Covariate-Dependent Clustering of Undirected Networks with Brain-Imaging Data

Sharmistha Guha

Assistant Professor, Department of Statistics, Texas A & M University,

3143 TAMU, College Station, TX 77843-3143, E-mail: sharmistha@tamu.edu

Rajarshi Guhaniyogi

Associate Professor, Department of Statistics, Texas A & M University,

3143 TAMU, College Station, TX 77843-3143 , E-mail: rajguhaniyogi@tamu.edu

May 27, 2024

Abstract

This article focuses on model-based clustering of subjects based on the shared relationships of subject-specific networks and covariates in scenarios when there are differences in the relationship between networks and covariates for different groups of subjects. It is also of interest to identify the network nodes significantly associated with each covariate in each cluster of subjects. To address these methodological questions, we propose a novel nonparametric 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 involve low-rankness and group sparsity within the low-rank structure. While the low-rank structure in the network coefficients adds parsimony and computational efficiency, the group sparsity within the low-rank structure enables drawing inference on network nodes and cells significantly associated with each scalar predictor. Being a principled Bayesian mixture modeling framework, our approach allows model-based identification

of the number of clusters, offers clustering uncertainty in terms of the co-clustering matrix and presents precise characterization of uncertainty in identifying network nodes significantly related to a predictor in each cluster. Empirical results in various simulation scenarios illustrate substantial inferential gains of the proposed framework in comparison with competitors. Analysis of a real brain connectome dataset using the proposed method provides interesting insights into the brain regions of interest (ROIs) significantly related to creative achievement in each cluster of subjects. Supplementary material shows the convergence rate for the posterior predictive density of the proposed model, additional simulation examples with model mis-specification, full conditional distributions to run the Markov Chain Monte Carlo (MCMC) algorithm and also presents traceplots for various model parameters to demonstrate convergence of the MCMC algorithm.

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

1 Introduction

In recent times, network data is regularly encountered in disciplines as diverse as neuroscience, genetics, finance and economics. This article is motivated by scenarios where an undirected network and several scalar variables are available for each subject in a multi-subject setting, and there are differences in the relationship between the network and covariates for different groups of subjects. In this context, we develop a novel non-parametric Bayesian regression framework with the network serving as the response, and scalar predictors. In particular, we focus on brain connectome data obtained using diffusion weighted magnetic resonance imaging (dMRI). Here, the human brain is segmented into structural regions of interest (ROIs), viewed as *nodes* of the brain network, and the number of fiber bundles connecting any pair of regions (an *edge* connecting a pair of nodes) is estimated. Thus the brain may be viewed as an undirected network expressed in the form of a symmetric matrix. Along with the brain networks, we have information on a brain related phenotype (creative achievement), as well as behavioral variables, for each subject, which serve as our 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. The clustering endeavor aims at model-based determination of the number and composition of clusters. Additionally, it offers clustering uncertainty in terms of the model-based estimation of the probability of a pair of observations being part of the same cluster. Second, inferential interest lies in identifying nodes and edges in the network significantly impacted by each predictor of interest in each cluster. To this end, an important inferential objective is to accurately characterize uncertainty in our inference on network nodes. For the brain connectome data, the latter objective amounts to drawing inference on brain regions of interest (ROIs) and their interconnections significantly associated with creative achievement in each cluster. We also offer uncertainty in terms of estimated probability of a node being identified as influential. Moreover, we would like to achieve these inferential goals with parsimony in the fitted model and computational efficiency in the model fitting process.

An overwhelming literature in network analysis aims at understanding the topological structure of a single network, rather than focusing on multiple network observations collected for different individuals. Some notable examples of single-network models include exponential random graph models (Frank and Strauss, 1986), social space models (Hoff, 2005) 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, a 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 approach cannot identify the impact of specific nodes on the predictor, which is of clear interest in our setting. In a newer work, Wang *et al.*, 2017 exploit the relational nature of the network response, though their approach does not offer any clustering mechanism for the subjects and is not designed to detect network nodes significantly related to a scalar predictor. Another approach in the tensor regression literature, which assumes a general version of the tensor and does not consider the symmetry constraint (inherent in an undirected network), applies

regression with a matrix/tensor response (Gahrooei *et al.*, 2021; Guhaniyogi and Spencer, 2021). Some recent articles on supervised stochastic block models (Kim and Levina, 2019; Pavlović *et al.*, 2020) focus on clustering the nodes of the network into groups, which is methodologically a different problem than ours, where the focus is on clustering subjects into groups.

While our framework treats the network as a response, a few recent approaches (Guha and Rodriguez, 2021; Relión *et al.*, 2019) treat the network as a predictor to predict a scalar response. This difference in the modeling approach leads to a different focus and interpretation. Network predictor regression focuses on understanding variations in a biological outcome as the network image varies, while network response regression aims to study changes in the network as predictors such as the creativity levels, age and sex vary. In a sense, their difference is comparable to that of multi-response regression and multi-predictor regression in the classical vector-valued regression context. Importantly, our network-response regression framework bypasses the need to invert any high dimensional matrix to draw Bayesian inference, thereby adding substantial computational gains over Guha and Rodriguez, 2021, especially in the analysis of networks with a moderately large to a large number of nodes, when computation in Guha and Rodriguez, 2021 may become quite expensive. Note that a recent approach (Guha and Guhaniyogi, 2021) discusses symmetric tensor-on-vector regression models; however, both Guha and Guhaniyogi, 2021 and Guha and Rodriguez, 2021 tacitly assume that the same set of network nodes influence the regression function in a similar manner for every subject.

The neuroscience literature provides substantial evidence of differences in the relationship between brain connectivity networks and phenotypic traits for different groups of individuals (Saad *et al.*, 2012; Meskaldji *et al.*, 2013, 2015). However, flexible statistical methods for identifying such subgroups and ascertaining subgroup differences have somewhat lagged behind the increasingly routine collection of such data. One possibility is to flatten the network to a high dimensional multivariate vector and employ a mixture of multivariate regression models. This idea can make use of the literature on mixtures of supervised parametric and semi-parametric linear and generalized linear models with continuous, binary and categorical responses and predictors (Müller *et al.*, 1996; Shahbaba and Neal, 2009; Duan *et al.*,

2007; Rodríguez *et al.*, 2009; Amewou-Atisso *et al.*, 2003; Hannah *et al.*, 2011). These approaches are less suitable for our problem of interest since they ignore the network topology in the model building process, and do not allow drawing inference on network nodes. In this context, one may invoke the literature on clustering of matrices or tensors into multiple groups (Huang *et al.*, 2009; Lee *et al.*, 2010; Chi and Lange, 2015; Chi *et al.*, 2017; Li *et al.*, 2014; Cao *et al.*, 2013; Wu *et al.*, 2016), but this literature is more pertinent to unsupervised clustering of networks, as opposed to our interest in supervised clustering of networks.

In this article, we propose a novel nonparametric Bayesian modeling approach to achieve the aforementioned inferential objectives simultaneously. A Dirichlet process (DP) mixture of network response regression models is developed and used to analyze 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 assumed to have a low-rank for parsimony and computational efficiency. We also impose a node-specific sparsity structure using a Bayesian spike-and-slab variable selection prior for identifying network nodes significantly associated with the predictors. Being a Bayesian mixture modeling framework, our method achieves model-based determination of the number of clusters and delivers model-based estimation of clustering uncertainty. Specifically, clustering uncertainty is quantified by estimating the posterior probability of any pair of observations being part of the same cluster. Additionally, the proposed method identifies network nodes significantly related to a predictor in each cluster, and characterizes uncertainty associated with identifying important network nodes in each cluster. The uncertainty is quantified by model-based estimation of posterior probability of a node being influentially related to a predictor. Unlike a frequentist approach, both the uncertainty in clustering as well as the uncertainty in the inference on nodes can be seamlessly estimated from the post burn-in Markov Chain Monte Carlo (MCMC) iterates of the proposed Bayesian approach. Such uncertainty characterizations are crucial in neuroscientific applications with a limited sample size, high dimensional parameters and low signal-to-noise ratio. Our framework does not involve any expensive matrix manipulation, allowing efficient computation with a large number of network nodes.

Our approach is fundamentally different from regression with a covariance matrix out-

come and scalar predictors. These approaches tacitly use positive-definite property of the outcome matrix (Hoff and Niu, 2012; Zou *et al.*, 2017). A similar modeling framework is developed by Zhao *et al.* (2021), wherein important principal components of covariance matrix outcomes related to scalar covariates are identified. In contrast, we do not impose any restrictions on our network matrix response matrix (e.g., positive definiteness) except symmetry, which is directly motivated by the scientific problem at hand. Additionally, an important modeling objective lies in identifying network nodes (with uncertainties) significantly related to scalar predictors. The aforementioned approaches are not designed to deliver such inference.

The rest of this article progresses as follows. Sections 2 and 3 describe model development and posterior computation, respectively. Empirical investigation of the model with simulation studies and the brain connectome data analysis are presented in Sections 4 and 5, respectively. Finally, Section 6 concludes the paper with an eye towards future work. Theoretical investigations of the model, showing convergence rate of the posterior predictive distribution, are described in the supplementary material. Supplementary material also shows simulation studies under model mis-specification, full conditional posterior distributions for running MCMC and convergence of MCMC chains for model estimation by displaying traceplots of model parameters.

2 Supervised Clustering of Undirected Networks: Model and Prior Formulation

2.1 Notations and Framework

For $i = 1, \ldots, n$, let $\boldsymbol{Y}_i \in \mathcal{Y} \in \mathbb{R}^{p \times p}$ denote the weighted undirected network response with p nodes, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{im})'$ be m predictors of interest and $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{il})'$ 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 article, 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 ($y_{i,(j_1,j_2)} = y_{i,(j_2,j_1)}$).

We assume that the relationship between the predictor vector of interest $\boldsymbol{x}_i$ and the

response varies in every cell (j_1, j_2). In contrast, an auxiliary predictor explains the response in every cell identically. Let $\mathcal{J} = \{\boldsymbol{j} = (j_1, j_2) : 1 \leq j_1 < j_2 \leq p\}$ be a set of indices. Since $\boldsymbol{Y}_i$ is symmetric with 0 diagonal entries, it suffices to build a probabilistic generative mechanism for $y_{i,\boldsymbol{j}}(\boldsymbol{j} \in \mathcal{J})$. This is a common practice in the modeling of undirected relational data (Hoff *et al.*, 2002; Hoff, 2005). Moreover, working with $(y_{i,\boldsymbol{j}} : \boldsymbol{j} \in \mathcal{J})$ is fundamentally different from the ordinary vectorization of $\boldsymbol{Y}_i$ for model fitting, 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 (i.e., position of the entry in the matrix), which will be crucial in the modeling development described below.

2.2 Model Development and Prior Distributions

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,

$$f(\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i, \sigma^2) = \int g(\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{B}_1, \ldots, \boldsymbol{B}_m, \gamma_0, \ldots, \gamma_l, \sigma^2) dG(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_m, \gamma_0, \gamma_1, \ldots, \gamma_l), \tag{1}$$

where $g(\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{B}_1, \ldots, \boldsymbol{B}_m, \gamma_0, \ldots, \gamma_l, \sigma^2)$ is suitably chosen from the class of exponential family distributions depending on whether each $y_{i,\boldsymbol{j}}$ is continuous, binary or categorical. In this article, we focus on the case where each network edge $y_{i,\boldsymbol{j}}$ is continuous, and choose $g(\cdot)$ such that each $y_{i,\boldsymbol{j}}(\boldsymbol{j} \in \mathcal{J})$ is normally distributed with its probabilistic generative mechanism given by

$$y_{i,\boldsymbol{j}} = \gamma_0 + \sum_{s=1}^{l} \gamma_s z_{is} + \sum_{s=1}^{m} B_{s,\boldsymbol{j}} x_{is} + e_{i,\boldsymbol{j}}, \;\; e_{i,\boldsymbol{j}} \sim N(0, \sigma^2). \tag{2}$$

Supplementary material presents performance of (2) when it is mis-specified, e.g., the continuous edge effects $y_{i,\boldsymbol{j}}$'s are generated from a heavy-tailed t-distribution in the truth. The coefficients $B_{1,\boldsymbol{j}}, \ldots, B_{m,\boldsymbol{j}}$ are the $\boldsymbol{j}$-th cells of the symmetric matrix coefficients $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_m$ of order $p \times p$ with zero diagonal entries, respectively. Here γ_0 is the intercept and $\gamma_1, ..., \gamma_l \in \mathbb{R}$ are coefficients corresponding to the auxiliary predictors. The model formulation implies a similar effect of any of the auxiliary variables $z_{i1}, ..., z_{il}$ on all cells of the response tensor in

each mixture component. In contrast, $B_{s,\boldsymbol{j}}$ determines the effect of x_{is} on the $\boldsymbol{j} = (j_1, j_2)$-th cell of the response in any mixture component. Equations (1) and (2) together can be seen as a mixture of undirected network response regression models with the mixing distribution given by $G(\cdot)$. Note that (1) is markedly different from building an ordinary mixture of linear regression models with vectorized network response $\boldsymbol{Y}_i$ and scalar predictors. Vectorization of the $\boldsymbol{Y}_i$ network matrix would have led to a long vector, which would be a collection of network edges. But such a construction would lose the crucial information regarding the pair of nodes that each edge is connected to. Since our interest lies in drawing inference on network nodes significantly associated with a scalar predictor, such a vectorization strategy is inadequate. In contrast, (2) develops a regression model for each edge on scalar predictors while keeping a tab on the nodes, j_1 and j_2, that the edge $y_{i,\boldsymbol{j}}$ is connected to. The edge coefficient $B_{s,\boldsymbol{j}}$ in (2) is suitably modeled, exploiting this information, to draw inference on network nodes significantly related to the predictors. It is important to note that the prior literature on network data modeling and tensor regression models also use a similar strategy (Hoff, 2005; Guhaniyogi *et al.*, 2017; Guha and Guhaniyogi, 2021). We further elaborate on this point as this section progresses.

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 $\boldsymbol{\Delta}_h^* = (\boldsymbol{B}_{1,h}^*, .., \boldsymbol{B}_{m,h}^*, \gamma_{0,h}^*, \gamma_{1,h}^*, .., \gamma_{l,h}^*) \sim G_0$. 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 $\alpha > 0$,

$$\omega_1 = v_1^*,\ \omega_2 = v_2^*(1 - v_1^*), .., \omega_{H-1} = v_{H-1}^* \prod_{h=1}^{H-2}(1 - v_h^*),\ \omega_H = \prod_{h=1}^{H-1}(1 - v_h^*),\ v_h^* \sim Beta(1, \alpha), \tag{3}$$

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 it 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

$$f(\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i, \sigma^2) = \sum_{h=1}^{H} \omega_h g(\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{B}^*_{1,h}, \ldots, \boldsymbol{B}^*_{m,h}, \gamma^*_{0,h}, \ldots, \gamma^*_{l,h}, \sigma^2). \tag{4}$$

The mixture components signify different relationships between the network response and scalar predictors in H different clusters. We introduce a cluster index $c_i \in \{1, .., H\}$ corresponding to the individual i, $\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i, c_i, \sigma^2 \sim g(\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{B}^*_{1,c_i}, \ldots, \boldsymbol{B}^*_{m,c_i}, \gamma^*_{0,c_i}, \ldots, \gamma^*_{l,c_i}, \sigma^2)$, 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 determine the number of clusters and constitution of each cluster.

In order to identify network nodes in different clusters significantly associated with predictors of interest, we first introduce a low-rank structure of the coefficient $\boldsymbol{B}^*_{s,h} = ((B^*_{s,h,\boldsymbol{j}}))^p_{j_1,j_2=1}$ corresponding to the sth predictor of interest in the hth cluster as

$$B^*_{s,h,\boldsymbol{j}} = \sum_{r=1}^{R} \lambda_{s,h,r} u^{(r)}_{s,h,j_1} u^{(r)}_{s,h,j_2}, \;\; h = 1, ..., H; \;\; s = 1, .., m, \;\; 1 \le j_1 < j_2 \le p. \tag{5}$$

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, and $\lambda_{s,h,r} \in \{-1, 0, 1\}$ determines if the rth summand in (5) is relevant in model fitting in the hth mixture component. Setting $\boldsymbol{U}_{s,h}$ as a $p \times R$ matrix with the k-th row as $\boldsymbol{u}_{s,h,k}$ $(k = 1, ..., p)$, and $\boldsymbol{\Lambda}_{s,h}$ a $R \times R$ diagonal matrix with the r-th diagonal entry as $\lambda_{s,h,r}$, (5) represents a low-rank decomposition of the symmetric matrix coefficient $\boldsymbol{B}^*_{s,h} = \boldsymbol{U}_{s,h}\boldsymbol{\Lambda}_{s,h}\boldsymbol{U}'_{s,h}$. Since the choice of R is arbitrary, allowing $\lambda_{s,h,r}$ to be 0 protects the model from over-fitting. The low-rank formulation of $\boldsymbol{B}^*_{s,h}$ is motivated by several considerations simultaneously. In practice, the matrix of coefficients $\boldsymbol{B}^*_{s,h}$ is expected to exhibit transitivity effects, i.e., we expect that if the interactions between nodes j_1 and j_2 and between nodes j_2 and j_3 are both influentially related to the sth predictor of interest, the interaction between nodes j_1 and j_3 is likely to

be influential as well (e.g., see Li *et al.*, 2013). Such an effect is a natural outcome of the low-rank formulation of $\boldsymbol{B}^*_{s,h}$. Second, the low-rank formulation allows inference on network nodes through the node-specific latent vectors $\boldsymbol{u}_{s,h,1}, \ldots, \boldsymbol{u}_{s,h,p}$, which can be interpreted 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. 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 $mHp(p-1)/2$ to $mHRp$, typically with $R \ll p$.

Depending on the structure of $\boldsymbol{\Lambda}_{s,h}$, the node specific latent variables $\boldsymbol{u}_{s,h,k}$'s may become unidentifiable. For example, when $\boldsymbol{\Lambda}_{s,h} = \boldsymbol{I}_R$, $\boldsymbol{B}^*_{s,h} = \boldsymbol{U}_{s,h}\boldsymbol{\Lambda}_{s,h}\boldsymbol{U}'_{s,h} = \boldsymbol{U}_{s,h}\boldsymbol{O}\boldsymbol{\Lambda}_{s,h}(\boldsymbol{U}_{s,h}\boldsymbol{O})'$, for any orthogonal matrix $\boldsymbol{O}$. While this implies that posterior inference on $\boldsymbol{u}_{s,h,k}$'s (without any constraint imposed on $\boldsymbol{u}_{s,h,k}$'s) may not always be meaningful, our focus is on the event $\{\boldsymbol{u}_{s,h,k} = \boldsymbol{0}\}$ for each k, which is indeed identifiable (since $\boldsymbol{0}$-valued latent vectors are invariant to orthogonal transformation) and is critical to drawing inference on the nodes related to the s-th predictor of interest, as we describe next. 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,

$$\boldsymbol{u}_{s,h,k} \sim \begin{cases} N(\boldsymbol{0}, \boldsymbol{M}_{s,h}), & \text{if } \xi_{s,h,k} = 1 \\ \delta_{\boldsymbol{0}}, & \text{if } \xi_{s,h,k} = 0 \end{cases}, \ \xi_{s,h,k} \sim Ber(\zeta_{s,h}), \ \boldsymbol{M}_{s,h} \sim IW(\nu, \boldsymbol{I}), \ \zeta_{s,h} \sim Beta(a, b). \tag{6}$$

where $\boldsymbol{M}_{s,h}$ is a covariance matrix of order $R \times R$. The parameter $\zeta_{s,h}$ corresponds to the probability of the nonzero mixture component in (6). 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. In order to learn how many summands in (5) are informative, we draw inference on $R_{eff,s} = \sum_{r=1}^{R} |\lambda_{s,h,r}|$ by assigning a hierarchical prior

$$\lambda_{s,h,r} \sim \begin{cases} 0, & \text{w.p. } \pi_{s,h,r,1}, \\ 1, & \text{w.p. } \pi_{s,h,r,2}, \\ -1, & \text{w.p. } \pi_{s,h,r,3}, \end{cases} \qquad (\pi_{s,h,r,1}, \pi_{s,h,r,2}, \pi_{s,h,r,3}) \sim Dirichlet(r^\eta, 1, 1), \quad \eta > 1.$$

The choice of hyper-parameters of the Dirichlet distribution is crucial. In particular, note that (i) $E[|\lambda_{s,h,r}|] = 2/(2+r^{\eta}) \to 0$ as $r \to \infty$ and that (ii) $\sum_{r=1}^{R} Var(|\lambda_{s,h,r}|) = \sum_{r=1}^{R}[\frac{2(r^{\eta}+1)}{(r^{\eta}+2)^2(r^{\eta}+3)} + \frac{2(r^{\eta}+1)}{(r^{\eta}+3)(r^{\eta}+4)}] < \infty$ as $R \to \infty$. Property (i) provides (weak) identifiability of the different latent dimensions, while property (ii) ensures that $\lim_{R\to\infty} var(R_{eff,s}) < \infty$. The parameters $\gamma^*_{0,h}, \gamma^*_{1,h}, ..., \gamma^*_{l,h}$ are assigned standard normal distributions and the error variance σ^2 is assigned IG(a_σ, b_σ) a-priori. With the construction specified above, the form of the base measure G_0 can be expressed as $G_0(\boldsymbol{\Delta}^*_h|\sigma^2) = \prod_{s=0}^{l} G_{0,1}(\gamma^*_{s,h}|\sigma^2) \prod_{s=1}^{m} G_{0,2}(\boldsymbol{B}^*_{s,h}|\sigma^2)$, where $G_{0,1}(\gamma^*_{s,h}|\sigma^2) = N(0,1)$, and $G_{0,2}(\boldsymbol{B}^*_{s,h}|\sigma^2)$ is expressed as follows:

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

The model and prior specifications allow clustering of individuals into a number of groups less than or equal to H. In each group, the network response and 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.

3 Posterior Computation

While fitting our proposed mixture model, we adopt a moderately large choice of H following the strategy in Rousseau and Mengersen (2011). If a specific choice of H leads to all of the clusters being occupied, then the analysis should be repeated for a larger H until there is at least one unoccupied cluster. Since all parameters except α have full conditional posterior distributions belonging to standard families, Gibbs sampling with Metropolis is implemented to empirically estimate posterior distributions. Details of the Markov chain Monte Carlo algorithm are presented in the supplementary material. 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.

Indicators to assess clustering performance. To assess inference from the proposed mixture model, we investigate (i) a point estimate of clustering denoted by $\hat{\boldsymbol{c}}$, (ii) a heatmap of the posterior probabilities of pairs of samples belonging to the same cluster, i.e., $P(c_i =$

$c_j|\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ (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 component-wise optimization) the expected loss function discussed in Lau and Green, 2007,

$$F(\hat{\boldsymbol{c}}) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} 1(\hat{c}_i = \hat{c}_j) \left[\frac{o_2}{o_1 + o_2} - P(c_i = c_j | \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n) \right], \tag{7}$$

where the ratio o_1/o_2 controls the relative loss due to incorrect clustering or separation of a pair of samples. Without any prior knowledge, we assume the loss due to incorrectly clustering and separating any pair of samples is the same, by setting $o_1/o_2 = 1$.

4 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 network response $\boldsymbol{Y}_i$, depending on the predictors $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, from the finite mixture model given by

$$\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i \sim \sum_{h=1}^{H_0} \omega_{h,0} g(\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i, B^*_{1,h,0}, ..., B^*_{m,h,0}, \gamma^*_{0,h,0}, ..., \gamma^*_{l,h,0}, \sigma^2_0), \;\; i = 1, ..., n + n_0. \tag{8}$$

The first n observations $\{(\boldsymbol{Y}_i, \boldsymbol{x}_i, \boldsymbol{z}_i) : i = 1, ..., n\}$ are used for model fitting and the rest n_0 observations are used for out-of-sample inference. Here $g(\cdot)$ is as described in equation (2), and $\boldsymbol{B}^*_{s,h,0}$, $s = 1, \ldots, m$, $h = 1, \ldots, H_0$ are mixture specific coefficients for x_{is}. Section 3 of the supplementary material shows performance of the proposed approach and its competitors under model mis-specification, i.e., when each $y_{i,\boldsymbol{j}}$ is generated from a heavy-tailed t-distribution. 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 $\boldsymbol{B}^*_{h,0}$, we draw p latent variables $\boldsymbol{u}_{h,k,0}$, each of dimension R_g,

from a mixture distribution given by

$$\boldsymbol{u}_{h,k,0} \sim \pi_0 N_{R_g}(\boldsymbol{u}_{h,m,g}, u^2_{h,v,g}\boldsymbol{I}_{R_g}) + (1-\pi_0)\delta_{\mathbf{0}};\ k \in \{1,...,p\}, \tag{9}$$

where $N_{R_g}(\boldsymbol{u}_{h,m,g}, u^2_{h,v,g}\boldsymbol{I}_{R_g})$ represents an R_g-variate normal distribution with mean vector $\boldsymbol{u}_{h,m,g}$ and covariance matrix $u^2_{h,v,g}\boldsymbol{I}_{R_g}$. Also, $(1-\pi_0)$ 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 10 simulation cases as following:

Cases 1-8: In Cases 1-8, 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 diagonal entries of $\boldsymbol{B}^*_{h,0}$ are set to zero, and the lower triangular part of $\boldsymbol{B}^*_{h,0}$ is a mirror image of the upper triangular part. The 8 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 9: In Case 9, we consider $H_0 = 2$, $\omega_{1,0} = 0.4, \omega_{3,0} = 0.6$, and $\boldsymbol{B}^*_{1,0}$ and $\boldsymbol{B}^*_{2,0}$ are simulated using two different strategies as following:

Simulating $\boldsymbol{B}^*_{1,0}$*:* 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. The diagonals of $\boldsymbol{B}^*_{1,0}$ are set to zero and the lower triangular part is a mirror image of the upper triangular part.

Simulating $\boldsymbol{B}^*_{2,0}$*:* Randomly set $(1-\pi_0) = 0.6$ proportion of upper triangular elements in $\boldsymbol{B}^*_{2,0}$ to zero, while the rest are simulated from $N(0,1)$. The diagonals of $\boldsymbol{B}^*_{2,0}$ are set to zero and the lower triangular part is a mirror image of the upper triangular part.

Case 10: Case 10 uses an identical construct as Case 9, 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

Table 1: Table presents specifications of Cases 1-8 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 present various combinations of the number of network nodes p, fitted sample size n, number of subjects n_0 for out-of-sample inference, network node sparsity $(1-\pi_0)$, true (R_g) and fitted (R) dimensions of the node specific latent variables.

Cases	p	n	n_0	R_g	R	$(1-\pi_0)$	H_0
1	30	100	20	2	5	0.6	3
2	30	100	20	2	5	0.3	3
3	30	100	20	3	5	0.6	4
4	80	100	20	2	5	0.6	3
5	80	100	20	2	5	0.3	3
6	80	100	20	3	5	0.6	2
7	30	100	20	2	5	0.6	1
8	30	100	20	2	5	0.3	1

and 2.

Notably, **Cases 1-6** represent scenarios with heterogeneous settings where $H_0 > 1$, i.e., the true data generating model is a mixture of more than one network response regression model. In contrast, Cases 7 and 8 correspond to the homogeneous setting with $H_0 = 1$, i.e., the true model constitutes a single network response regression model. Also, it is important to note that **Cases 1-8** represent the true model being included in the class of fitted models. In contrast, **Cases 9** and **10** show departure of the true model from the fitted models. In particular, the last two cases include specifications where the network coefficient in a cluster is full rank, whereas the fitted model assumes a low-rank structure for network coefficients in all the clusters. This will allow assessing the performance of our approach under model mis-specification. Notably, for any p (i.e., the number of network nodes), the model needs to estimate $mHRp$ parameters, so that even moderately large values of p in Cases 1-10 lead to high dimensional regression settings. In all simulations, we set $p < n$, which is a sufficient condition for convergence of the fitted predictive density to the true predictive density (see Theorem 1.1 in the Supplementary Material).

4.1 Choice of Hyper-parameters

All simulation settings 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 represent a reasonable prior belief. Moderately perturbing hyper-parameters yields practically identical inference, as described in Section 4.5.

4.2 Competitors and Metrics of Evaluation

NRMM is fitted in all simulations with $H = 15$ mixture components and it is allowed to identify the unknown and true number of mixtures from the data. 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. However, we also include simulation cases 7 and 8 where data are simulated from the NRR model (i.e., there is only one mixture component), and hence data are likely to favor NRR over NRMM. Additionally, we compare our approach with a frequentist higher order low-rank regression (HOLRR) method (Rabusseau and Kadri, 2016) popularly used in machine learning. This approach also assumes a homogeneous setting and a low-rank decomposition for matrix coefficients without the ability of identifying influential network nodes with uncertainties.

The competitors are assessed based on their ability to estimate the true regression mean function $E_0[y_{i,\boldsymbol{j}}|\boldsymbol{x}_i, \boldsymbol{z}_i] = \sum_{h=1}^{H_0} \omega_{h,0} \left(\gamma^*_{0,h,0} + \sum_{s=1}^{l} \gamma^*_{s,h,0} z_{is} + \sum_{s=1}^{m} B^*_{s,h,0,\boldsymbol{j}} x_{is}\right)$ for both in-sample and out-of-sample observations. In particular, the mean squared error (MSE) of estimating the true regression mean function over the in-sample fitted data and out-of-sample

observations are given by,

$$\text{MSE} = \frac{2}{np(p-1)} \sum_{i=1}^{n} ||E_0[\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i] - E[\widehat{\boldsymbol{Y}_i|\boldsymbol{x}_i}, \boldsymbol{z}_i]||^2,$$
$$\text{MSE(out)} = \frac{2}{n_0 p(p-1)} \sum_{i=n+1}^{n+n_0} ||E_0[\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i] - E[\widehat{\boldsymbol{Y}_i|\boldsymbol{x}_i}, \boldsymbol{z}_i]||^2, \tag{10}$$

where $E[\widehat{\boldsymbol{Y}_i|\boldsymbol{x}_i}, \boldsymbol{z}_i]$ denotes the posterior mean of the regression function for a competing method. While MSE offers an evaluation of the point estimation by competitors in-sample, MSE(out) evaluates the point estimation by competitors out-of-sample. The uncertainty in estimating the true regression mean function in-sample is measured using the coverage and length of 95% credible intervals obtained from NRMM and NRR. HOLRR being a frequentist method, coverage and length of 95% confidence intervals from HOLRR are not readily available. Thus, we invoke normal approximation wherein the 95% confidence interval for $E_0[y_{i,\boldsymbol{j}}|\boldsymbol{x}_i, \boldsymbol{z}_i]$ is constructed by $(E[\widehat{y_{i,\boldsymbol{j}}|\boldsymbol{x}_i}, \boldsymbol{z}_i] - 1.96\ se(E[y_{i,\boldsymbol{j}}|\boldsymbol{x}_i, \boldsymbol{z}_i]), E[\widehat{y_{i,\boldsymbol{j}}|\boldsymbol{x}_i}, \boldsymbol{z}_i] + 1.96\ se(E[y_{i,\boldsymbol{j}}|\boldsymbol{x}_i, \boldsymbol{z}_i])$, where $se(E[y_{i,\boldsymbol{j}}|\boldsymbol{x}_i, \boldsymbol{z}_i])$ denotes the standard error of estimating the regression mean function in-sample.

In addition to reporting the posterior distribution of the number of clusters and the uncertainty associated with clustering through $P(c_i = c_j|\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$, 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. ARI is upper bounded by 1, with values close to 1 indicating more agreement between cluster configurations.

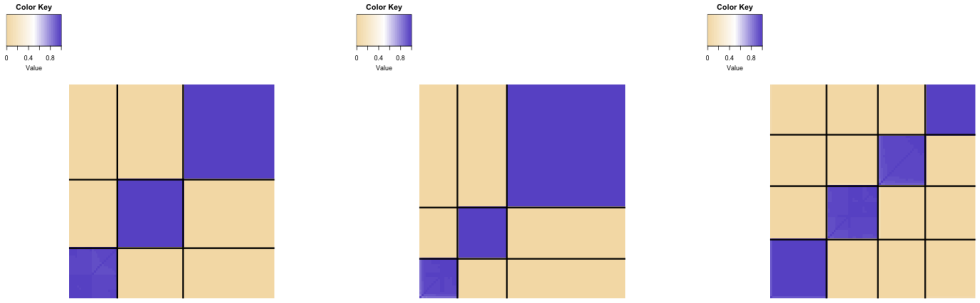
4.3 Simulation Results

All model parameters show excellent convergence with fairly uncorrelated post burn-in samples to draw posterior inference. To demonstrate this, we present the effective sample size (ESS) corresponding to 10000 post burn-in samples from NRMM for all simulation examples (see Table 2). Section 4 of the supplementary material shows traceplots for a few parameters, and demonstrate excellent convergence. Table 2 and Figure 1 provide insights

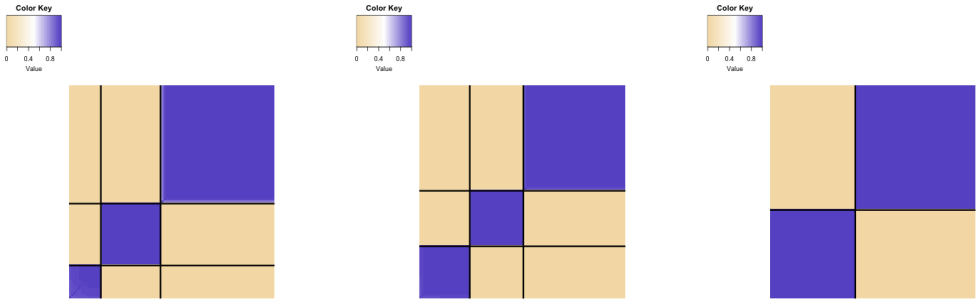
into the estimates of the cluster structure and associated uncertainty by displaying the discrepancy between the true and estimated number of clusters and heat maps of posterior probabilities of pairs of subjects belonging to the same cluster. To facilitate visualization in Figure 1, subjects are ordered according to their true cluster configurations in the heatmap. In all cases, the model successfully recovers the true cluster structure, with little uncertainty associated with the estimator. Importantly, the mixture model also successfully identifies only one cluster when the true data generation setting is homogeneous (cases 7-8). The most challenging cases among all are cases 9 and 10, 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 9 and 0.95 in case 10. It appears from Figure 1 that the clustering performance improves nominally with decreasing sparsity of $\boldsymbol{B}^*_{h,0}$, the impact of sparsity being a little more prominent under model mis-specification (compare cases 9 and 10). The uncertainty in clustering for a few individuals also appears to be higher in case 8, where the true data generating model is homogeneous with $H_0 = 1$.

The posterior distributions of the number of identified clusters are also presented in the form of barplots in Figure 2. The posterior distribution of the number of clusters appears to concentrate around the true number of clusters H_0 in all cases except case 9, where the model mildly overestimates the number of clusters. Notably, case 9 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 10). 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 8 (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.

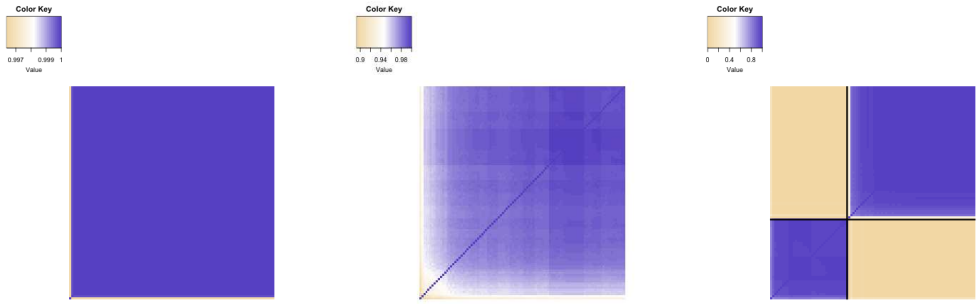
Table 2 presents in-sample MSE and out-of-sample MSE(out) for estimating the regression mean function for all competitors. Further, coverage and average length of 95% credible intervals for NRMM and NRR, and 95% confidence intervals for HOLRR are provided to assess uncertainty quantification for regression mean functions in-sample. 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



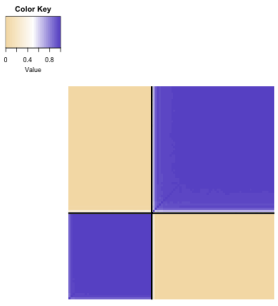
(a) Case 1,$H_0 = 3$ (b) Case 2,$H_0 = 3$ (c) Case 3,$H_0 = 4$



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

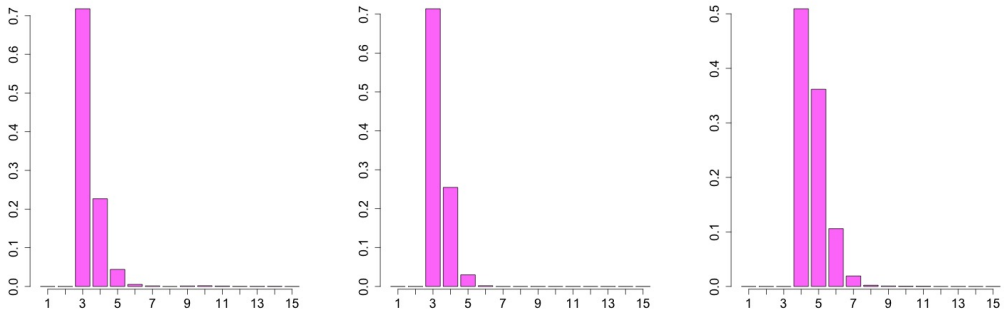


(g) Case 7,$H_0 = 1$ (h) Case 8,$H_0 = 1$ (i) Case 9,$H_0 = 2$

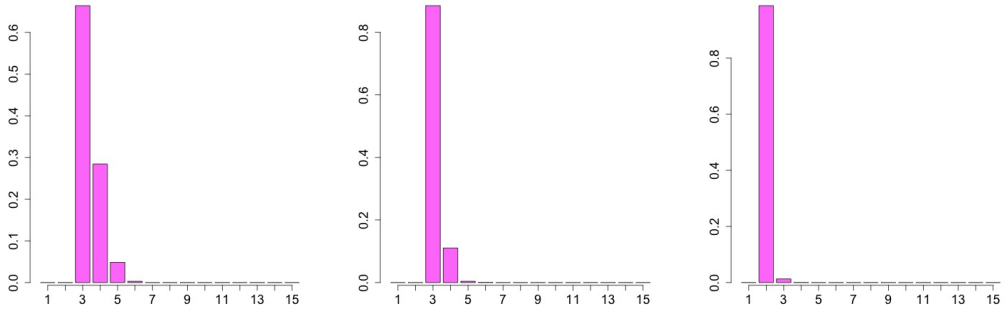


(j) Case 10,$H_0 = 2$

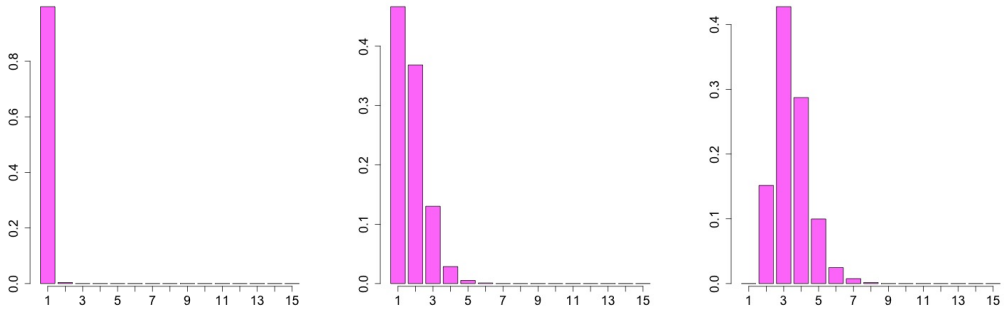
Figure 1: Plots showing uncertainty in estimating clusters in simulation cases 1-10. Boldfaced horizontal and vertical lines indicate the true clustering.



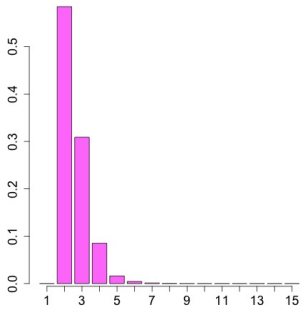
(a) Case 1,$H_0 = 3$ (b) Case 2,$H_0 = 3$ (c) Case 3,$H_0 = 4$



(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,$H_0 = 1$ (i) Case 9,$H_0 = 2$



(j) Case 10,$H_0 = 2$

Figure 2: Plots showing the posterior distribution of the number of clusters in the simulation cases 1-10.

Table 2: The first column presents Effective sample size (ESS) for NRMM corresponding to the 10000 post burn-in iterations to assess the convergence of the MCMC sampler for NRMM. 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. MSE and MSE(out) for NRMM, NRR and HOLRR are presented for cases 1-10. The lowest MSE in each case is boldfaced. Coverage and length of 95% credible interval are provided for NRMM, NRR and HOLRR.

	NRMM					Competitors		
Case	ESS	ARI	TPR	FPR		NRMM	NRR	HOLRR
1	8006	0.99	0.87	0.08	MSE(out)	**0.07**	0.47	0.10
					MSE	**0.02**	0.40	0.08
					Coverage of 95% CI	0.89	0.02	0.99
					Length of 95% CI	0.54	0.22	3.67
2	7985	0.99	0.90	0.05	MSE(out)	**0.12**	0.61	0.26
					MSE	**0.03**	0.54	0.14
					Coverage of 95% CI	0.96	0.05	0.98
					Length of 95% CI	0.58	0.44	4.12
3	7942	0.98	0.71	0.00	MSE(out)	**0.23**	0.42	0.51
					MSE	**0.14**	0.32	0.44
					Coverage of 95% CI	0.69	0.29	0.98
					Length of 95% CI	0.64	0.39	4.67
4	7235	0.99	0.95	0.02	MSE(out)	**0.05**	0.09	0.12
					MSE	**0.01**	0.07	0.09
					Coverage of 95% CI	0.99	0.15	0.99
					Length of 95% CI	0.47	0.15	3.64
5	7451	0.99	0.93	0.02	MSE(out)	**0.09**	0.16	0.18
					MSE	**0.04**	0.06	0.11
					Coverage of 95% CI	0.93	0.44	0.98
					Length of 95% CI	0.55	0.34	3.92
6	7324	0.99	1.00	0.00	MSE(out)	**0.09**	0.58	0.22
					MSE	**0.05**	0.30	0.17
					Coverage of 95% CI	0.99	0.10	0.99
					Length of 95% CI	0.61	0.28	4.05
7	8106	0.97	0.92	0.00	MSE(out)	0.17	**0.06**	0.33
					MSE	0.12	**0.008**	0.32
					Coverage of 95% CI	0.86	0.97	0.96
					Length of 95% CI	0.37	0.07	2.86
8	8390	1.00	0.89	0.02	MSE(out)	0.19	**0.07**	0.38
					MSE	0.14	**0.02**	0.36
					Coverage of 95% CI	0.84	0.95	0.96
					Length of 95% CI	0.40	0.06	2.74
9	8195	0.93	–	–	MSE(out)	**0.16**	1.33	0.19
					MSE	**0.10**	1.30	0.13
					Coverage of 95% CI	0.84	0.07	0.96
					Length of 95% CI	0.51	0.36	3.69
10	7839	0.95	–	–	MSE(out)	**0.20**	0.69	0.21
					MSE	**0.17**	0.54	0.19
					Coverage of 95% CI	0.74	0.09	0.97
					Length of 95% CI	0.70	0.39	3.81

the sparsity parameter $(1 - \pi_0)$. Results from cases 9 and 10 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. The out-of-sample inference on regression mean function follow very similar trends as suggested by the MSE(out) values. As expected, out-of-sample MSE values are always higher than in-sample MSE values. Additionally, in most cases, NRMM shows higher coverage levels, often close to nominal coverage, compared to NRR. The less than nominal coverage in cases 9 and 10 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 obstacles in model estimation. Note that under cases 7 and 8, only one mixture component (homogeneous setting) is used to simulate the data, and so the data favors NRR over NRMM. Consequently, NRR yields 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. The construction of 95% confidence intervals for HOLRR based on the asymptotic normal approximation result in much wider intervals with coverage close to 1.

Note that inference on each cluster is not readily available from the mixture model due to the clusters not being 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 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 1 and 2, for cases 1-8, 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

Table 3: Computation time (in seconds) per MCMC iteration of the NRMM model with $H = 15$ mixture components. The number of network nodes (p) and the sample size (n) are varied.

p	20	40	80	160	200	250
$n = 50$	0.17	0.32	1.08	3.63	5.97	7.63
$n = 100$	0.26	0.43	1.14	4.10	6.41	13.40
$n = 150$	0.40	0.72	1.70	6.08	9.49	16.31

Rates (TPR)= $\frac{\text{TP}}{\text{TP+FP}}$ and False Positive Rates (FPR)= $\frac{\text{FP}}{\text{TN+FP}}$ of identifying influential network nodes over all clusters, where TP, FP and TN denote the total number of true positives, false positives and true negatives, respectively. 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.

4.4 Computational Complexity and Time

In our framework, the Gibbs sampler for model estimation does not involve any expensive matrix inversion or multiplication, leading to fast computation. In fact, the Gibbs sampler can be suitably parallelized since the updates of $\boldsymbol{u}_{s,h,k}$ can be performed over different processors in parallel. Computation times (in seconds) per MCMC iteration for the NRMM model (for varying number of network nodes, p, and sample size n) without parallelization are provided in Table 3. The entries in the table are recorded corresponding to $H = 15$ mixture components fitted to the data.

4.5 Sensitivity Analysis

To check 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) $a = 1, b = 1, \nu = 50$. Note 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

Table 4: Sensitivity Analysis: ARI, MSE, MSE(out), coverage of 95% CI and length of 95% CI for NRMM in 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
MSE(out)	0.15	0.12	0.15
Coverage of 95% CI	0.93	0.96	0.95
Length of 95% CI	0.61	0.57	0.50

for $\xi_{h,k}$ which encourages a higher number of activated nodes. Combination (iii) presents a variation of the hyperparameter ν in the Inverse-Wishart distribution of $\boldsymbol{M}_h$. Table 4 shows the posterior mean of ARI in case 2 under the three different hyper-parameter settings. We additionally present MSE, MSE(out), 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 on the inference, 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)), both MSE and MSE(out) are 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 little higher as compared to case 2 in Table 2. In contrast, combinations (ii) and (iii) yield practically identical results when compared with 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 to moderate perturbations of hyper-parameters a_σ and b_σ.

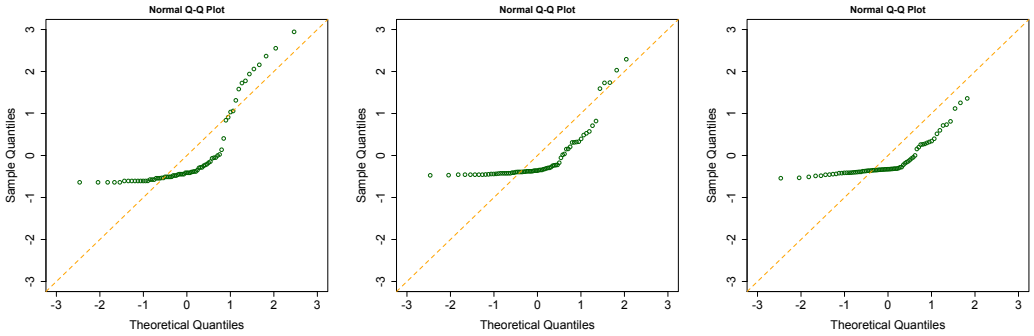
5 Brain Connectome Dataset with the Creative Achievement Questionnaire (CAQ)

Our dataset of interest consists of brain connectome information on several subjects collected using a brain imaging technique called *Diffusion Weighted Magnetic Resonance Imaging* (dMRI). It is openly available in the `Templeton 114` repository at `https://neurodata.io/mri`. Note that dMRI is a magnetic resonance imaging technique that measures the restricted diffusion of water in brain tissues in order to produce neural tract images which are then pre-processed using the NDMG pre-processing pipeline (Kiar *et al.*, 2016). In the

context of dMRI, 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 the 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, cingulate* and *insula* lobes.

Using dMRI, 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 *feature of interest*. Creative achievement can be perceived as an aggregate of creative products of an individual during his/her lifetime (Carson *et al.*, 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 brain network information and CAQ, age and sex are also available and are treated as *auxiliary features* for $n = 73$ subjects in our dataset of interest. While there is earlier literature suggesting an effect of age on brain connectivity (Baum *et al.*, 2017), all subjects in our dataset belong to the age group of 18-29 years with very little variation, which prompts us to ignore ROI specific age effects. We also find in the analysis in Section 5.1 that the age effects are practically insignificant in almost all the clusters, which further justifies our argument.

The main objective of the data analysis lies in supervised clustering of brain networks of the 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 on CAQ, age and sex. The model offers inference



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

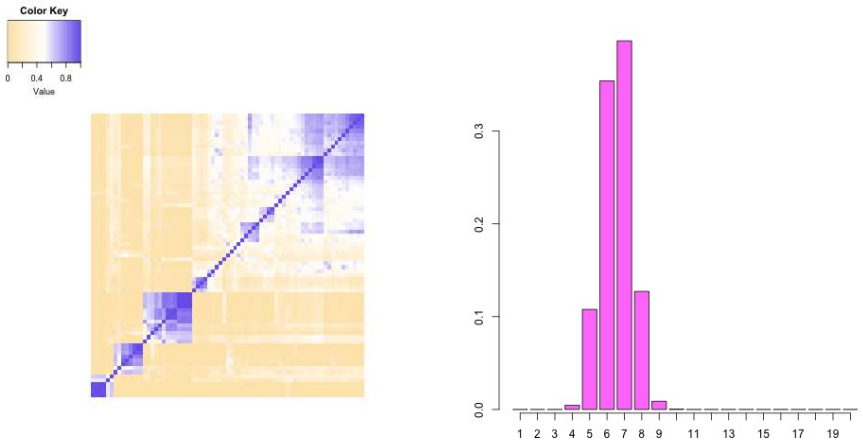
Figure 3: QQ-plot of residuals corresponding to linear regression fitted on three representative cells (edges in the brain network) with $n = 73$ subjects of the CAQ dataset.

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. As a byproduct of our clustering exercise using the network mixture model, the normality assumption on the errors of the network response matrix is automatically relaxed. This is deemed appropriate for this dataset, since after fitting linear regression models independently on each cell of the network response matrix with CAQ, age and sex as predictors, we have observed visible non-normality in the standardized residuals (refer to the QQ plots of the standardized residuals for three representative cells in Figure 3).

5.1 Findings from CAQ Brain Connectome Data

We fit NRMM with $H = 20$. NRMM, on the CAQ dataset, identifies 7 clusters with 25, 13, 6, 6, 7, 8 and 8 subjects in the clusters, respectively. 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 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. 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 the performances of NRMM and NRR with respect to the Posterior Predictive Loss Criterion (Gelfand and Ghosh, 1998), which is



(a) Uncertainty in Clustering (b) Posterior Dist. of no. of Clusters

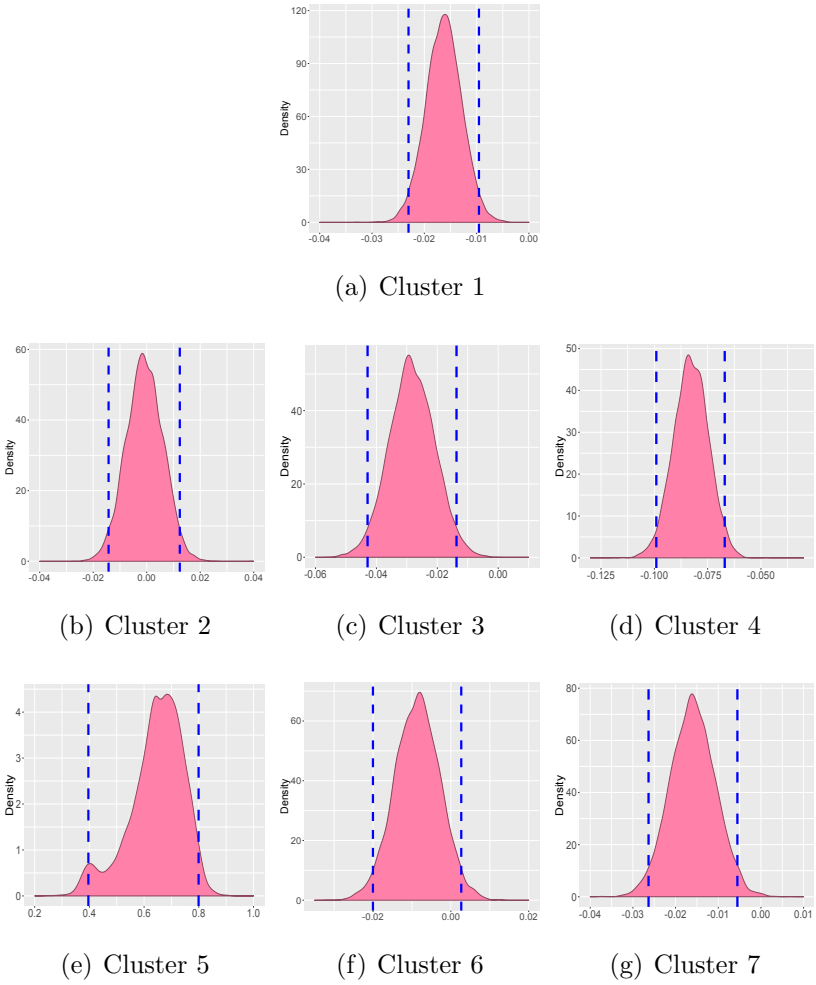
Figure 4: **CAQ Data**: Figure (a) shows the uncertainty in estimating the clusters. Figure (b) shows the barplot corresponding to the posterior distribution of the estimated number of clusters. The inference is presented for $H = 20$.

Table 5: Model fitting statistics for NRMM and NRR.

Competing Models	G	P	D
NRMM	101921	99711	201632
NRR	101727	101734	203461

calculated as $D = G + P$, such that a model corresponding to a lower value of D is preferred. Referring to Table 5, the model fitting statistic D shows a better performance of NRMM over NRR. HOLRR, being a frequentist method, is not included in this comparison. We also compute leave-one-out of sample MSPE for the three competitors and they turn out to be $0.64, 0.73, 0.71$ for NRMM, NRR and HOLRR, respectively.

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 6 depicts satisfactory point prediction along with an excellent characterization of predictive uncertainty. Referring to the high degree of non-normality in the error distributions discussed in Section 5, it is instructive to see if the mixture modeling



(a) Cluster 1

(b) Cluster 2 (c) Cluster 3 (d) Cluster 4

(e) Cluster 5 (f) Cluster 6 (g) Cluster 7

Figure 5: Plots of age coefficient in each cluster. The 95% posterior credible intervals are shown through the space between the two dotted lines.

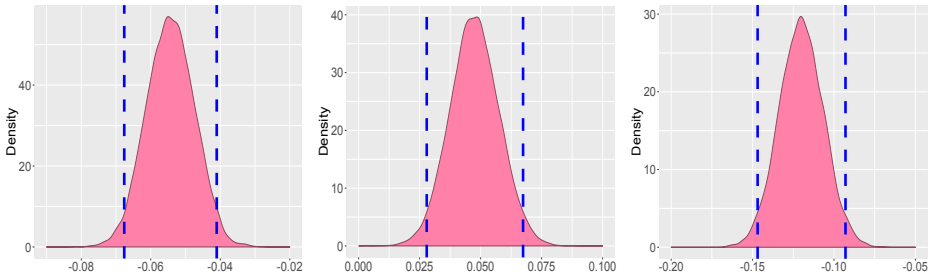
framework justifies the normality assumption on the error distribution in each cluster. To check this, cell by cell Kolmogorov-Smirnov tests are conducted by comparing the discrepancy between the posterior mean of residuals and the normal distribution. Out of 2278 network matrix cells in each cluster, 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 for Clusters 2 and 6, all other age coefficients turn out to be significant. Digging a bit deeper, we find that Clusters 2 and 6 show significantly lower variability in the ages of the subjects included, compared to the other clusters, which explains the age coefficient

Table 6: 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

being statistically insignificant in these clusters. Also, except for Cluster 5, the posterior mean of the age coefficients are found to be negative in all other clusters, implying a negative association between creativity and age. In all six clusters where gender is added as a variable, it is found to be significantly related to creativity (see Figure 6).



(a) Cluster 1 (b) Cluster 2 (c) Cluster 4

(d) Cluster 5 (e) Cluster 6 (f) Cluster 7

Figure 6: Plots of sex coefficient in each cluster. The 95% posterior credible intervals are shown through the space between the two dotted lines.

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 posterior probability of being 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 in the 7 clusters. A considerable proportion of ROIs detected in each cluster are part of the *frontal*, *cingulate*

and *temporal* lobes in both hemispheres. The frontal lobe has been scientifically associated with divergent thinking, problem solving ability, spontaneity, memory, language, judgement and social behavior (Stuss *et al.*, 1985; Razumnikova, 2007; Miller and Milner, 1985; Kolb and Milner, 1981). Prior scientific studies also suggest that increased cortical volume of the superior frontal gyrus, and ventromedial prefrontal cortex, and decreased volume of the dorsal and rostral anterior cingulate cortex are associated with CAQ (Chen *et al.*, 2014). Finkelstein *et al.*, 1991 also report *de novo* artistic expression to be associated with the frontal and temporal regions. Our identification of the middle frontal gyrus and the inferior occipital gyrus being associated with CAQ is also supported by earlier scientific studies (e.g., Shi *et al.*, 2017). Additionally, CAQ being related to the medial superior frontal gyrus and orbitofrontal insula have also been supported by previous studies (Chen *et al.*, 2014).

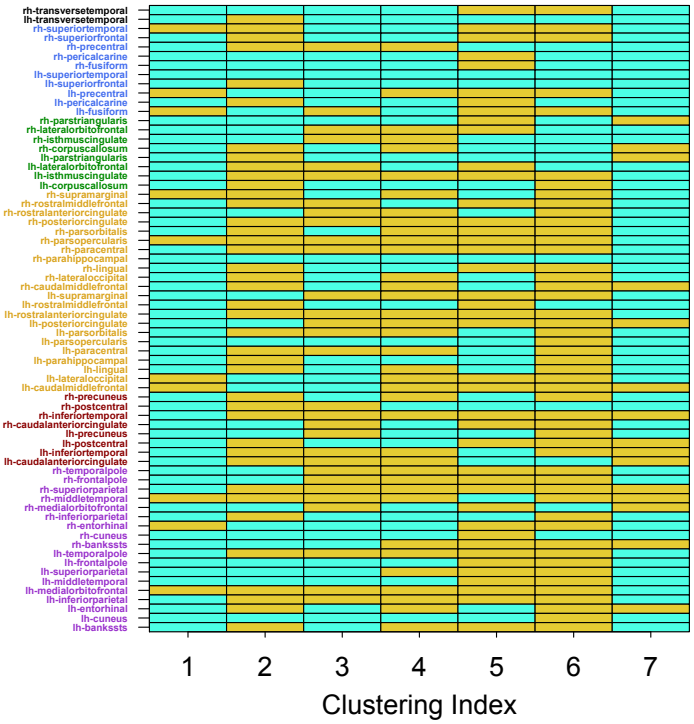


Figure 7: **CAQ Data**: Figure plots a 68×7 matrix with the rows and columns corresponding to the ROIs and clusters, respectively. A green cell in the (k, h)th entry of the matrix implies that the kth ROI in the hth cluster is not significantly related to creativity. Prefix 'lh-' and 'rh-' in the ROI 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. From bottom to top, the group of ROIs under the same color correspond to the temporal, cingulate, frontal, occipital, parietal and insula lobes.

6 Conclusion and Future Work

This article is motivated by the need to develop a flexible relationship between the brain network and creativity from subjects in a brain connectome dataset. Viewing the brain image for each subject as an undirected network, we propose a 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 network nodes significantly related to a predictor. Empirical investigations validate our NRMM framework and yield superior inference over relevant competitors. The NRMM framework 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.

7 Acknowledgements

The second author acknowledges funding from National Science Foundation Grant DMS-2210672, DMS-2220840, and National Institute Of Neurological Disorders And Stroke of the National Institutes of Health under Award Number R01NS131604.

Supplementary Material

Narrative supplement. The narrative supplementary material discusses convergence properties of the proposed NRMM with proofs of the theoretical results. It shows performance of NRMM and its competitors under mis-specification. It also presents full conditional distributions of the model parameters for running MCMC and shows MCMC traceplots for selected model parameters. Finally, this supplementary material presents full conditional distributions for implementing MCMC for the proposed model.

Code supplement. We also include code and data used in this paper as a separate supplementary material. The codes reproduce Table 5 in the article.

References

Amewou-Atisso, M., Ghosal, S., Ghosh, J. K., Ramamoorthi, R., *et al.* (2003). Posterior consistency for semi-parametric regression problems. *Bernoulli*, **9**(2), 291–312.

Baum, G. L., Ciric, R., Roalf, D. R., Betzel, R. F., Moore, T. M., Shinohara, R. T., Kahn, A. E., Vandekar, S. N., Rupert, P. E., Quarmley, M., *et al.* (2017). Modular segregation of structural brain networks supports the development of executive function in youth. *Current Biology*, **27**(11), 1561–1572.

Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews. Neuroscience*, **10**(3), 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**(1), 37–50.

Chen, Q., Yang, W., Li, W., Wei, D., Li, H., Lei, Q., Zhang, Q., and Qiu, J. (2014). Association of creative achievement with cognitive flexibility by a combined voxel-based morphometry and resting-state functional connectivity study. *Neuroimage*, **102**, 474–483.

Chi, E. C. and Lange, K. (2015). Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, **24**(4), 994–1013.

Chi, E. C., Allen, G. I., and Baraniuk, R. G. (2017). Convex biclustering. *Biometrics*, **73**(1), 10–19.

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

Duan, J. A., Guindani, M., and Gelfand, A. E. (2007). Generalized spatial dirichlet process models. *Biometrika*, **94**(4), 809–825.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, **90**(430), 577–588.

Finkelstein, Y., Vardi, J., and Hod, I. (1991). Impulsive artistic creativity as a presentation of transient cognitive alterations. *Behavioral medicine*, **17**(2), 91–94.

Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, **81**(395), 832–842.

Gahrooei, M. R., Yan, H., Paynabar, K., and Shi, J. (2021). Multiple tensor-on-tensor regression: an approach for modeling processes with heterogeneous sources of data. *Technometrics*, **63**(2), 147–159.

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

Guha, S. and Guhaniyogi, R. (2021). Bayesian generalized sparse symmetric tensor-on-vector regression. *Technometrics*, **63**(2), 160–170.

Guha, S. and Rodriguez, A. (2021). Bayesian regression with undirected network predictors with an application to brain connectome data. *Journal of the American Statistical Association*, **116**(534), 581–593.

Guhaniyogi, R. and Spencer, D. (2021). Bayesian tensor response regression with an application to brain activation studies. *Bayesian Analysis*, **16**(4), 1221–1249.

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

Hannah, L. A., Blei, D. M., and Powell, W. B. (2011). Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, **12**(6).

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

Hoff, P. D. and Niu, X. (2012). A covariance regression model. *Statistica Sinica*, pages 729–753.

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**(460), 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**(488), 1609–1620.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, **2**(1), 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**(3), 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**(3), 398–409.

Kiar, G., Gray Roncal, W., Mhembere, D., Bridgeford, E., Burns, R., and Vogelstein, J. (2016). ndmg: Neurodata's MRI graphs pipeline.

Kim, Y. and Levina, E. (2019). Graph-aware modeling of brain connectivity networks. *arXiv preprint arXiv:1903.02129*.

Kolb, B. and Milner, B. (1981). Performance of complex arm and facial movements after focal brain lesions. *Neuropsychologia*, **19**(4), 491–503.

Lau, J. W. and Green, P. J. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, **16**(3), 526–558.

Lee, M., Shen, H., Huang, J. Z., and Marron, J. (2010). Biclustering via sparse singular value decomposition. *Biometrics*, **66**(4), 1087–1095.

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**(5), 1411–1423.

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**(9), e73186.

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**(3), 371–379.

Müller, P., Erkanli, A., and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, **83**(1), 67–79.

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic block structures. *Journal of the American Statistical Association*, **96**(455), 1077–1087.

Pavlović, D. M., Guillaume, B. R., Towlson, E. K., Kuek, N. M., Afyouni, S., Vértes, P. E., Yeo, B. T., Bullmore, E. T., and Nichols, T. E. (2020). Multi-subject stochastic blockmodels for adaptive analysis of individual differences in human brain network cluster structure. *NeuroImage*, page 116611.

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

Razumnikova, O. M. (2007). Creativity related cortex activity in the remote associates task. *Brain Research Bulletin*, **73**(1), 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**(3), 1648–1677.

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2009). Bayesian nonparametric functional data analysis through density estimation. *Biometrika*, **96**(1), 149–162.

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**(5), 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**(1), 25–32.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650.

Shahbaba, B. and Neal, R. (2009). Nonlinear models using dirichlet process mixtures. *Journal of Machine Learning Research*, **10**(8).

Shi, B., Cao, X., Chen, Q., Zhuang, K., and Qiu, J. (2017). Different brain structures associated with artistic and scientific creativity: a voxel-based morphometry study. *Scientific reports*, **7**(1), 42911.

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**(1), 41–47.

Wang, L., Durante, D., Jung, R. E., and Dunson, D. B. (2017). Bayesian network–response regression. *Bioinformatics*, **33**(12), 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*, pages 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*, pages 138–149. Springer.

Zhao, Y., Wang, B., Mostofsky, S. H., Caffo, B. S., and Luo, X. (2021). Covariate assisted principal regression for covariance matrix outcomes. *Biostatistics*, **22**(3), 629–645.

Zou, T., Lan, W., Wang, H., and Tsai, C.-L. (2017). Covariance regression analysis. *Journal of the American Statistical Association*, **112**(517), 266–281.