix with CAQ, age and sex as predictors, we observe non-normality in the standardized residuals (refer to the QQ plots of the standardized residuals for three representative cells in Figure 1).

3. Supervised Clustering of Undirected Networks: Model and Prior Formulation. For $i = 1,...,n$, let $\boldsymbol{Y}_i \in \mathcal{Y} \in \mathbb{R}^{p \times p}$ denote the weighted network response with p nodes, $\boldsymbol{x}_i = (x_{i1},...,x_{im})'$ be m predictors of interest and $\boldsymbol{z}_i = (z_{i1},...,z_{il})'$ be l auxiliary predictors corresponding to the ith individual. Mathematically, this amounts to $\boldsymbol{Y}_i$ being a $p \times p$ matrix, with the (j_1, j_2)-th entry of $\boldsymbol{Y}_i$ denoted by $y_{i,(j_1,j_2)} \in \mathbb{R}$. In this paper, we focus on networks that contain no self relationship, i.e., $y_{i,(j_1,j_2)} \equiv 0$ when $j_1 = j_2$, and are undirected ($y_{i,(j_1,j_2)} = y_{i,(j_2,j_1)}$). In the context of the data described in Section 2, CAQ is the predictor of interest, whereas age and sex are considered auxiliary predictors.

We assume that the relationship between $\boldsymbol{x}_i$ and the response varies in every cell (j_1, j_2). In contrast, an auxiliary predictor explains the response in every cell identically. Since $\boldsymbol{Y}_i$ is symmetric with 0 diagonal entries, it suffices to build a probabilistic generative mechanism for the upper triangular vector $\boldsymbol{y}_i = (y_{i,\boldsymbol{j}} : 1 \le j_1 < j_2 \le p)$ of dimension $q = \frac{p(p-1)}{2}$. This is a common practice in the undirected relational data modeling (Hoff, 2005). Moreover, working with $\boldsymbol{y}_i$ is fundamentally different from the exercise of vectorizing the upper triangle of the matrix $\boldsymbol{Y}_i$, since every element $y_{i,\boldsymbol{j}}$ of $\boldsymbol{y}_i$ keeps a tab on the cell index $\boldsymbol{j} = (j_1, j_2)$ of the entry, which will be crucial in the modeling development described below.

To develop a sufficiently flexible relationship between $\boldsymbol{y}_i$ and predictors $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, we propose to model the conditional distribution of $\boldsymbol{y}_i \,|\, \boldsymbol{x}_i, \boldsymbol{z}_i, \sigma^2$, denoted by $f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i, \sigma^2)$ as a mixture model given by,

(1)

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{z}_i,\sigma^2) = \int N_q\left(\boldsymbol{y}_i|\boldsymbol{1}\gamma_0 + \sum_{s=1}^{m}\boldsymbol{\beta}_s x_{is} + \boldsymbol{1}\sum_{s=1}^{l}\gamma_s z_{is}, \sigma^2\boldsymbol{I}_q\right) dG(\boldsymbol{\beta}_1,..,\boldsymbol{\beta}_m,\gamma_0,\gamma_1,..,\gamma_l),$$

where $\boldsymbol{1}$ denotes a $q-$dimensional vector with each entry as 1, γ_0 is the intercept and $\gamma_1,...,\gamma_l \in \mathbb{R}$ are coefficients corresponding to the auxiliary predictors. Here, $N_q(\cdot,\cdot)$ stands for a $q-$variate normal distribution and the $q-$dimensional parameter $\boldsymbol{\beta}_s$ is envisioned as the upper triangular vector of a $p \times p$ symmetric matrix $\boldsymbol{B}_s = ((B_{s,\boldsymbol{j}}))$, $s = 1,...,l$, i.e., $\boldsymbol{\beta}_s = (B_{s,\boldsymbol{j}} : 1 \le j_1 < j_2 \le p)$. Equation (1) can be seen as a mixture of undirected network response regression models with the mixing distribution given by $G(\cdot)$.

The random probability measure $G(\cdot)$ is taken to be a discrete distribution of the form $G = \sum_{h=1}^{H}\omega_h\delta_{\boldsymbol{\Delta}_h^*}$, with atoms $\boldsymbol{\Delta}_h^* = (\boldsymbol{\beta}_{1,h}^*,..,\boldsymbol{\beta}_{m,h}^*,\gamma_{0,h}^*,\gamma_{1,h}^*,..,\gamma_{l,h}^*) \sim G_0$. Here, G_0 is the base measure and $\delta_{\boldsymbol{\Delta}_h^*}$ corresponds to the Dirac-delta function at $\boldsymbol{\Delta}_h^*$. Such a specification contains a broad class of species sampling priors, including the Dirichlet process (DP) prior and the Pitman-Yor process prior through the popular stick breaking construction (Sethuraman, 1994). In this work, we adopt the stick breaking construction to jointly model cluster inclusion probabilities. More precisely, for $h = 1,...,H-1$, and $\alpha > 0$,

(2)

$$\omega_1 = v_1^*,\ \omega_2 = v_2^*(1-v_1^*),..,\omega_{H-1} = v_{H-1}^*\prod_{h=1}^{H-2}(1-v_h^*),\ \omega_H = \prod_{h=1}^{H-1}(1-v_h^*),\ v_h^* \sim Beta(1,\alpha),$$

where H is an upper bound on the number of clusters. As $H \to \infty$, this choice leads to the classical Dirichlet process prior (Ishwaran and James, 2002). The parameter α is crucial in determining the number of clusters and is assigned a $Gamma(a_\alpha, b_\alpha)$ prior distribution.

From (1) and the discrete prior on G imposed by the stick breaking construction, the conditional distribution of $\boldsymbol{y}_i$ can be written as

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{z}_i,\sigma^2) = \sum_{h=1}^{H}\omega_h \mathrm{N}_q(\boldsymbol{y}_i|\boldsymbol{1}\gamma_{0,h}^* + \sum_{s=1}^{m}\boldsymbol{\beta}_{s,h}^* x_{is} + \boldsymbol{1}\sum_{s=1}^{l}\gamma_{s,h}^* z_{is}, \sigma^2\boldsymbol{I}_q). \tag{3}$$

Note that the mixture components signify different relationships between the network response and scalar predictors in H different clusters. Introducing a cluster index $c_i \in \{1,..,H\}$ corresponding to the individual i, we obtain $\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{z}_i,c_i,\sigma^2 \sim \mathrm{N}_q(\boldsymbol{y}_i|\mathbf{1}\gamma^*_{0,c_i} + \mathbf{1}\sum_{s=1}^{l}\gamma^*_{s,c_i}z_{is} + \sum_{s=1}^{m}\boldsymbol{\beta}^*_{s,c_i}x_{is},\sigma^2)$, with $P(c_i = h) = \omega_h$, for $h = 1,...,H$. This conditional independence structure, given the cluster indices of the individuals, facilitates computation, while still allowing a flexible dependence structure among the different components marginally. Additionally, inference on cluster indices determines the number of clusters and the constitution of each cluster.

Next, we turn to identifying network nodes in different clusters significantly associated with the predictors of interest. For this purpose, we first introduce a low-rank structure of the coefficient $\boldsymbol{B}^*_{s,h}$ corresponding to the sth predictor of interest in the hth cluster as

$$(4) \qquad B^*_{s,h,\boldsymbol{j}} = \sum_{r=1}^{R}\lambda_{s,h,r}u^{(r)}_{s,h,j_1}u^{(r)}_{s,h,j_2}, \;\; h = 1,...,H;\;\; s = 1,..,m,\;\; 1 \leq j_1 < j_2 \leq p.$$

Here $\boldsymbol{u}_{s,h,k} = (u^{(1)}_{s,h,k},...,u^{(R)}_{s,h,k})' \in \mathbb{R}^R$, for $k = 1,...,p$, is a collection of R-dimensional h-th mixture specific latent variables, one for each node and each predictor of interest, such that $\boldsymbol{u}_{s,h,k}$ corresponds to node k and predictor x_s in the h-th mixture component. Here, $\lambda_{s,h,r} \in \{0,1\}$ is the binary inclusion variable determining if the rth summand in (4) is relevant in model fitting in the hth mixture component. Drawing intuition from the random dot product graph models (Young and Scheinerman, 2007), we can interpret the latent vectors $\boldsymbol{u}_{s,h,1},\ldots,\boldsymbol{u}_{s,h,p}$ as the positions of the nodes in a latent space, with the strength of the association $\boldsymbol{B}^*_{s,h}$ being controlled by the inner product or the angular distance between the vectors. We expect the matrix of coefficients $\boldsymbol{B}^*_{s,h}$ (which itself can be regarded as describing a weighted network) to exhibit transitivity effects, i.e., we expect that if the interactions between regions j_1 and j_2 and between regions j_2 and j_3 are both influentially related to the sth predictor of interest, the interaction between regions j_1 and j_3 is likely to be influential as well (e.g., see Li et al., 2013). The structure proposed in (4) is commonly used to model social and biological networks because of its ability to capture these transitive effects. The assumed low-rank structure on $\boldsymbol{B}^*_{1,h},...,\boldsymbol{B}^*_{m,h}$ additionally offers parsimony by reducing the number of estimable parameters from mHq to $mHRp$, typically with $R \ll p$.

To infer on the network nodes significantly related to the predictors of interest in each cluster, we assign a spike-and-slab prior on node specific latent variables as below

$$(5)$$
$$\boldsymbol{u}_{s,h,k} \sim \begin{cases} N(\mathbf{0},\boldsymbol{M}_{s,h}), & \text{if } \xi_{s,h,k} = 1 \\ \delta_{\mathbf{0}}, & \text{if } \xi_{s,h,k} = 0 \end{cases},\;\; \xi_{s,h,k} \sim Ber(\zeta_{s,h}),\;\; \boldsymbol{M}_{s,h} \sim IW(\nu,\boldsymbol{I}),\;\; \zeta_{s,h} \sim Beta(a,b).$$

Here $\boldsymbol{M}_{s,h}$ is a covariance matrix of order $R \times R$. The parameter $\zeta_{s,h}$ corresponds to the probability of the nonzero mixture component in (5). Importantly, $\xi_{s,h,k} = 0$ implies that the kth network node in the response is not related to the sth predictor in the hth cluster of subjects. It needs to be emphasized that the model is invariant to the rotation of latent variables $\boldsymbol{u}_{s,h,k}$'s, and hence these latent variables are not directly identifiable. However, our inferential objective of identifying the set of nodes $\{k : \boldsymbol{u}_{s,h,k} = \mathbf{0}\}$ which are not significantly related to the sth predictor is achievable since a $\mathbf{0}$-valued latent vector is invariant under rotation. The parameters $\gamma^*_{0,h},\gamma^*_{1,h},...,\gamma^*_{l,h}$ are assigned standard normal distributions a-priori. We assign a hierarchical prior $\lambda_{s,h,r} \sim Ber(\pi_{s,h,r})$, $\pi_{s,h,r} \sim Beta(1,r^\eta)$, $\eta > 1$, and σ^2 is assigned an IG(a_σ,b_σ) prior. With the construction specified as above, the form of the base measure G_0 can be expressed as $G_0(\boldsymbol{\Delta}^*_h|\sigma^2) = \prod_{s=0}^{l}G_{0,1}(\gamma^*_{s,h}|\sigma^2)\prod_{s=1}^{m}G_{0,2}(\boldsymbol{\beta}^*_{s,h}|\sigma^2)$, where

$G_{0,1}(\gamma^*_{s,h}|\sigma^2) = N(0,1)$, and $G_{0,2}(\boldsymbol{\beta}^*_{s,h}|\sigma^2)$ is expressed as follows:

$$G_{0,2}(\boldsymbol{\beta}^*_{s,h}|\sigma^2) = \int \prod_{k=1}^{p} \pi(\boldsymbol{u}_{s,h,k}) \prod_{r=1}^{R} \pi(\lambda_{s,h,r}) \prod_{r=1}^{R} d\lambda_{s,h,r} \prod_{k=1}^{p} d\boldsymbol{u}_{s,h,k}.$$

The model and prior specification allow clustering of individuals into a number of groups less than or equal to H. In each group, the network response and the scalar predictors share separate regression structures, and thus subjects belonging to different clusters may have different sets of network nodes significantly related to the predictors of interest, as desired.

4. Posterior Computations. While fitting our proposed mixture model, we adopt a moderately large choice of H. Note that, according to Rousseau and Mengersen (2011), a similar choice of prior as ours is effective in the deletion of redundant mixture components not needed to characterize the data. If H is chosen to be too small, then none of the clusters will be unoccupied, and the analysis should be repeated for a larger H. Since all parameters except α have full conditional posterior distributions lying in standard families of distributions, Gibbs sampling with Metropolis is implemented to empirically estimate the posterior distributions. Details of the Markov chain Monte Carlo algorithm are presented in Appendix A. 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.

To assess inference from the proposed mixture model, we look at (i) the point estimate of cluster membership indices denoted by $\hat{\boldsymbol{c}}$, (ii) a heatmap of the posterior probability of any two samples belonging to the same cluster, $P(c_i = c_j|\boldsymbol{y})$ (which provides a measure of the uncertainty associated with the clustering), and (iii) a histogram of the posterior distribution of the number of identified clusters. The point estimate $\hat{\boldsymbol{c}}$ is obtained by minimizing (using iterative componentwise optimization) the expected loss function discussed in Lau and Green (2007),

$$F(\hat{\boldsymbol{c}}) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} 1(\hat{c}_i = \hat{c}_j) \left[\frac{w_2}{w_1 + w_2} - P(c_i = c_j|\boldsymbol{y}) \right], \tag{6}$$

where the ratio w_1/w_2 controls the relative loss of incorrectly clustering or separating a pair of samples. In our illustrations we set $w_1/w_2 = 1$. The posterior inference is based on 10000 suitably thinned samples from the MCMC sampler after a burn-in of 10000 samples. The time to compute 20000 MCMC iterations with $V = 20$ and $V = 50$ nodes (both with $H = 15$ and $n = 100$) took around 5.31 hours and 20.83 hours, respectively. All simulation examples and the real data example show very good convergence of the MCMC chain with fairly uncorrelated post burn-in MCMC iterates. The average effective sample size (ESS) for coefficients corresponding to the predictors of interest are provided for the simulation studies and the real data analysis.

5. Simulation Studies. This section studies the relative performance of our proposed network response mixture model (NRMM) vis-a-vis its competitors. To study all competitors under various data generation schemes, we simulate the response $\boldsymbol{y}_i$ depending on the predictors $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ from the finite mixture model given by

$$\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i \sim \sum_{h=1}^{H_0} \omega_{h,0} N(\boldsymbol{1}\gamma^*_{0,h,0} + \sum_{s=1}^{m} \boldsymbol{\beta}^*_{s,h,0} x_{is} + \boldsymbol{1} \sum_{s=1}^{l} \gamma^*_{s,h,0} z_{is}, \sigma_0^2 \boldsymbol{I}_q), \tag{7}$$

TABLE 1

Table presents specifications of Cases 1-7 in the simulation study. The parameter H_0 refers to the true number of mixture components in the Bayesian network response mixture model (NRMM). Different cases also present various combinations of the number of network nodes p, sample size n, network node sparsity $(1-\pi_0)$, true (R_g) and fitted dimensions (R) of the node specific latent variables.

Cases	p	n	R_g	R	$(1-\pi_0)$	H_0
1	20	100	2	5	0.6	3
2	20	100	2	5	0.3	3
3	20	100	3	5	0.6	4
4	50	100	2	5	0.6	3
5	50	100	2	5	0.3	3
6	50	100	3	5	0.6	2
7	20	100	2	5	0.6	1

where $\boldsymbol{\beta}^*_{s,h,0}$, $h = 1,...,H_0$ are mixture specific coefficients for x_{is}. The parameter $\gamma^*_{0,h,0}$ is the hth mixture specific intercept and $\gamma^*_{1,h,0},...,\gamma^*_{l,h,0}$ are the hth mixture specific coefficients corresponding to $z_{i1},...,z_{il}$, respectively. We set $m = 1$ and $l = 2$ for the simulations, which mimics the real data application scenario. Since $m = 1$, the subscript s will be omitted from variables related to the predictor of interest hereon. The predictors x_i , z_{i1} and z_{i2} are simulated i.i.d. from N(0,1).

To simulate the coefficients $\boldsymbol{\beta}^*_{h,0}$, we draw p latent variables $\boldsymbol{u}_{h,k,0}$, each of dimension R_g, from a mixture distribution given by

$$\boldsymbol{u}_{h,k,0} \sim \pi_0 N_{R_g}(\boldsymbol{u}_{h,m,g}, u^2_{h,v,g}) + (1-\pi_0)\delta_{\mathbf{0}};\ k \in \{1,...,p\}, \tag{8}$$

where $(1-\pi_0)$ is the probability of any $\boldsymbol{u}_{h,k,0}$ being zero in the truth, $h = 1,...,H_0$, and is referred to as the *network node sparsity*. We consider nine simulation cases as following:

Cases 1-7: In Cases 1-7, we assume $\boldsymbol{\beta}^*_{h,0}$ is the upper triangular vector of a symmetric matrix $\boldsymbol{B}^*_{h,0}$, i.e., $\boldsymbol{\beta}^*_{h,0} = (B^*_{h,0,\boldsymbol{j}} : j_1 < j_2)'$. The $\boldsymbol{j} = (j_1, j_2)$th element $(j_1 < j_2)$ of $\boldsymbol{B}^*_{h,0}$ corresponding to the h-th mixture component is constructed using a low-rank approach $B^*_{h,0,\boldsymbol{j}} = \boldsymbol{u}'_{h,j_1,0}\boldsymbol{u}_{h,j_2,0}$, accounting for the interaction between the j_1th and j_2th network nodes, for all $h = 1,...,H_0$. The 7 different cases are obtained by varying the number of true mixture components (H_0), number of network nodes (p), sample size (n), true dimension of latent variables (R_g), fitted dimension of latent variables (R) and network node sparsity $(1-\pi_0)$, as summarized in Table 1.

Case 8: In Case 8, we consider $H_0 = 2$, $\omega_{1,0} = 0.4, \omega_{3,0} = 0.6$, and $\boldsymbol{\beta}^*_{1,0}$ and $\boldsymbol{\beta}^*_{2,0}$ are simulated using two different strategies as following:

Simulating $\boldsymbol{\beta}^*_{1,0}$: The $\boldsymbol{j} = (j_1, j_2)$th element $(j_1 < j_2)$ of $\boldsymbol{B}^*_{1,0}$ is constructed using a low-rank approach $B^*_{1,0,\boldsymbol{j}} = \boldsymbol{u}'_{1,j_1,0}\boldsymbol{u}_{1,j_2,0}$, where the sparsity $(1-\pi_0)$ in generating the latent variables is set at 0.6.

Simulating $\boldsymbol{\beta}^*_{2,0}$: Randomly set $(1-\pi_0) = 0.6$ proportion of elements in $\boldsymbol{\beta}^*_{2,0}$ to be zero, and the rest are simulated from $N(0,1)$.

Case 9: Case 9 uses an identical construct as described in Case 8, except that $(1-\pi_0)$ is set at 0.3.

The intercept $\gamma^*_{s,h,0}$, $h = 1,...,H_0$, $s = 1,2$ in each mixture component is drawn from $N(-2,2)$, while σ^2_0 is fixed at 0.5.

In all cases, each component of the mean vector $\boldsymbol{u}_{h,m,g}$ is randomly generated to lie between $(-2,2)$ and the standard deviation $u_{h,v,g}$ is set randomly at a number between 0.3 and 2.

Notably, **Cases 1-7** represent the true model being included in the class of fitted models. In contrast, **Cases 8** and **9** show departure of the true model from the fitted models. This will allow assessment of the performance of our approach under model mis-specification.

5.1. *Choice of Hyper-parameters.* All simulation studies and the real data analysis are presented with the hyper-parameters chosen as $a = 1, b = 1, a_\sigma = 1, b_\sigma = 1$ and $\nu = 20$. The choice of $a_\sigma = b_\sigma = 1$ ensures that the prior on σ^2 is sufficiently flat with an infinite mean. The choice of $a = b = 1$ leads to a-priori uniform distribution on the number of network nodes related to each predictor in each cluster. Setting $\nu = 20$ implies that the prior distribution of $\boldsymbol{M}_h$ is concentrated around a scaled identity matrix. Since the model is invariant to rotations of the latent positions $\boldsymbol{u}_{h,k}$, the prior on $\boldsymbol{u}_{h,k}$'s should ideally be invariant under rotation. Centering $\boldsymbol{M}_h$ around a matrix that is proportional to the identity satisfies such a requirement. Finally, we choose a_α, b_α following Escobar and West (1995) such that the mean number of clusters is approximately 2.5 a priori. Since in most applications of the mixture model, the true number of clusters is small, our choice of a_α and b_α present a reasonable prior belief. Moderately perturbing hyper-parameters yields practically identical inference, as described in Section 5.4.

5.2. *Competitors and Metrics of Evaluation.* NRMM is fitted in all simulations with $H = 15$ mixture components. As a competitor to our model, we employ the *network response regression* (NRR), which is essentially our proposed framework with only one mixture component, i.e., $H = 1$. Thus NRR assumes (a) the same set of network nodes is significantly related to the predictors of interest for every individual, and, (b) normality for the distribution of each cell in the network response. Comparison with NRR will highlight any relative advantages of NRMM when these assumptions do not hold true. Additionally, we compare our approach with a frequentist higher order low-rank regression (HOLRR) method (Rabusseau and Kadri, 2016) popularly used in machine learning.

The competitors are assessed based on their ability to estimate the true regression mean function $E_0[\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i] = \sum_{h=1}^{H_0} \omega_{h,0} \left(\mathbf{1}\gamma^*_{0,h,0} + \mathbf{1}\sum_{s=1}^{l} \gamma^*_{s,h,0} z_{is} + \sum_{s=1}^{m} \boldsymbol{\beta}^*_{s,h,0} x_{is}\right)$. In particular, we compute the mean squared error (MSE) of estimating the true regression mean function over all data points, given by $\frac{1}{nq}\sum_{i=1}^{n} ||E_0[\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i] - E[\widehat{\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i}]||^2$, where $E[\widehat{\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i}]$ denotes the posterior mean of the regression function from a competing method. While MSE offers an evaluation of the point estimation by competitors, the uncertainty in estimating the true regression mean function is measured using the coverage and length of 95% credible intervals obtained from NRMM and NRR. We do not report coverage and length of 95% credible intervals from HOLRR since they are not readily available.

In addition to reporting the posterior distribution of the number of clusters and the uncertainty associated with clustering through $P(c_i = c_j|\boldsymbol{y})$, we also evaluate the ability of the models to identify clusters using the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) of the posterior cluster configurations with respect to the known cluster configuration. The ARI evaluates the agreement in cluster assignment between two cluster configurations. It ranges between -1 and 1, with larger values indicating more agreement between cluster configurations.

5.3. *Simulation Results.* All simulation examples show fairly uncorrelated post burn-in samples for drawing posterior inference. In fact, the effective sample size for 10000 post burn-in samples in simulation cases $1 - 9$ are found to be 8006, 7985, 7942, 7235, 7451, 7324, 8106, 8195 and 7839, respectively. Table 2 and Figure 2 provide insights into the estimates of the cluster structure and associated uncertainty by displaying the discrepancy between the true and estimated number of clusters and heat maps of posterior probabilities of pairs of subjects belonging to the same cluster. To facilitate visualization in Figure 2, subjects are ordered according to their true cluster configurations in the heatmap. In all cases, the model successfully recovers the true cluster structure, with little uncertainty associated

TABLE 2

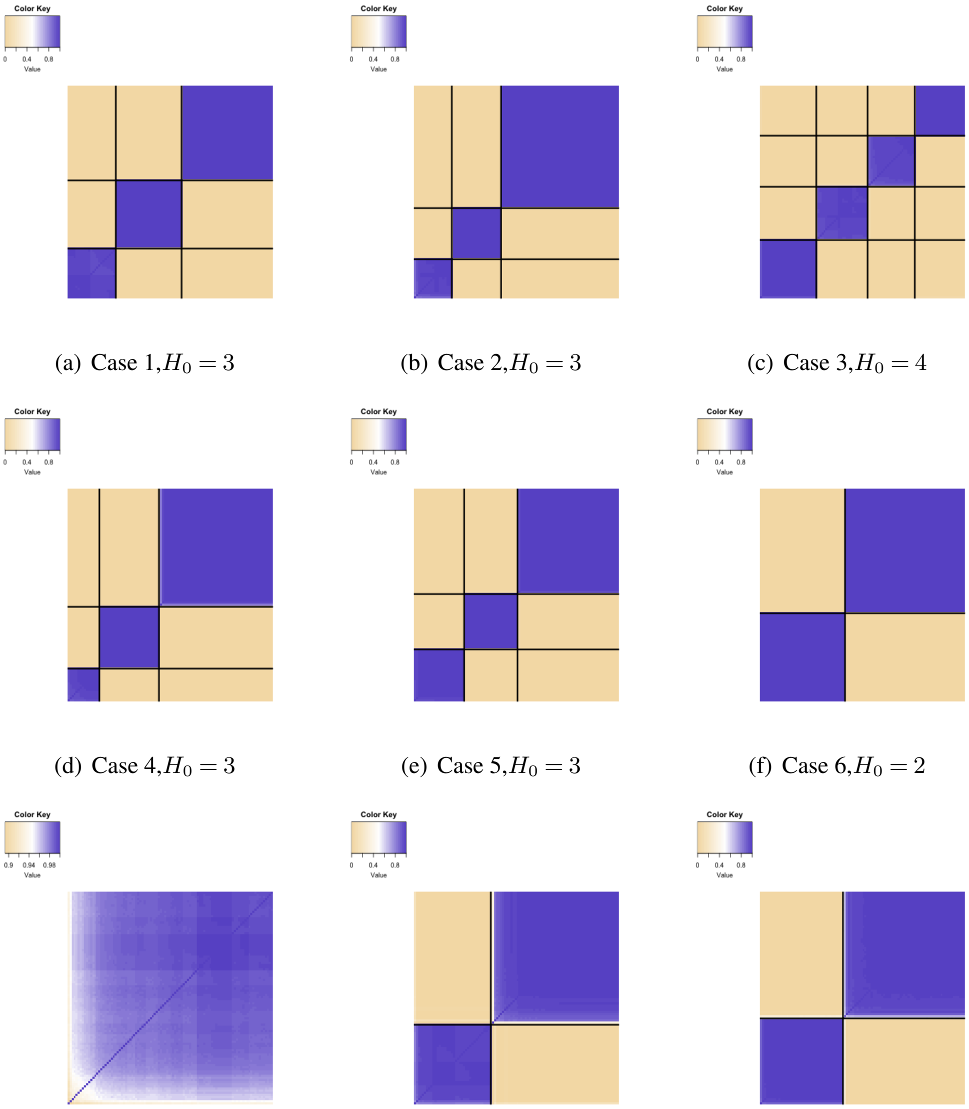
The second column presents ARI values to assess the clustering accuracy of NRMM. The next two columns present True Positive Rates (TPR) and False Positive Rates (FPR) in identifying network nodes related to the predictor of interest in NRMM. Mean Squared Error (MSE) for NRMM, NRR and HOLRR are presented for cases 1-9. The lowest MSE in each case is boldfaced. Coverage and length of 95% credible interval are provided for NRMM and NRR only, since the corresponding values for HOLRR are not readily available.

	NRMM				Competitors		
Case	ARI	TPR	FPR		NRMM	NRR	HOLRR
1	0.99	0.87	0.08	MSE	**0.02**	0.40	0.08
				Coverage of 95% CI	0.89	0.02	–
				Length of 95% CI	0.54	0.22	–
2	0.99	0.90	0.05	MSE	**0.03**	0.94	0.14
				Coverage of 95% CI	0.96	0.05	–
				Length of 95% CI	0.58	0.44	–
3	0.98	0.71	0.00	MSE	**0.14**	0.32	0.44
				Coverage of 95% CI	0.69	0.29	–
				Length of 95% CI	0.64	0.39	–
4	0.99	0.95	0.02	MSE	**0.01**	0.07	0.09
				Coverage of 95% CI	0.99	0.15	–
				Length of 95% CI	0.47	0.15	–
5	0.99	0.93	0.02	MSE	**0.04**	0.06	0.11
				Coverage of 95% CI	0.93	0.44	–
				Length of 95% CI	0.55	0.34	–
6	0.99	1.00	0.00	MSE	**0.05**	0.30	0.17
				Coverage of 95% CI	0.99	0.10	–
				Length of 95% CI	0.61	0.28	–
7	0.97	0.92	0.00	MSE	0.12	**0.008**	0.40
				Coverage of 95% CI	0.86	0.97	–
				Length of 95% CI	0.37	0.07	–
8	0.93	–	–	MSE	**0.10**	1.30	0.13
				Coverage of 95% CI	0.84	0.07	–
				Length of 95% CI	0.51	0.36	–
9	0.95	–	–	MSE	**0.17**	0.54	0.19
				Coverage of 95% CI	0.74	0.09	–
				Length of 95% CI	0.70	0.39	–

with the estimator. The most challenging cases among all are cases 8 and 9, which correspond to model mis-specification. Even with model mis-specification, there is a minor deterioration in the performance, with ARI dropping to around 0.93 in case 8 and 0.95 in case 9. It appears from Figure 2 that the clustering performance improves nominally with decreasing sparsity of $\boldsymbol{\beta}^*_{h,0}$, the impact of sparsity being a little more prominent under model mis-specification (compare cases 8 and 9). The uncertainty in clustering for a few individuals also appears to be higher in case 7, where the true data generating model sets $H_0 = 1$.

The posterior distributions of the number of identified clusters are also presented in the form of barplots in Figure 3. Consistent with the story presented so far, the posterior distribution of the number of clusters appears to concentrate around the true number of clusters H_0 in all cases except case 8, where the model overestimates the number of clusters. Notably, case 8 corresponds to model mis-specification with a higher node sparsity parameter $(1 - \pi_0)$. As the node sparsity parameter $(1 - \pi_0)$ decreases, the posterior distribution of the number of clusters concentrates around H_0 even under model mis-specification (case 9). The results also reveal a somewhat bi-modal structure of the posterior distribution of the number of clusters under cases 3 (with $H_0 = 4$) and 7 (with $H_0 = 1$). Importantly, out of H assigned clusters, most are not populated in each case, justifying the choice of $H = 15$ in each case.

Table 2 presents mean squared errors (MSE) for estimating the regression mean function under each of the competitors. Further, coverage and average length of 95% credible intervals



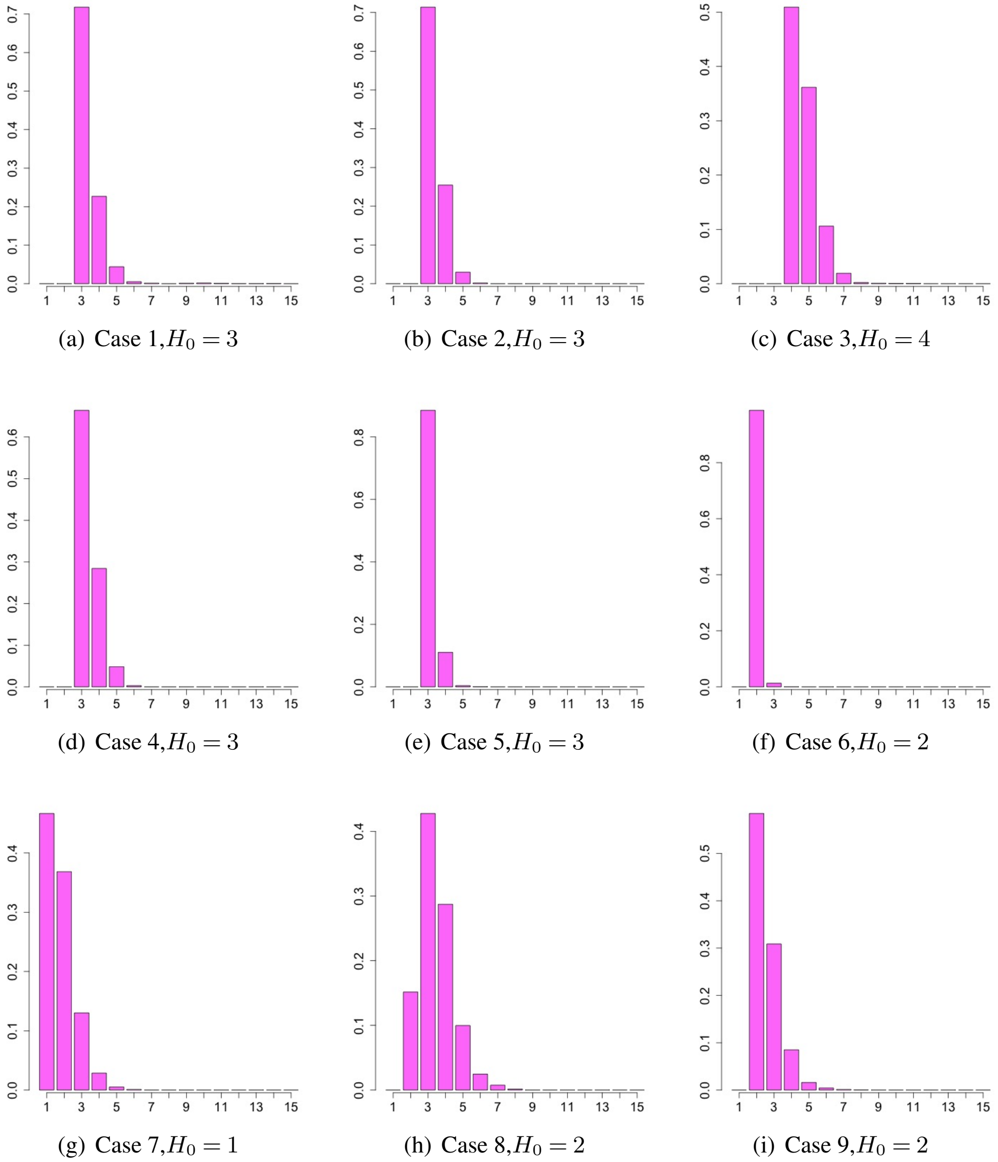
(a) Case 1,$H_0 = 3$ (b) Case 2,$H_0 = 3$ (c) Case 3,$H_0 = 4$

(d) Case 4,$H_0 = 3$ (e) Case 5,$H_0 = 3$ (f) Case 6,$H_0 = 2$

(g) Case 7,$H_0 = 1$ (h) Case 8,$H_0 = 2$ (i) Case 9,$H_0 = 2$

FIG 2. *Plots showing uncertainty in estimating clusters in simulation cases 1-9. Boldfaced horizontal and vertical lines indicate the true clustering.*

are provided to assess the uncertainty quantification from NRMM and NRR. A few interesting observations emerge from Table 2. Comparing cases 1 and 2 (and also comparing cases 4 and 5), it turns out that NRMM yields marginally lower MSE with increased values of the sparsity parameter $(1 - \pi_0)$. Results from cases 8 and 9 present a similar trend, even under model mis-specification. Also, keeping n fixed and increasing p moderately does not have any significant impact on MSE. Increasing the number of true mixture components H_0 has an adverse effect on the performance of NRMM, which becomes evident by comparing results from case 3 with cases 1 and 2. Additionally, in most cases, NRMM shows higher coverage levels, often close to nominal coverage, compared to NRR. The less than nominal coverage in cases 8 and 9 can be attributed to model mis-specification, whereas the under-coverage in case 3 could be due to the larger number of mixture components, which presents



(a) Case 1,$H_0 = 3$ (b) Case 2,$H_0 = 3$ (c) Case 3,$H_0 = 4$

(d) Case 4,$H_0 = 3$ (e) Case 5,$H_0 = 3$ (f) Case 6,$H_0 = 2$

(g) Case 7,$H_0 = 1$ (h) Case 8,$H_0 = 2$ (i) Case 9,$H_0 = 2$

FIG 3. *Plots showing the posterior distribution of the number of clusters in simulation cases 1-9.*

obstacles to model estimation. Note that under case 7, only one mixture component is used to simulate the data, and so the data favors NRR over NRMM. Consequently, NRR yields considerably smaller MSE and close to nominal coverage in this case. Under all other cases with $H_0 > 1$, NRR demonstrates inferior performance to NRMM with a higher MSE and considerable under-coverage of the mean function. HOLRR offers a higher MSE compared to NRMM under all simulation scenarios.

Note that inference on each cluster is not readily available from the mixture model due to the clusters being not identifiable. Thus, to draw inference on which network nodes are influential in each cluster, we fix the cluster membership indicator c_i for the ith sample at $\hat{c}_i$ (the estimated cluster indicator) and run the model once more without updating the cluster membership indicator c_i at any MCMC iteration. With the clusters remaining fixed in every iteration, it is possible to draw inference on the influential network nodes in each cluster. In

TABLE 3
ARI, MSE, coverage of 95% CI and length of 95% CI for NRMM under Case 2 with different hyper-parameter combinations are provided.

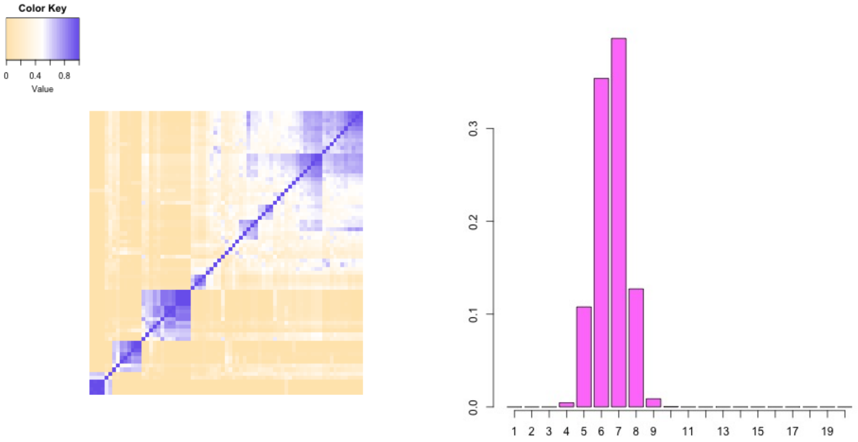
Combinations	$(i)\ a=1,b=5,\nu=20$	$(ii)\ a=5,b=1,\nu=20$	$(iii)\ a=1,b=1,\nu=50$
ARI	0.99	0.99	0.99
MSE	0.08	0.03	0.05
Coverage of 95% CI	0.93	0.96	0.95
Length of 95% CI	0.61	0.57	0.50

particular, the kth node is deemed influential for the hth cluster, if the empirically estimated posterior probability of the event $\{\boldsymbol{u}_{h,k} \neq \mathbf{0}\}$ exceeds 0.5. As demonstrated in Figures 2 and 3, for cases 1-7, our proposed model correctly identifies each cluster in every simulation, and hence inference on influential network nodes in each cluster as mentioned above can be directly compared to the truly influential nodes in each cluster for these simulation cases (i.e., under no model mis-specification). In this regard, Table 2 presents the True Positive Rates (TPR) and False Positive Rates (FPR) of identifying influential network nodes over all clusters. The results indicate high TPR and low FPR in all cases, except in case 3, which shows a comparatively lower TPR than the rest, but still a very low FPR. This observation may be attributed to a higher number of true clusters, where the model detects some influential nodes as uninfluential, resulting in decrease of TPR. Overall, the simulation studies indicate good performance of NRMM.

5.4. *Sensitivity Analysis.* To check the sensitivity of inference to the choice of hyper-parameters, we consider a representative case (case 2) and re-analyze the same simulated data with different combinations of hyper-parameters. In particular, we consider three different hyper-parameter settings for case 2 and compare the inference with the results on case 2 presented earlier. The three combinations are given by, (i) $a=1, b=5, \nu=20$; (ii) $a=5, b=1, \nu=20$; (iii) $a=1, b=1, \nu=50$. Notice that (i) presents a low prior mean of 0.2 for each $\xi_{h,k}$ encouraging less number of activated nodes a-priori, whereas (ii) presents a higher prior mean of 5 for $\xi_{h,k}$ which encourages a higher number of activated nodes. Additionally, (iii) presents a variation of the hyperparameter ν in the Inverse-Wishart distribution of $\boldsymbol{M}_h$. Table 3 shows the posterior mean of ARI in case 2 under the three different hyper-parameter settings. We additionally present MSE, coverage and length of 95% credible intervals for these hyper-parameter combinations and compare these results with the result presented for case 2 in Table 2. Of all the parameters, only variations in a and b seem to have an effect in the inferences, but this effect is found to be very small. More specifically, when the prior mean of the number of activated nodes is small (combination (i)), MSE is found to be a little higher than what is presented in Table 2 under case 2. Similarly, the coverage is found to be a little lower and length a little higher as compared to case 2 in Table 2. In contrast, combinations (ii) and (iii) yield practically identical results when compared to case 2 in Table 2. The clustering accuracy is found to be unaffected by the perturbation in hyper-parameters, with all three combinations resulting in similar values of ARI. The results are also found not to be sensitive at all to moderate perturbation of hyper-parameters a_σ and b_σ.

6. Findings from CAQ Brain Connectome Data.

This section reports the analysis of the CAQ brain connectome dataset described in Section 2. We fit NRMM with $H=20$, with the same set of hyper-parameters used in the simulation studies. NRMM, when applied to the CAQ dataset, identifies 7 clusters with 25, 13, 6, 6, 7, 8 and 8 subjects included in the clusters, respectively. Similar to simulation studies, the uncertainty in clustering is measured by the posterior probability of pairs of subjects lying in the same cluster, which is displayed through a heatmap in Figure 4(a). The figure indicates three distinct cluster assignments, with



(a) Uncertainty in Clustering

(b) Posterior Dist. of No. of Clusters

FIG 4. ***CAQ Data***: *Figure 4(a) shows the uncertainty in estimating the clusters. Figure 4(b) shows the barplot corresponding to the posterior distribution of the estimated number of clusters. The inference is presented for* $H = 20$.

TABLE 4

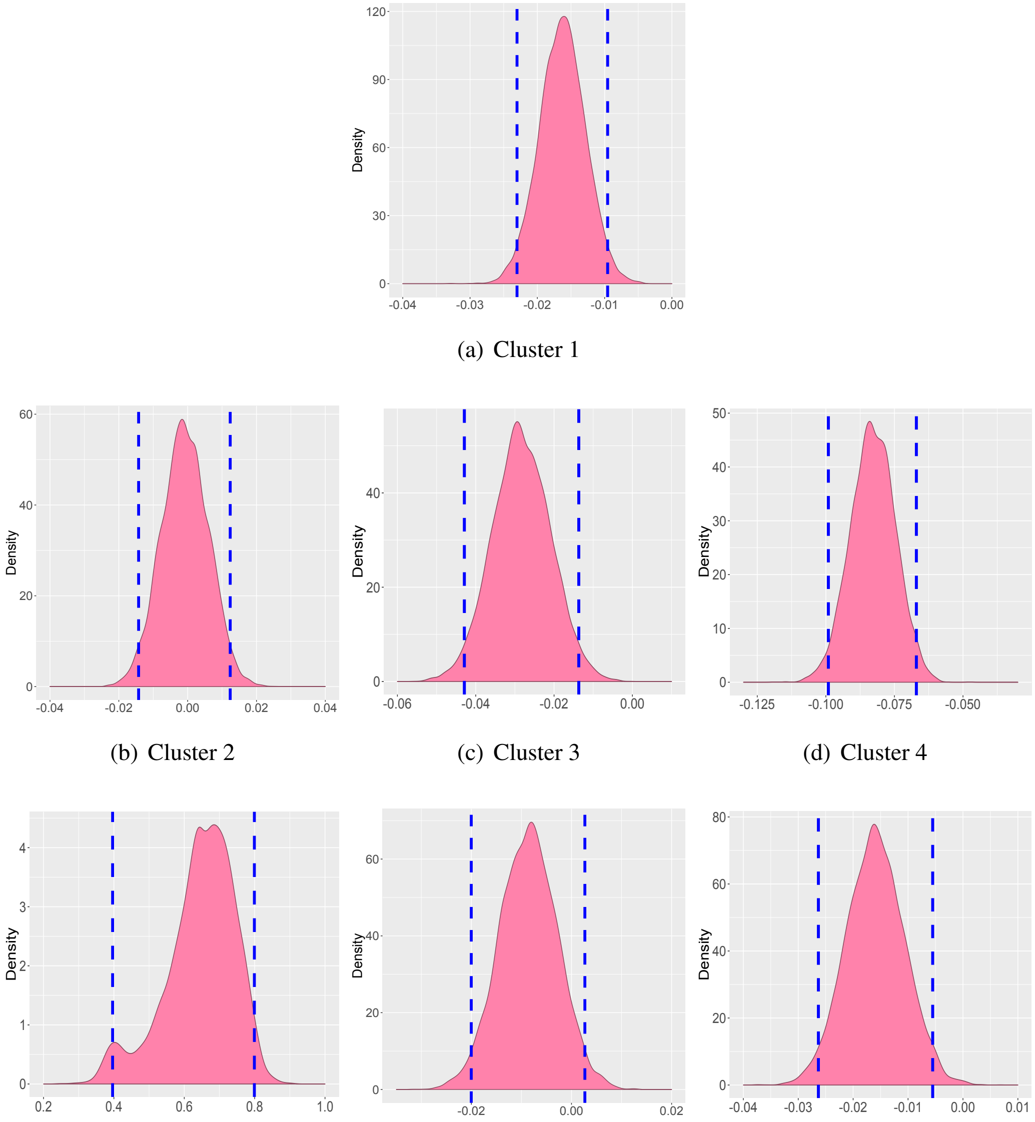
MSPE, average coverage of 95% predictive intervals and average length of 95% predictive intervals for the seven clusters are provided.

Cluster size	25	13	6	6	7	8	8
MSE	0.66	0.43	0.28	0.92	0.64	0.83	0.54
Coverage of 95% CI	0.95	0.97	0.97	0.94	0.95	0.94	0.96
Length of 95% CI	3.02	3.02	3.03	3.03	3.04	3.03	3.02

a somewhat higher degree of uncertainty among the pairs lying outside these three clusters. The posterior distribution of the number of clusters (see Figure 4(b)) demonstrates some bimodality with modes at 6 and 7. Importantly, there is no posterior probability of having more than 9 clusters, suggesting that $H = 20$ is appropriate for this analysis.

In the absence of any ground truth, we compare performances of NRMM and NRR with respect to the Posterior Predictive Loss Criterion statistic (Gelfand and Ghosh, 1998), which is calculated as $D = G + P$, such that a model corresponding to a lower value of D is preferred. The G values, representing a measure of model fit, turn out to be 98163.8 and 101738.7 for NRMM and NRR, respectively. The P values, indicative of model complexity, are 101722 and 101489.2 for NRMM and NRR, respectively. Thus, the overall model fitting statistic D shows a better performance of NRMM compared to NRR. HOLRR, being a frequentist method, is not included in this comparison. The effective sample size averaged over all cells of the network matrix coefficient of interest turns out to be 8270, indicating fairly uncorrelated post burn-in samples.

Similar to the simulation studies, we supply the model with the estimated cluster indicators and run it again to draw further inference on the influential nodes in the seven clusters. Notably, Cluster 3 includes individuals who are all male. Hence analysis of Cluster 3 does not include gender as a variable. To assess the model fit in each cluster, we calculate the mean squared prediction error (MSPE), average coverage of 95% predictive intervals and average length of 95% predictive intervals averaged over all cells of the network response matrix and all subjects in a cluster. Table 4 depicts satisfactory point prediction along with very good characterization of predictive uncertainty. Referring to the presence of non-normality in the error distributions discussed in Section 2, it is instructive to see if the mixture mod-



(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

(d) Cluster 4

(e) Cluster 5

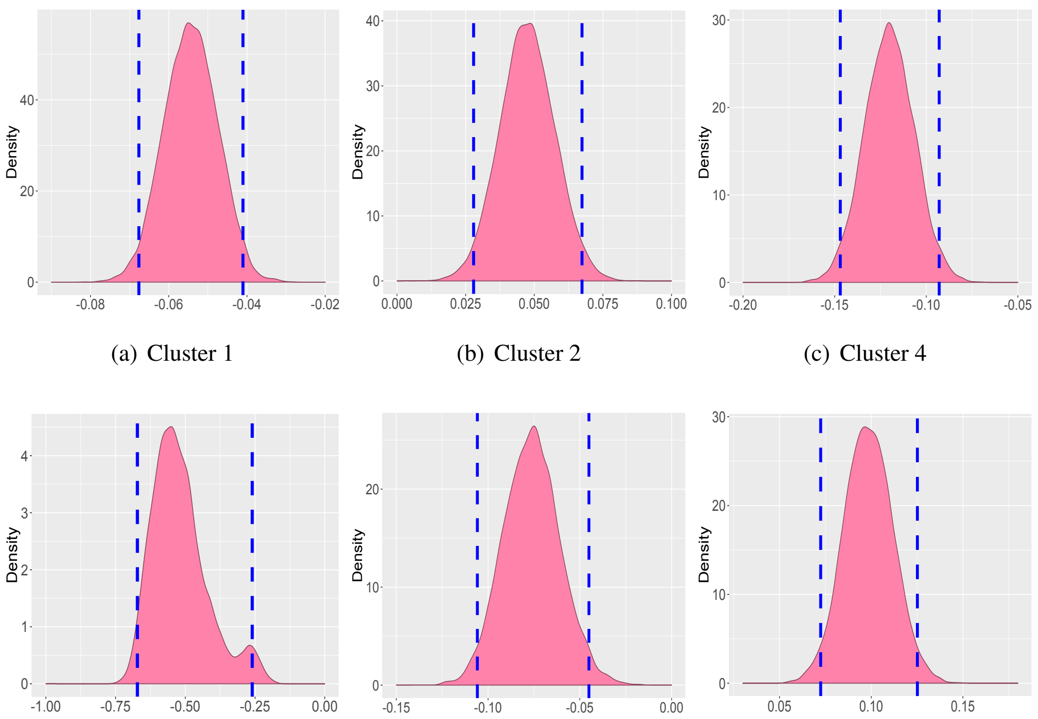
(f) Cluster 6

(g) Cluster 7

FIG 5. *Plots of age coefficient in each cluster. The 95% posterior credible interval is denoted by the interval between the two dotted lines.*

eling framework justifies the normality assumption on the error distribution in each cluster. To check this, cell-by-cell Kolmogorov-Smirnov tests are conducted by comparing the discrepancy between the posterior mean of residuals and the normal distribution. Out of 2278 network matrix cells in each cluster, residuals in $51\%, 62\%, 18\%, 96\%, 91\%, 89\%$ and 97% cells in clusters $1-7$, respectively, show statistically significant normality. Therefore, the normality assumption on the errors in each cluster is reasonable except for Cluster 3.

Figure 5 displays posterior densities of the age coefficients for all seven clusters. Except Cluster 2, all other age coefficients turn out to be significant. Digging a bit deeper, we find that Cluster 2 shows significantly lower variability in the ages of the subjects included compared to the other clusters, which explains the age coefficient being statistically insignificant in this cluster. Also, except Cluster 5, the posterior means of age coefficients are found to be negative in all other clusters, implying a negative association between creativity and age. Similarly, in



(a) Cluster 1

(b) Cluster 2

(c) Cluster 4

(d) Cluster 5

(e) Cluster 6

(f) Cluster 7

FIG 6. *Plots of sex coefficient in each cluster. The 95% posterior credible interval is denoted by the interval between the two dotted lines.*

all six clusters where gender is added as a variable, it is found to be significantly affecting the creativity (see Figure 6).

To assess which nodes are related to creativity (as measured by CAQ) in each cluster, we run the analysis in each cluster 10 times and report the nodes which have a posterior probability of being active greater than 0.5 for at least five of the replications. Figure 7 records the 10, 40, 30, 37, 41, 49 and 15 ROIs significantly related to CAQ for the 7 clusters of individuals. A considerable proportion of ROIs detected in each cluster are part of the *frontal*, *cingulate* and *temporal* lobes in both hemispheres. This finding concurs with results presented previously in the literature. The frontal lobe has been scientifically associated with divergent thinking, problem solving ability, spontaneity, memory, language, judgement, impulse control and social behavior (Stuss et al., 1985; Razumnikova, 2007; Miller and Milner, 1985; Kolb and Milner, 1981). Finkelstein, Vardi and Hod, 1991 also report *de novo* artistic expression to be associated with the frontal and temporal regions.

7. Conclusion and Future Work. This article is motivated by the need to develop a flexible relationship between the brain network and creativity, as measured by CAQ, from subjects in a brain connectome dataset. Viewing the brain image for each subject as an undirected network, we propose a novel Bayesian mixture of regression models with a network response and scalar predictors. Our proposed framework clusters subjects into groups, with individuals in the same group sharing an identical relationship between the network response and scalar predictors. A spike-and-slab variable selection prior is assigned on the network node specific latent variables in each mixture component to deliver inference on influential network nodes significantly related to a specific predictor of interest. Empirical investigations

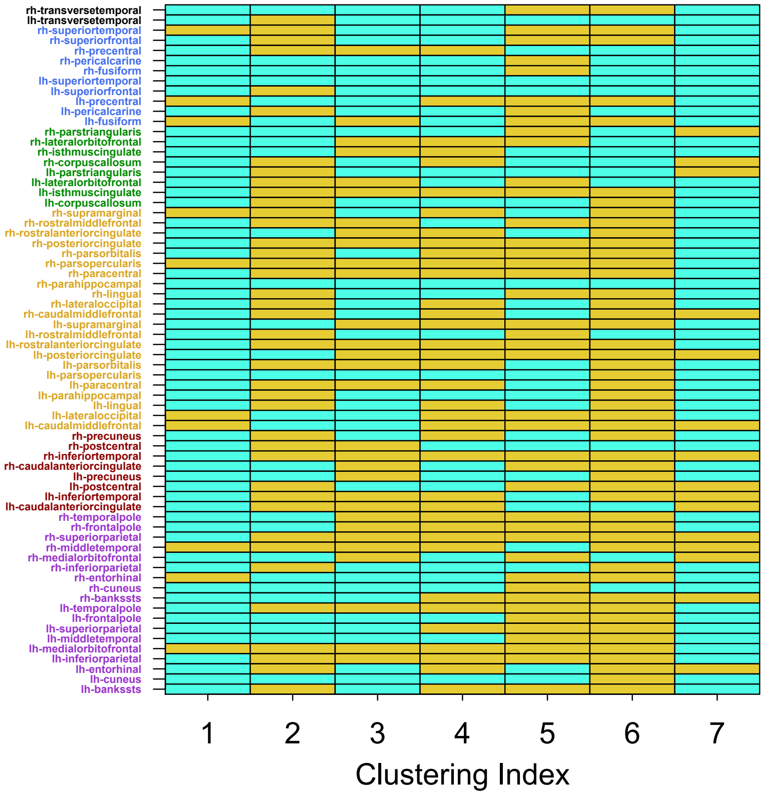


FIG 7. ***CAQ Data***: *Plot shows a 68×7 matrix, with the rows and columns corresponding to the ROIs and clusters, respectively. A green cell in the (k, h)th entry of the matrix implies that the kth ROI in the hth cluster is not significantly related to creativity. Prefix 'lh-' and 'rh-' in the ROI names on the y-axis denote their positions in the left and right hemispheres of the brain, respectively. The ROI names are color-coded according to the lobes they belong to. The groups of ROIs (according to color coding), from bottom to top, represent the* ***temporal***, ***cingulate***, ***frontal***, ***occipital***, ***parietal*** *and* ***insula*** *lobes, respectively.*

with simulation studies validate our network response mixture modeling (NRMM) framework and yield superior inference over relevant competitors. The NRMM framework, when applied to a real brain connectome dataset, finds clusters of individuals sharing similar relationships between their brain networks and creativity, identifying brain ROIs significantly related to creativity in each cluster.

As part of future work, we envision investigating the performance of our model with a more flexible non-local prior structure on the node specific latent variables. We also plan to extend our framework with each mixture component fitting a generalized linear model with a symmetric network/tensor response and scalar predictors.

APPENDIX A: POSTERIOR FULL CONDITIONALS

Let $\mathcal{I}_h = \{i : c_i = h\}$, n_h denote the cardinality of $\mathcal{I}_h$, and $\boldsymbol{y}_h = (\boldsymbol{y}_i : c_i = h)^T$, $h = 1, ..., H$. Further assume $\mathcal{J}_k = \{\boldsymbol{j} \in \mathcal{J} : j_{s_1} = k, \text{for some } s_1\}$. The full conditionals are in closed form and hence allow a Gibbs sampling procedure to sample posteriors. They are listed as the following:

- $\gamma^*_{0,h}| - \sim N\left[\frac{\sum_{i\in\mathcal{I}_h}\mathbf{1}^T(\boldsymbol{y}_i - \sum_{s=1}^m \boldsymbol{\beta}^*_{s,h}x_{is} - \mathbf{1}\sum_{s=1}^l \gamma^*_{s,h}z_{is})/\sigma^2}{(n_h q)/\sigma^2+1}, \frac{1}{(n_h q)/\sigma^2+1}\right], h = 1, ..., H.$

- $\gamma^*_{s,h}|- \sim N\left(\frac{\sum_{i\in\mathcal{I}_h} z_{is}^2 \mathbf{1}^T(\boldsymbol{y}_i - \sum_{h_2=1}^m \boldsymbol{\beta}^*_{h_2,h} x_{ih_2} - \mathbf{1}\sum_{h_2=1,h_2\neq s}^l \gamma^*_{h_2,h} z_{ih_2})/\sigma^2 + a_\beta/b_\beta}{q\sum_{i\in\mathcal{I}_h} z_{is}^2/\sigma^2 + 1/b_\beta}, \frac{1}{q\sum_{i\in\mathcal{I}_h} z_{is}^2/\sigma^2+1/b_\beta}\right)$, $s=1,...,l;\ h=1,...,H.$
- $\sigma^2|- \sim IG(a_\sigma + (nq)/2, b_\sigma + \sum_{h=1}^H \sum_{i\in\mathcal{I}_h} ||\boldsymbol{y}_i - \sum_{s=1}^m \boldsymbol{\beta}^*_{s,h} x_{is} - \mathbf{1}\sum_{s=1}^l \gamma^*_{s,h} z_{is}||^2/2)$
- $\boldsymbol{M}_{s,h}|- \sim IW\left[(\boldsymbol{S} + \sum_{k:\boldsymbol{u}_{s,h,k}\neq \mathbf{0}} \boldsymbol{u}_{s,h,k}\boldsymbol{u}_{s,h,k}{}^T), (\nu + \{\#k : \boldsymbol{u}_{s,h,k} \neq \mathbf{0}\})\right]$
- $\pi_{s,h,r}|- \sim Beta[(1+\lambda_{s,h,r}), (r^\eta + 1 - \lambda_{s,h,r})]$
- $\lambda_{s,h,r}|- \sim Ber(p_{s,h,r})$, where $p_{s,h,r} = \frac{\pi_{s,h,r} J(\boldsymbol{\Lambda}_{s,h})_{(\lambda_{s,h,r}=1)}}{\pi_{s,h,r} J(\boldsymbol{\Lambda}_{s,h})_{(\lambda_{s,h,r}=1)} + (1-\pi_{s,h,r}) J(\boldsymbol{\Lambda}_{s,h})_{(\lambda_{s,h,r}=0)}}$ and $J(\boldsymbol{\Lambda}_{s,h}) = \prod_{i\in\mathcal{I}_h} N(\boldsymbol{y}_i|\gamma^*_{0,h}\mathbf{1} + \sum_{s=1}^m \boldsymbol{\beta}^*_{s,h} x_{is} + \mathbf{1}\sum_{s=1}^l \gamma^*_{s,h} z_{is}, \sigma^2 I)$. $J(\boldsymbol{\Lambda}_{s,h})_{(\lambda_{s,h,r}=1)}$ denotes $J(\boldsymbol{\Lambda}_{s,h})$ evaluated at $\lambda_{s,h,r} = 1$. Here $\boldsymbol{\Lambda}_{s,h}$ is the collection of $\{\lambda_{s,h,r} : r = 1,...,R\}$.
- $\boldsymbol{u}_{s,h,k}|- \sim w_{\boldsymbol{u}_{s,h,k}} \delta_0(\boldsymbol{u}_{s,h,k}) + (1 - w_{\boldsymbol{u}_{s,h,k}}) N(\boldsymbol{u}_{s,h,k}|\boldsymbol{m}_{\boldsymbol{u}_{s,h,k}}, \boldsymbol{\Sigma}_{\boldsymbol{u}_{s,h,k}})$, where $\boldsymbol{U}_{s,h,\mathcal{J}_k} = [\boldsymbol{U}^T_{1,s,h,\mathcal{J}_k} : \cdots : \boldsymbol{U}^T_{n_h,s,h,\mathcal{J}_k}]^T$, $\boldsymbol{U}^T_{i,s,h,\mathcal{J}_k}$ has rows $(x_{is}\lambda_{s,h,1}\prod_{s_1=1,j_{s_1}\neq k}^D u^{(1)}_{s,h,j_{s_1}}, ..., x_{is}\lambda_{s,h,R}\prod_{s_1=1,j_{s_1}\neq k}^D u^{(R)}_{s,h,j_{s_1}})$. Further assume $\tilde{y}^s_{i,j} = y_{i,j} - \gamma^*_{0,h} - \sum_{h_1=1}^l \gamma^*_{h_1,h} z_{ih_1} - \sum_{h_2=1,h_2\neq s}^m \beta_{h_2,h,\boldsymbol{j}} x_{ih_2}$, $\tilde{y}^s_{i,\mathcal{J}_k}$ is a vector of collections of $\tilde{y}^s_{i,\boldsymbol{j}}$ over $\boldsymbol{j} \in \mathcal{J}_k$ and $\tilde{\boldsymbol{y}}^s_{\mathcal{J}_k}$ is a vector consisting of $\tilde{y}^s_{i,\mathcal{J}_k}$ over $i \in \mathcal{I}_h$. Also,

$$\boldsymbol{\Sigma}_{\boldsymbol{u}_{s,h,k}} = \left(\boldsymbol{U}^T_{s,h,\mathcal{J}_k}\boldsymbol{U}_{s,h,\mathcal{J}_k}/\sigma^2 + \boldsymbol{M}^{-1}_{s,h}\right)^{-1},\ \boldsymbol{m}_{\boldsymbol{u}_{s,h,k}} = \boldsymbol{\Sigma}_{\boldsymbol{u}_{s,h,k}}\boldsymbol{U}^T_{s,h,\mathcal{J}_k}\tilde{\boldsymbol{y}}^s_{\mathcal{J}_k}/\sigma^2$$

$$w_{\boldsymbol{u}_{s,h,k}} = \frac{(1-\zeta_{s,h})N(\tilde{\boldsymbol{y}}^s_{\mathcal{J}_k}|0,\sigma^2 I)}{(1-\zeta_{s,h})N(\tilde{\boldsymbol{y}}^s_{\mathcal{J}_k}|0,\sigma^2 I) + \pi N(\tilde{\boldsymbol{y}}^s_{\mathcal{J}_k}|0,\sigma^2\boldsymbol{I} + \boldsymbol{U}_{s,h,\mathcal{J}_k}\boldsymbol{M}_{s,h}\boldsymbol{U}^T_{s,h,\mathcal{J}_k})}$$

- $\xi_{s,h,k}|- \sim Ber(1 - w_{u_{s,h,k}})$
- $\zeta_{s,h}|- \sim Beta(\sum_{k=1}^p \xi_{s,h,k} + 1, \sum_{k=1}^p (1-\xi_{s,h,k}) + 1).$
- $P(c_i = h\,|-) = \frac{\omega_h N(\boldsymbol{y}_i|\gamma^*_{0,h}\mathbf{1} + \sum_{s=1}^m \boldsymbol{\beta}^*_{s,h} x_{is} + \mathbf{1}\sum_{s=1}^l \gamma^*_{s,h} z_{is}, \sigma^2 I)}{\sum_{d'=1}^H \omega_{d'} N(\boldsymbol{y}_i|\gamma^*_{0,d'}\mathbf{1} + \sum_{s=1}^m \boldsymbol{\beta}^*_{s,d'} x_{is} + \mathbf{1}\sum_{s=1}^l \gamma^*_{s,d'} z_{is}, \sigma^2 I)}$, for $h = 1,..,H$.
- $v^*_{l_1}\,| - Beta(1 + \#\{i : c_i = l_1\}, \alpha + \sum_{ss=l_1+1}^H \#\{i : c_i = ss\}),\ l_1 = 1,...,H-1,$ $\omega_1 = v^*_1$, $\omega_2 = v^*_2(1-v^*_1), .., \omega_{H-1} = v^*_{H-1}\prod_{l_1=1}^{H-2}(1-v^*_{l_1})$, $\omega_H = \prod_{l_1=1}^{H-1}(1-v^*_{l_1})$
- Parameter α is updated using a Metropolis-Hastings algorithm.

REFERENCES

Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews. Neuroscience* **10** 186–198.

Cao, X., Wei, X., Han, Y., Yang, Y. and Lin, D. (2013). Robust tensor clustering with non-greedy maximization. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

Carson, S. H., Peterson, J. B. and Higgins, D. M. (2005). Reliability, validity, and factor structure of the creative achievement questionnaire. *Creativity Research Journal* **17** 37–50.

Chi, E. C., Allen, G. I. and Baraniuk, R. G. (2017). Convex biclustering. *Biometrics* **73** 10–19.

Chi, E. C. and Lange, K. (2015). Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics* **24** 994–1013.

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** 968–980.

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

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association* **90** 577–588.

Finkelstein, Y., Vardi, J. and Hod, I. (1991). Impulsive artistic creativity as a presentation of transient cognitive alterations. *Behavioral medicine* **17** 91–94.

FRANK, O. and STRAUSS, D. (1986). Markov graphs. *Journal of the American Statistical Association* **81** 832–842.

GELFAND, A. E. and GHOSH, S. K. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika* **85** 1–11.

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., QAMAR, S. and DUNSON, D. B. (2018). Bayesian conditional density filtering. *Journal of Computational and Graphical Statistics* **27** 657–672.

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

HOFF, P. D. (2009). A hierarchical eigenmodel for pooled covariance estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71** 971–992.

HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association* **97** 1090–1098.

HUANG, J. Z., SHEN, H. and BUJA, A. (2009). The analysis of two-way functional data using two-way regularized singular value decompositions. *Journal of the American Statistical Association* **104** 1609–1620.

HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *Journal of classification* **2** 193–218.

ISHWARAN, H. and JAMES, L. F. (2002). Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical statistics* **11** 508–532.

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** 398–409.

KIAR, G., GORGOLEWSKI, K. and KLEISSAS, D. (2017). Example Use Case of sic with the ndmg Pipeline (sic: ndmg). *GigaScience Database*.

KIAR, G., GRAY RONCAL, W., MHEMBERE, D., BRIDGEFORD, E., BURNS, R. and VOGELSTEIN, J. (2016). ndmg: NeuroData's MRI graphs pipeline.

KIAR, G., GORGOLEWSKI, K. J., KLEISSAS, D., RONCAL, W. G., LITT, B., WANDELL, B., POLDRACK, R. A., WIENER, M., VOGELSTEIN, R. J., BURNS, R. et al. (2017). Science In the Cloud (SIC): A Use Case in MRI Connectomics. *Giga Science* **6** 1–10.

KOLB, B. and MILNER, B. (1981). Performance of complex arm and facial movements after focal brain lesions. *Neuropsychologia* **19** 491–503.

LAU, J. W. and GREEN, P. J. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics* **16** 526–558.

LEE, M., SHEN, H., HUANG, J. Z. and MARRON, J. (2010). Biclustering via sparse singular value decomposition. *Biometrics* **66** 1087–1095.

LI, L. and ZHANG, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association* **112** 1131–1146.

LI, Y., QIN, Y., CHEN, X. and LI, W. (2013). Exploring the functional brain network of Alzheimer's disease: based on the computational experiment. *PloS one* **8** e73186.

LI, R., ZHANG, W., ZHAO, Y., ZHU, Z. and JI, S. (2014). Sparsity learning formulations for mining time-varying data. *IEEE Transactions on Knowledge and Data Engineering* **27** 1411–1423.

MESKALDJI, D. E., FISCHI-GOMEZ, E., GRIFFA, A., HAGMANN, P., MORGENTHALER, S. and THIRAN, J.-P. (2013). Comparing connectomes across subjects and populations at different scales. *NeuroImage* **80** 416–425.

MESKALDJI, D.-E., VASUNG, L., ROMASCANO, D., THIRAN, J.-P., HAGMANN, P., MORGENTHALER, S. and VAN DE VILLE, D. (2015). Improved statistical evaluation of group differences in connectomes by screening–filtering strategy with application to study maturation of brain connections between childhood and adolescence. *NeuroImage* **108** 251–264.

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

NOWICKI, K. and SNIJDERS, T. A. B. (2001). Estimation and prediction for stochastic block structures. *Journal of the American Statistical Association* **96** 1077–1087.

RABUSSEAU, G. and KADRI, H. (2016). Low-rank regression with tensor responses. In *Advances in Neural Information Processing Systems* 1867–1875.

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

RELIÓN, J. D. A., KESSLER, D., LEVINA, E., TAYLOR, S. F. et al. (2019). Network classification with applications to brain connectomics. *The Annals of Applied Statistics* **13** 1648–1677.

ROUSSEAU, J. and MENGERSEN, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73** 689–710.

SAAD, Z. S., GOTTS, S. J., MURPHY, K., CHEN, G., JO, H. J., MARTIN, A. and COX, R. W. (2012). Trouble at rest: how correlation patterns and group differences become distorted after global signal regression. *Brain connectivity* **2** 25–32.

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statistica sinica* 639–650.

STUSS, D., ELY, P., HUGENHOLTZ, H., RICHARD, M., LAROCHELLE, S., POIRIER, C. and BELL, I. (1985). Subtle neuropsychological deficits in patients with good recovery after closed head injury. *Neurosurgery* **17** 41–47.

SUN, W. W. and LI, L. (2017). STORE: sparse tensor response regression and neuroimaging analysis. *The Journal of Machine Learning Research* **18** 4908–4944.

WANG, L., DURANTE, D., JUNG, R. E. and DUNSON, D. B. (2017). Bayesian network–response regression. *Bioinformatics* **33** 1859–1866.

WU, T., BENSON, A. R. and GLEICH, D. F. (2016). General tensor spectral co-clustering for higher-order data. In *Advances in Neural Information Processing Systems* 2559–2567.

YOUNG, S. J. and SCHEINERMAN, E. R. (2007). Random dot product graph models for social networks. In *International Workshop on Algorithms and Models for the Web-Graph* 138–149. Springer.