Bayesian Supervised Clustering of Undirected Networks with Cluster Specific Inference on significant Nodes and Edges Related to Predictors

Sharmistha Guha

Postdoctoral Associate, Department of Statistical Science,

Duke University, Old Chemistry Building, Durham, NC 27708, E-mail: sg516@duke.edu

Rajarshi Guhaniyogi

Associate Professor, Department of Statistics,

UC Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, E-mail: rguhaniy@ucsc.edu

April 10, 2021

Abstract

Motivated by the connectome datasets acquired from various imaging modalities, this article focuses on model based clustering of subjects according to the shared relationship of subject-specific networks and covariates. Additionally, it is of interest to identify 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 are embedded with low-rankness and group sparsity within the low-rank structure. While the low-rank structure on 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 framework allows precise characterization of uncertainty in identifying significant network nodes in each cluster.Theoretically, we establish convergence of the posterior predictive density from the proposed model to the true data generating density at a rate very close to the finite dimensional optimal rate of $n^{-1/2}$. Empirical results in various simulation scenarios illustrate substantial inferential gains of the proposed framework in comparison with competitors. Analysis of a brain connectome data with the proposed model reveals interesting insights into the brain regions of interest (ROIs) significantly related to creative achievement in each cluster of subjects.

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

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.

The dataset offers interesting opportunities to characterize the relationship between brain networks and brain related phenotypes for subjects included in the analysis. Motivated

by such neuro-scientific applications, we undertake modeling endeavor primarily aimed 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. Additionally, inferential interest lies in identifying nodes and edges 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) and interconnections between them significantly associated with creative achievement in each cluster. Moreover, the objective also lies in achieving these inferential goals with parsimony in the fitted model and computational efficiency in the model fitting process.

We propose a novel nonparametric Bayesian modeling approach to achieve the aforementioned inferential objectives simultaneously. To be more specific, a Dirichlet process (DP) mixture of network response regression models 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 assumed to have a low-rank for parsimony and computational efficiency. We additionally impose a node-wise sparsity structure using a Bayesian spike-and-slab variable selection prior for identifying network nodes significantly associated with the 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. Our framework does not involve any expensive matrix manipulation and allows parallelization for efficient computation with a large number of network nodes.

An important contribution of this article is proving the near optimal contraction rate for the predictive density of the mixture of network regression models. The literature on the theory of posterior contraction rates for high dimensional linear and generalized linear regression models have observed significant development in the last decade (Castillo *et al.*, 2012; Belitser and Nurushev, 2015; Jeong and Ghosal, 2020). Similarly, there is a well developed literature on the posterior contraction rate for Gaussian and non-Gaussian mixture models in both supervised and unsupervised settings (Genovese *et al.*, 2000; Ghosal *et al.*,

2007; Choi, 2008). In contrast, to the best of our knowledge, there is no theoretical literature on the posterior contraction rates for the mixture of network regression models. This article lays down sufficient conditions on the number of network nodes, ranks of network coefficients and the number of fitted mixture components as a function of the sample size to obtain a *near optimal* convergence rate for the posterior predictive density of the proposed mixture of network regression models. On a related note, a few recent articles invest in studying posterior contraction properties of linear regressions and generalized linear models involving high dimensional tensor response and predictors (Guhaniyogi, 2017; Guhaniyogi *et al.*, 2018; Guha and Guhaniyogi, 2020), though none of them consider mixture of regression models involving network response. We establish the novelty of our proposal in the light of the existing literature discussed below.

Rather than focusing on multiple network observations collected over different individuals, an overwhelming literature with 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 *et al.*, 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). 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 (Wang *et al.*, 2017) exploits the relational nature of the network response, though it does not offer clusters of subjects and is not designed to detect network nodes significantly related to a scalar predictor. On a related note, there is an emerging literature on supervised stochastic block models (Kim and Levina, 2019; Pavlović *et al.*, 2020) focusing on clustering nodes of the network into groups, which is scientifically/metjodologically a different problem than our focus of clustering subjects into groups.

Viewing networks as symmetric tensors, our inferential problem can also be formulated

under a tensor response regression framework with a symmetric tensor response and scalar predictors. While an overwhelming literature on tensor response regression does not enforce any symmetry constraint on the tensor response (Guhaniyogi *et al.*, 2017, 2018; Spencer *et al.*, 2020), there are recent efforts (Sun and Li, 2017; Guha and Guhaniyogi, 2020) to devise new classes of models which are equipped to incorporate a symmetry constraint for the tensor response in the modeling framework. However, these approaches are based on two assumptions both of which may appear to be restrictive for a variety of neuro-scientific applications. First, the variance of the response for all tensor cells are free of the predictors. Second, the same set of network nodes influence the regression function in a similar manner for every individual.

While our framework treats the network as a response, a few recent approaches (Guha and Rodriguez, 2018; 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 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 in Guha and Rodriguez (2018) may become severely 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.

In fact the earlier literature in neuroscience provides substantial 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 identifying such subgroups and ascertaining subgroup differences have somewhat lagged behind the increasingly routine collection of such data. One possibility

is to reshape the network as 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; Dunson *et al.*, 2007; Duan *et al.*, 2007; Rodríguez *et al.*, 2009; Amewou-Atisso *et al.*, 2003; Hannah *et al.*, 2011; DeYoreo and Kottas, 2018). These approaches are less suitable to our problem of interest since they ignore the network topology in the process of model building and do not allow drawing inference on network nodes. In this context, it is also possible to invoke the literature on clustering of matrices or higher order tensor objects 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; 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.

The rest of the article progresses as following. Section 6 provides a brief description of the brain connectome data and the inferential objectives. Sections 2 and 4 describe the model development and posterior computation, respectively. Empirical investigation of the model with simulation studies and the brain connectome data analysis are presented in Sections 5 and 6.1, respectively. Finally, Section 7 concludes the paper with an eye towards future work.

2 Supervised Clustering of Undirected Networks: Model and Prior Formulation

2.1 Notations and Framework

For $i = 1, ..., n$, let $\boldsymbol{Y}_i \in \mathcal{Y} \in \mathbb{R}^{p \times p}$ denote the weighted undirected network response with p nodes, $\boldsymbol{x}_i = (x_{i1}, ..., x_{im})'$ 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 ($y_{i,(j_1,j_2)} = y_{i,(j_2,j_1)}$). 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. *Since $\boldsymbol{Y}_i$ is symmetric with 0 diagonal entries, it suffices to build a probabilistic generative mechanism for the upper triangular vector, or the vector of edges for the undirected network given by, $\boldsymbol{y}_i = (y_{i,\boldsymbol{j}} : 1 \leq j_1 < j_2 \leq p)'$ 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 ordinary reshaping $\boldsymbol{Y}_i$ 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,

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

where $\mathbf{1}_q$ 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., $\boldsymbol{\beta}_s = (B_{s,\boldsymbol{j}} : 1 \leq j_1 < j_2 \leq p)'$. Equation (1) 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 reshaped network response $\boldsymbol{Y}_i$ and scalar predictors. While such an approach would have lost information on the nodes each edge is connected to, $B_{s,\boldsymbol{j}}$ coefficients in our modeling framework (1) allows us to draw inference on network nodes significantly related to the predictors. We further elaborate 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 $\boldsymbol{\Delta}_h^* = (\boldsymbol{\beta}_{1,h}^*, .., \boldsymbol{\beta}_{m,h}^*, \gamma_{0,h}^*, \gamma_{1,h}^*, .., \gamma_{l,h}^*) \sim G_0$. 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 $\alpha > 0$,

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

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

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

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

Next, we turn into identifying network nodes in different clusters significantly associated with 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

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

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, $\lambda_{s,h,r} \in \{-1, 0, 1\}$ determines if the rth summand in (4) 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 $\lambda_{s,h,r}$, (4) represents a low-rank decomposition of the symmetric matrix coefficient $\boldsymbol{B}^*_{s,h} = \boldsymbol{U}_{s,h} \boldsymbol{\Lambda}_{s,h} \boldsymbol{U}'_{s,h}$, which is able to approximate any matrix to an arbitrary level of accuracy for appropriate choices of R. Since the choice of R is arbitrary, allowing $\lambda_{s,h,r}$ to be 0 protects the model from over-fitting. 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 both are 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 $R \ll p$.

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 the posterior inference on $\boldsymbol{u}_{s,h,k}$'s (without any constraint imposed on $\boldsymbol{u}_{s,h,k}$'s) may not be always meaningful, our focus is on the event $\{\boldsymbol{u}_{s,h,k} = \boldsymbol{0}\}$ for each k, which remains identifiable (since $\boldsymbol{0}$-valued latent vectors are invariant under orthogonal transformation) and is critical to drawing inference on the

nodes related to the s-th predictor of interest, as we describe next. In fact, 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

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

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. The parameters $\gamma^*_{0,h}, \gamma^*_{1,h}, ..., \gamma^*_{l,h}$ are assigned standard normal distributions. In order to learn which summands in (4) are informative, we assign a hierarchical prior

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

The choice of hyper-parameters of the beta distribution is crucial. In particular, note that $E[\delta_{\lambda_{s,h,r} \in \{-1,1\}}] = 2/(2 + r^\eta) \to 0$ as $r \to \infty$ and that $\sum_{r=1}^R var(\delta_{\lambda_{s,h,r} \in \{-1,1\}}) = \sum_{r=1}^R [\frac{2(r^\eta+1)}{(r^\eta+2)^2(r^\eta+3)} + \frac{2(r^\eta+1)}{(r^\eta+3)(r^\eta+4)}] < \infty$ as $R \to \infty$. The first property provides (weak) identifiability of the different latent dimensions, while the second ensures that $\lim_{R\to\infty} var(R_{eff}) < \infty$. The error variance σ^2 is assigned a 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:

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

The model and prior 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.

3 Convergence Rate for Predictive Densities

This section presents posterior convergence properties of the proposed network response mixture model (NRMM). We adopt the framework outlined in Jiang *et al.* (2007), with some important differences. While Jiang *et al.* (2007) studies the convergence rate of the posterior predictive distribution with a scalar response and a high dimensional vector predictor without considering any mixture of distribution, we focus on mixture of densities involving a network response and a vector predictor. The novel model development and the prior structure described in Section 2 of the main article present theoretical challenges which are unique and very different from Jiang *et al.* (2007).

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

3.1 Framework and Main Results

In what follows, we assume $m = 1$ predictor of interest (hence get rid of the subscript s for all parameters) and no auxiliary predictor for simplifying calculations, though the results assume straightforward extension to cases where $m > 1$ and $l > 1$. Without loss of generality, the predictor x satisfies $|x_i| < 1$ for all i. Let p_n denote the number of nodes and R_n denote the dimension of the node specific latent variables in presence of sample size n. We assume that p_n and R_n are both non-decreasing functions of n, with $R_n < p_n$ for all large n. Denote $\mathcal{J} = \{\boldsymbol{j} : 1 \leq j_1 < j_2 \leq p_n\}$ as the set of all indices. Hence, the number of elements in $\mathcal{J}$,

given by $q_n = p_n(p_n-1)/2$, also naturally becomes a function of n. This paradigm attempts to capture the fact that q_n grows with n, and a higher rank CP decomposition of $\boldsymbol{B}$ can be estimated more precisely in presence of a larger sample size n. We also add the subscript n to $\boldsymbol{B}_h$ and $\boldsymbol{u}_{h,k}$ to denote them by $\boldsymbol{B}_{n,h}$ and $\boldsymbol{u}_{n,h,k}$. The true density and the predictive density of the fitted model assume the form of Gaussian mixture distributions with the same number of mixture components as given below,

$$f(\boldsymbol{Y}|x) = \sum_{h=1}^{H_n} \omega_h \prod_{\boldsymbol{j}\in\mathcal{J}} f_{\boldsymbol{j}}(Y_{\boldsymbol{j}}|x, B_{n,h,\boldsymbol{j}}),\ f_{\boldsymbol{j}}(Y_{\boldsymbol{j}}|x) = \frac{1}{\sqrt{2\pi}}\exp\{-(Y_{\boldsymbol{j}} - B_{n,h,\boldsymbol{j}}x)^2/2\}$$
$$f_T(\boldsymbol{Y}|x) = \int \prod_{\boldsymbol{j}\in\mathcal{J}} f_{\boldsymbol{j}}(Y_{\boldsymbol{j}}|x, B_{T,n,\boldsymbol{j}})dG_T(B_{T,n,\boldsymbol{j}}),\ G_T = \sum_{h=1}^{H_n} \omega_{h,T}\delta_{B_{T,n,h,\boldsymbol{j}}}. \tag{6}$$

To show the theoretical results, we make a number of simplifications to our model setting as dicussed in the next paragraph. We emphasize that our analysis on posterior contraction rate of can be extended without such simplifications, though it will require substantially more algebraic manipulations.

Similar to each $\boldsymbol{B}_{n,h}$, the true tensor coefficients $\boldsymbol{B}_{T,n,h}$ (having the $\boldsymbol{j}$th cell as $B_{T,n,h,\boldsymbol{j}}$, $\boldsymbol{j} \in \mathcal{J}$) also assumes symmetric matrix decomposition with rank R_n, i.e., $B_{T,n,h,\boldsymbol{j}} = \sum_{r=1}^{R_n} u^{(r)}_{T,n,h,j_1} u^{(r)}_{T,n,h,j_2}$, for $\boldsymbol{j} \in \mathcal{J}$. Although this is a somewhat restrictive assumption, it has been frequently employed in earlier theoretical literature on tensor regressions for simplifying calculations (Guhaniyogi *et al.*, 2017, 2018). With $\boldsymbol{B}_{n,h}$ having the same rank with $\boldsymbol{B}_{T,n,h}$, no rank selection is necessary in our framework. Thus, we assume $\lambda_r = 1$ for all r. Additionally, we assume that the fitted mixture weights $(\omega_1, ..., \omega_{H_n})$ follows a Dirichlet distribution in model fitting for simplifying calculations, though with little extra algebra, our results can be extended to the setting where ω_h's assume a stick breaking representation. Finally, we set $\boldsymbol{M}_h = \boldsymbol{I}$ for all h for simplifying calculations.

For two sequences c_n and d_n, let $c_n \prec d_n$ signifies $c_n/d_n \to 0$ as $n \to \infty$. With these notations, we state the following theorem, the proof of which can be found in the Appendix A.

Theorem 3.1 *For a sequence ϵ_n satisfying $0 < \epsilon_n < 1$ and $n\epsilon_n^2 \to \infty$ and sequences C_n and D_n, let the following conditions hold*

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

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

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

(iv) $H_n R_n p_n \log(C_n) \prec n\epsilon_n^2$

(v) $\limsup_{n\to\infty} \sum_{k=1}^{p_n} ||\boldsymbol{u}_{T,n,h,k}|| < \infty$, *where* $\boldsymbol{u}_{T,n,h,k} = (u^{(1)}_{T,n,h,k}, ..., u^{(R_n)}_{T,n,h,k})'$, *for all* $h = 1, ..., H_n$.

Then, $\lim_{n\to\infty} P_{f_T}\left[\Pi\{\mathcal{D}_H(f, f_T) > 4\epsilon_n | \{\boldsymbol{Y}_i, x_i\}_{i=1}^n\} < 2e^{-n\epsilon_n^2}\right] = 1.$

The assumptions in Theorem 3.1 lead to the following convergence rate result for the predictive density of the fitted model.

Corollary 3.2 *Assume* p_n *grows at a rate slower than* n^θ, $\theta < 1$ *(i.e.* $p_n \leq C_1^* n^\theta$*), the tensor rank* R_n *grows at a much slower rate of* $(\log n)^{k_1}$ *for some* k_1 *(i.e.* $R_n \leq C_2^* (\log n)^{k_1}$*) and the number of fitted mixture components* H_n *also grows at a much slower rate of* $(\log n)^{k_2}$ *for some* k_2 *(i.e.* $H_n \leq C_3^* (\log n)^{k_2}$*). Choose* C_n *such that* $n^{\mu_1} \leq C_n \leq n^{\mu_2}$, *for some* μ_1, μ_2 *satisfying* $\theta/2 < \mu_1 < \mu_2$. *Then, the convergence rate* ϵ_n *can be taken as* $\epsilon_n \sim n^{-(1-\theta)/2} (\log n)^{(k_1+k_2)/2+2}$.

It is evident that the convergence rate is a function of how the number of tensor nodes, the rank of the true tensor (same as the rank of the fitted tensor) and the number of fitted mixture components grow with n. Intuitively, p_n should grow at a much faster rate than R_n, and both should be bounded by an appropriate function of n to achieve a good convergence rate. Finally, it is worth noting that for any value of k_1 and k_2, $(\log n)^{(k_1+k_2)/2+1} \prec n^{\theta/2}$. Thus the convergence rate $\epsilon_n \sim n^{-1/2+\theta}$ which is close to the "finite dimensional" optimal rate of $n^{-1/2}$ when θ is very close to 0.

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 posterior distributions. Details of the Markov chain Monte Carlo algorithm are presented in Appendix B. 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 look at (i) the point estimate of clustering 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),

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

where the ratio o_1/o_2 controls the relative loss of incorrectly clustering or separating a pair of samples. In our illustrations we set $o_1/o_2 = 1$.

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

$$\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i \sim \sum_{h=1}^{H_0} \omega_{h,0} N(\mathbf{1}\gamma^*_{0,h,0} + \mathbf{1}\sum_{s=1}^{l} \gamma^*_{s,h,0} z_{is} + \sum_{s=1}^{m} \boldsymbol{\beta}^*_{s,h,0} x_{is}, \sigma_0^2 \boldsymbol{I}_q), \tag{8}$$

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

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

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

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	30	100	2	5	0.6	3
2	30	100	2	5	0.3	3
3	30	100	3	5	0.6	4
4	80	100	2	5	0.6	3
5	80	100	2	5	0.3	3
6	80	100	3	5	0.6	2
7	30	100	2	5	0.6	1

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_0^2 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. In particular, the last two cases include specifications where the network coefficient in a cluster is full rank, where as the fitted model assumes a low-rank structure for network coefficients in all the clusters. This will allow assessing 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 5.5.

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 $E_0[\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i] = \sum_{h=1}^{H_0} \omega_{h,0} \left(\mathbf{1}_q \gamma^*_{0,h,0} + \mathbf{1}_q \sum_{s=1}^{l} \gamma^*_{s,h,0} z_{is} + \sum_{s=1}^{m} \boldsymbol{\beta}^*_{s,h,0} x_{is} \right)$. In particular, we compute the mean squared error (MSE) of estimating the true regression mean function over all data points, given by $\frac{1}{nq} \sum_{i=1}^{n} ||E_0[\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i] - \widehat{E[\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i]}||^2$, where $\widehat{E[\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i]}$ 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

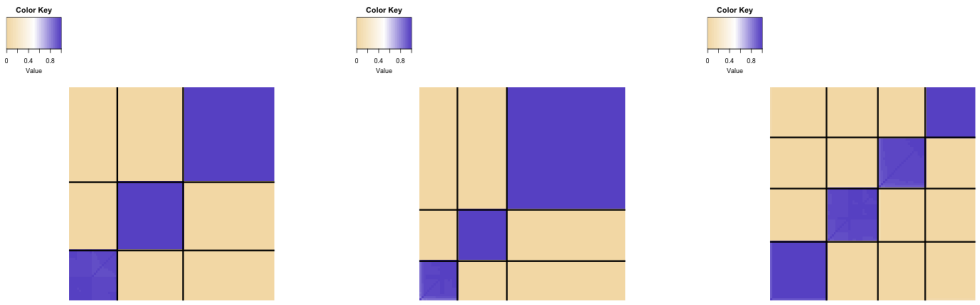
All model parameters show excellent convergence with fairly uncorrelated post burn-in samples to draw posterior inference. To demonstrate this, we present the effective sample size (ESS) corresponding to 10000 post burn-in samples from NRMM for all simulation examples (see Table 2). Trace-plots for MCMC chains for a few representative parameters are presented in Appendix B. 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. 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 1 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 2. 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 mildly 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

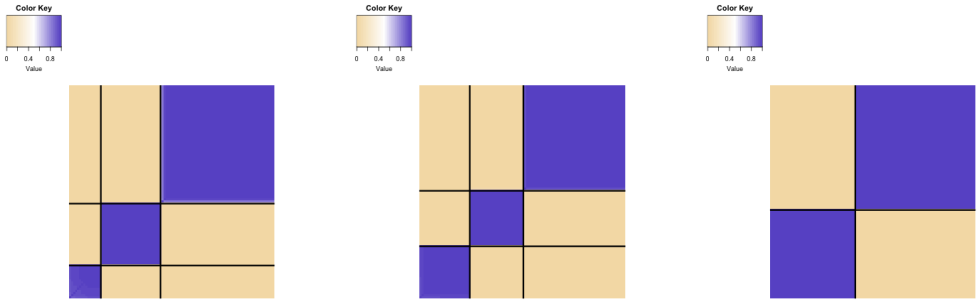
of H assigned clusters, most are not populated in each case, justifying the choice of $H = 15$ in each case.

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 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 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.

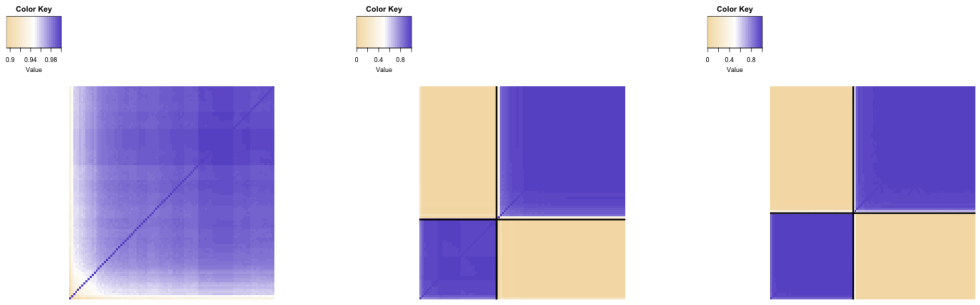
Note that inference on each cluster is not readily available from the mixture model due to the clusters being not identifiable. Thus, to draw inference on which network nodes are influential in each cluster, we fix the cluster membership indicator c_i for the ith sample at $\hat{c}_i$ (the estimated 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 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 1-7, our proposed model correctly identifies each cluster in every simulation,



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

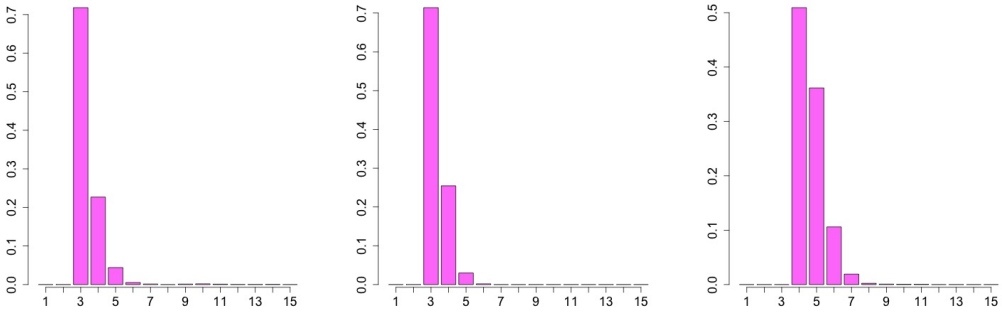


(d) Case 4,$H_0 = 3$ (e) Case 5,$H_0 = 3$ (f) Case 6,$H_0 = 2$

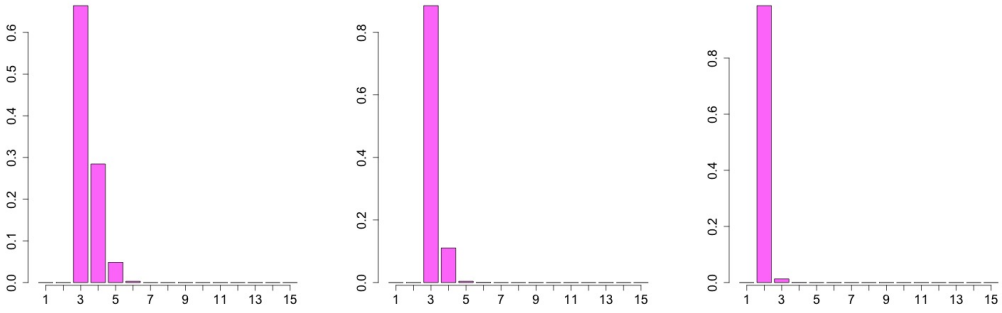


(g) Case 7,$H_0 = 1$ (h) Case 8,$H_0 = 2$ (i) Case 9,$H_0 = 2$

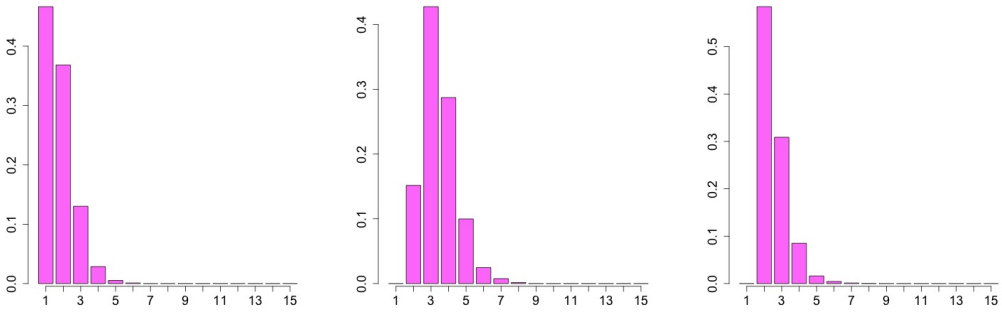
Figure 1: Plots showing uncertainty in estimating clusters in simulation cases 1-9. Boldfaced horizontal and vertical lines indicate the true clustering.



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

Figure 2: Plots showing the posterior distribution of the number of clusters in the simulation cases 1-9.

Table 2: The first column presents Effective sample size for NRMM corresponding to the 10000 post burn-in iterations to assess the convergence of the MCMC sampler for NRMM. 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	ESS	ARI	TPR	FPR		NRMM	NRR	HOLRR
1	8006	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	7985	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	7942	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	7235	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	7451	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	7324	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	8106	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	8195	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	7839	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	–

Table 3: Computation time (in seconds) per MCMC iteration of the NRMM model with $H = 15$ mixture components.

V	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

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)= $\frac{\text{TP}}{\text{TP+FP}}$ 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.

5.4 Computational complexity and time

The Gibbs sampler for model estimation does not involve any expensive matrix inversion or multiplication, leading to fast computation. Further, the Gibbs sampler can be suitably parallelized since the updates of $\boldsymbol{u}_{s,h,k}$ can be performed over different processors in parallel. The computation time (in seconds) per MCMC iteration for NRMM model is provided in Table 3. The entries in the table are recorded assuming $H = 15$ mixture components are fitted to the data.

5.5 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)

Table 4: 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

$a = 1, b = 1, \nu = 50$. 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 higher prior mean of 5 for $\xi_{h,k}$ which encourages higher number of activated nodes. (iii) presents 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, 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 number of activated nodes is small (combination (i)), MSE is found to be little higher than what is presented in Table 2 under case 2. Similarly, the coverage is found to be 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 the similar value of ARI. The results are also found to be not sensitive at all with the moderate perturbation of hyper-parameters a_σ and b_σ.

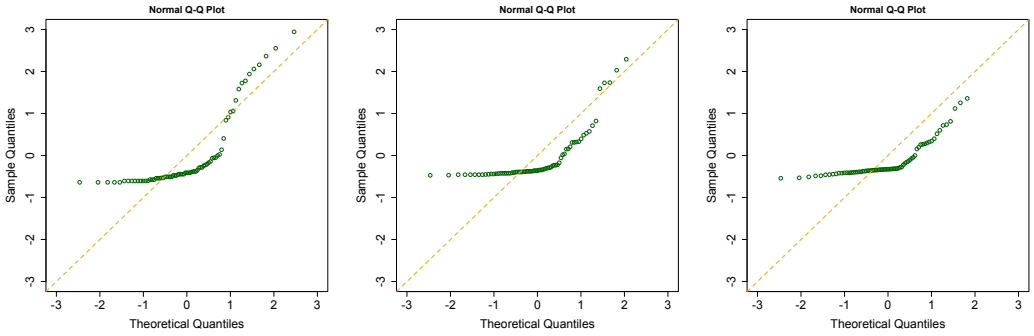
6 Brain Connectome Dataset with the Creative Achievement Questionnaire (CAQ)

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 in the repository named `Templeton 114` at `https://neurodata.io/mri`. Note that DWI is a magnetic resonance imaging technique that mea-

sures 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 *et al.*, 2017a; Kiar *et al.*, 2017b). 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 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 *feature of interest*. Creative achievement can be perceived as the 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 the 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 are earlier literature suggesting 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 prompted us to ignore ROI specific age effects. We also found in the analysis in Section 6.1 that the age effects are closely 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 from 73 subjects. The Bayesian mixture model of network objects proposed in this article



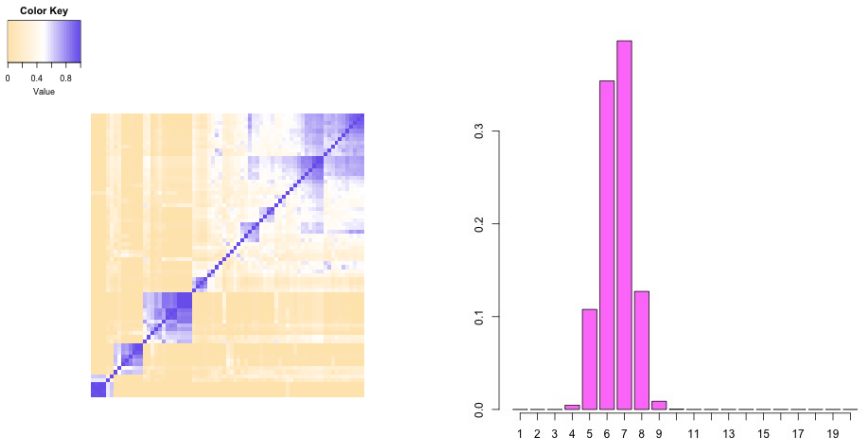
(a) QQ Plot: Cell 1 (b) QQ Plot: Cell 2 (c) QQ Plot: Cell 3

Figure 3: 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.

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 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 to our clustering exercise using the network mixture model, the normality assumption on the errors of the network response matrix is automatically relaxed. This 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 observe visible non-normality in the standardized residuals (refer to the QQ plots of the standardized residuals for three representative cells in Figure 3).

6.1 Findings from CAQ Brain Connectome Data

This section reports analysis of the CAQ brain connectome dataset described in Section 6. 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 somewhat higher degree of uncertainty among the pairs lying outside these three clusters. The posterior distribution of the number of clusters



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

Figure 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 $H = 20$.

(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. We also compute leave-one-out of sample mean squared prediction error (MSPE) for the three competitors and they turned 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

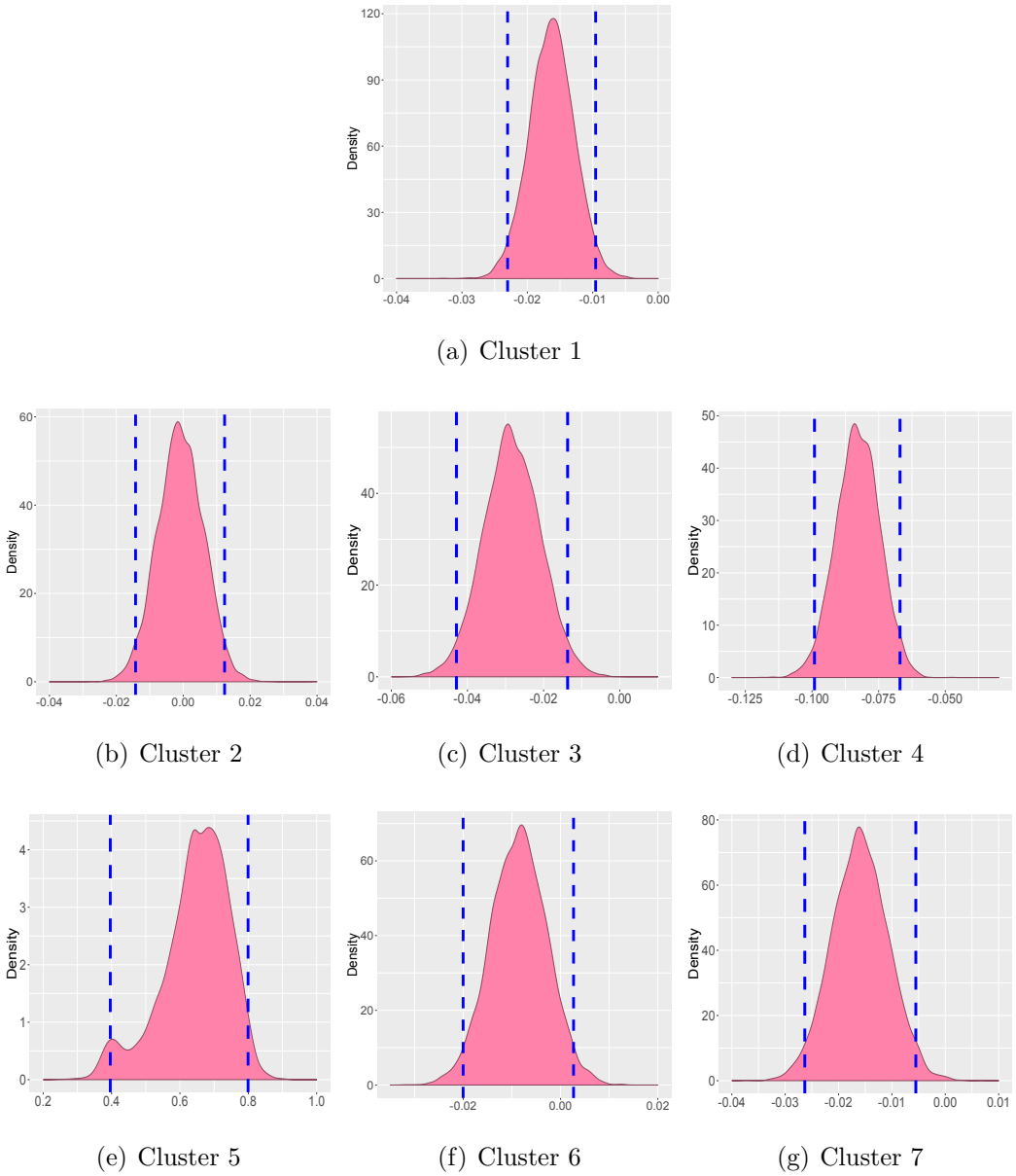
Table 5: MSPE, average coverage of 95% predictive intervals and average length of 95% predictive intervals for the seven clusters are provided.

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

average length of 95% predictive intervals averaged over all cells of the network response matrix and all subjects in a cluster. Table 5 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 6, it is instructive to see if the mixture modeling framework justifies normality assumption on the error distribution in each cluster. To check this, cell by cell Kolmogorov-Smirnov test 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 Cluster 2, all other age coefficients turn out to be significant. Digging a bit deeper, we found that Cluster 2 shows significantly lower variability in the ages of the subjects included compared to the other clusters, which explains age coefficient being statistically insignificant in this cluster. Also, except for Cluster 5, the poster mean of age coefficients are found to be negative in all other clusters, implying a negative association between creativity and age. Similarly, in 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 posterior probability of being active is 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



(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

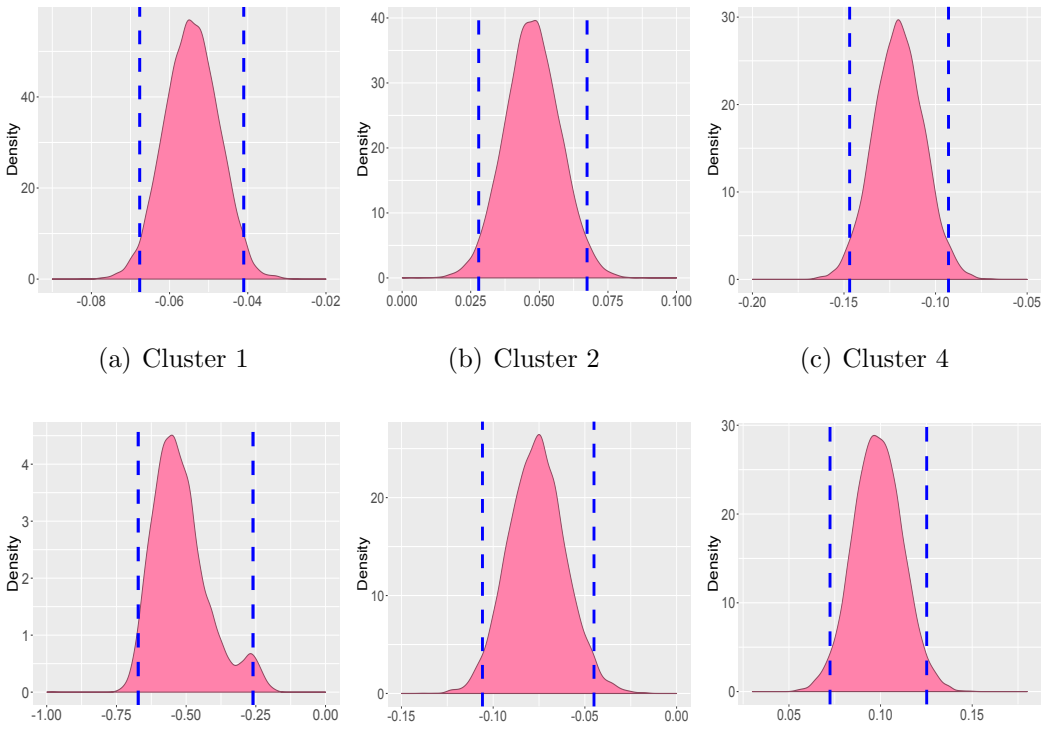
(d) Cluster 4

(e) Cluster 5

(f) Cluster 6

(g) Cluster 7

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



(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. 95% posterior credible intervals are shown through the space between the two dotted lines.

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 *et al.*, 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 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.

A Appendix

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

$$x(x + ||\boldsymbol{u}_{T,n,h,j_2}||) + x||\boldsymbol{u}_{T,n,h,j_1}|| = \delta_n. \tag{10}$$

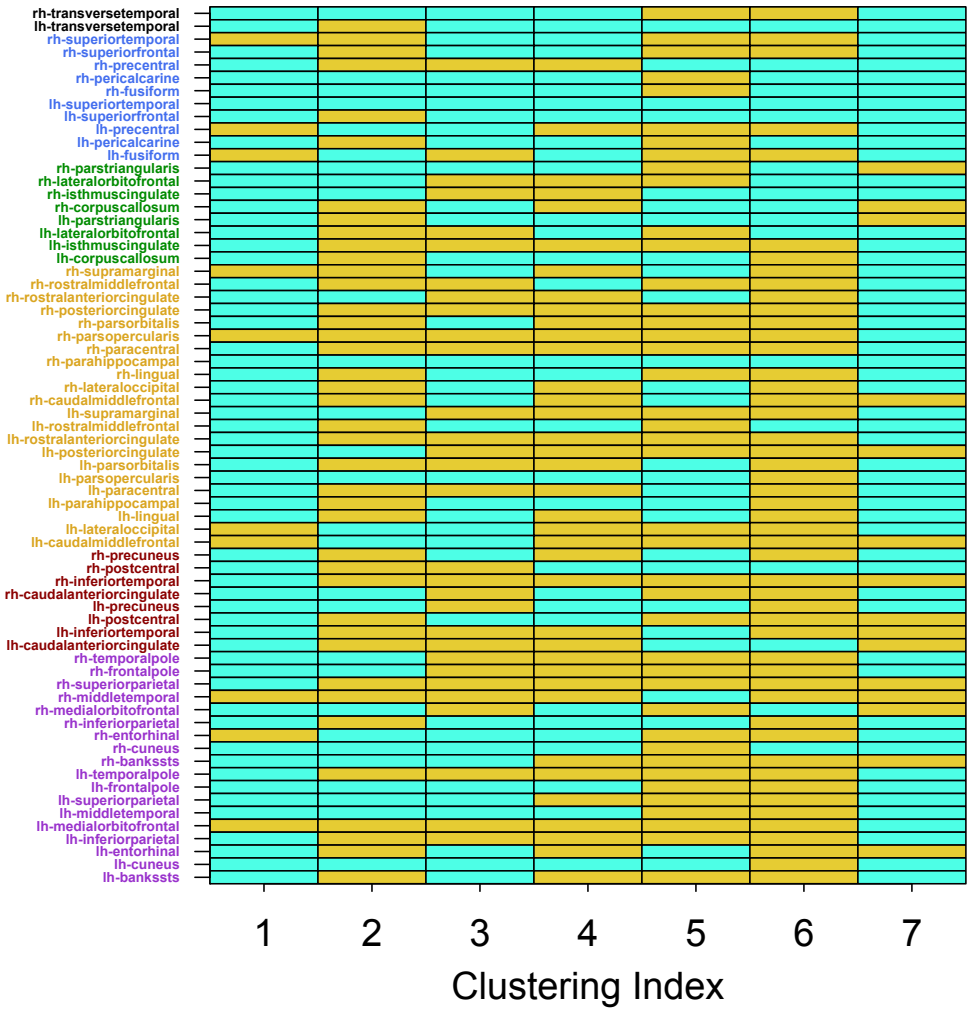


Figure 7: **CAQ Data**: Plots a 68×7 matrix with the rows and columns corresponding to the ROIs and clusters, respectively. A green cell in the (k, h)th entry of the matrix implies that the kth ROI in the hth cluster is not significantly related to creativity. Prefix 'lh-' and 'rh-' in the ROI names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

Assume $\gamma_{n,h} = min_{\boldsymbol{j}\in\mathcal{J}}\, \gamma_{n,h,\boldsymbol{j}}$. *Then,* $\Pi(||\boldsymbol{B}_{n,h} - \boldsymbol{B}_{T,n,h}||_\infty \leq \delta_n) \geq \Pi(||\boldsymbol{u}_{n,h,k} - \boldsymbol{u}_{T,n,h,k}|| \leq \gamma_{n,h}, k = 1, .., p_n)$, $h = 1, ..., H_n$.

Proof For $\boldsymbol{j} \in \mathcal{J}$,
$|B_{n,h,\boldsymbol{j}} - B_{T,n,h,\boldsymbol{j}}| = |\sum_{r=1}^{R_n} u_{n,h,j_1}^{(r)} u_{n,h,j_2}^{(r)} - \sum_{r=1}^{R_n} u_{T,n,h,j_1}^{(r)} u_{T,n,h,j_2}^{(r)}| = |\sum_{r=1}^{R_n} (u_{n,h,j_1}^{(r)} - u_{T,n,h,j_1}^{(r)}) u_{n,h,j_2}^{(r)}| +$
$|\sum_{r=1}^{R_n} (u_{n,h,j_2}^{(r)} - u_{T,n,h,j_2}^{(r)}) u_{T,n,h,j_1}^{(r)}| \leq ||\boldsymbol{u}_{n,h,j_1} - \boldsymbol{u}_{T,n,h,j_1}||||\boldsymbol{u}_{n,h,j_2}|| + ||\boldsymbol{u}_{n,h,j_2} - \boldsymbol{u}_{T,n,h,j_2}||||\boldsymbol{u}_{T,n,h,j_1}||$
If $||\boldsymbol{u}_{n,h,k} - \boldsymbol{u}_{T,n,h,k}|| \leq \gamma_{n,h},\ k = 1, .., p_n$, the above inequality implies that $|B_{n,h,\boldsymbol{j}} - B_{T,n,h,\boldsymbol{j}}| \leq \gamma_{n,h}(\gamma_{n,h} + ||\boldsymbol{u}_{T,n,h,j_2}||) + \gamma_{n,h}||\boldsymbol{u}_{T,n,h,j_1}|| \leq \delta_n$.
Thus $\Pi(||\boldsymbol{B}_{n,h} - \boldsymbol{B}_{T,n,h}||_\infty \leq \delta_n) \geq \Pi(||\boldsymbol{u}_{n,h,k} - \boldsymbol{u}_{T,n,h,k}|| \leq \gamma_{n,h},\ k = 1, .., p_n)$.

Lemma A.2 *With* $\gamma_{n,h}$ *and* $\boldsymbol{u}_{T,n,h,k}$ *defined as in Lemma A.1, for all* $h = 1, ..., H_n$,

$$\Pi(||\boldsymbol{B}_{n,h} - \boldsymbol{B}_{T,n,h}||_\infty \leq \delta_n) \geq e^{-\sum_{k=1}^{p_n} ||\boldsymbol{u}_{T,n,h,k}||^2/2} \left(\frac{1}{\sqrt{2\pi}}\right)^{R_n p_n} \frac{R_n p_n}{R_n p_n + 1} \left(\frac{2\gamma_{n,h}}{R_n}\right)^{R_n p_n} e^{-p_n \gamma_{n,h}^2/R_n}.$$

Proof For $h = 1, ..., H_n$,

$$\begin{aligned}
&\Pi(||\boldsymbol{B}_{n,h} - \boldsymbol{B}_{T,n,h}||_\infty \leq \delta_n) \geq \Pi(||\boldsymbol{u}_{n,h,k} - \boldsymbol{u}_{T,n,h,k}|| \leq \gamma_{n,h},\ k = 1, .., p_n) \\
&\geq E\left[\Pi(||\boldsymbol{u}_{n,h,k} - \boldsymbol{u}_{T,n,h,k}|| \leq \gamma_{n,h},\ k = 1, .., p_n|\xi)\right] \\
&\geq E\left[\prod_{k=1}^{p_n} \left\{e^{-||\boldsymbol{u}_{T,n,h,k}||^2/2} \Pi(||\boldsymbol{u}_{n,h,k}|| \leq \gamma_{n,h}|\xi)\right\}\right] \\
&= e^{-\sum_{k=1}^{p_n} ||\boldsymbol{u}_{T,n,h,k}||^2/2} E\left[\prod_{k=1}^{p_n} \Pi(||\boldsymbol{u}_{n,h,k}|| \leq \gamma_{n,h}|\xi)\right], \qquad (11)
\end{aligned}$$

where the first inequality follows from Lemma A.1 and the second inequality follows from the Anderson's Lemma. We will now make use of the fact that $\int_{-a}^{a} e^{-x^2/2} dx \geq e^{-a^2} 2a$ to conclude

$$\begin{aligned}
&\Pi(||\boldsymbol{u}_{n,h,k}|| \leq \gamma_{n,h}|\xi) \geq \prod_{r=1}^{R_n} \Pi(|u_{n,h,k}^{(r)}| \leq \frac{\gamma_{n,h}}{R_n}|\xi) = \prod_{r=1}^{R_n} \left((1-\xi) + \left(\frac{\xi}{\sqrt{2\pi}}\right) \int_{-\gamma_{n,h}/R_n}^{\gamma_{n,h}/R_n} e^{-x^2/2}\right) \\
&\geq \prod_{r=1}^{R_n} \left((1-\xi) + \left(\frac{\xi}{\sqrt{2\pi}}\right) e^{-\gamma_{n,h}^2/R_n^2} \frac{2\gamma_{n,h}}{R_n}\right) \geq \left[(1-\xi) + \frac{\xi}{\sqrt{2\pi}} e^{-\gamma_{n,h}^2/R_n^2} \frac{2\gamma_{n,h}}{R_n}\right]^{R_n}.
\end{aligned}$$

$$\prod_{k=1}^{p_n} \Pi(||\boldsymbol{u}_{n,h,k}|| \leq \gamma_{n,h}) \geq E\left[(1-\xi) + \frac{\xi}{\sqrt{2\pi}} \exp\left(-\frac{\gamma_{n,h}^2}{R_n^2}\right) \frac{2\gamma_{n,h}}{R_n}\right]^{R_n p_n}$$

$$= E\left[\sum_{h_1=0}^{R_n p_n} \binom{R_n p_n}{h_1} (1-\xi)^{h_1} \left(\frac{\xi}{\sqrt{2\pi}}\right)^{R_n p_n - h_1} \left(\frac{2\gamma_{n,h}}{R_n}\right)^{R_n p_n - h_1} \exp(-\frac{(R_n p_n - h_1)\gamma_{n,h}^2}{R_n^2})\right]$$

$$\geq \left(\frac{1}{\sqrt{2\pi}}\right)^{R_n p_n} \sum_{h_1=0}^{R_n p_n} \binom{R_n p_n}{h_1} Beta(R_n p_n - h_1 + 1, h_1 + 1) \left(\frac{2\gamma_{n,h}}{R_n}\right)^{R_n p_n - h_1} \exp\left(-\frac{(R_n p_n - h_1)\gamma_{n,h}^2}{R_n^2}\right)$$

$$\geq \left(\frac{1}{\sqrt{2\pi}}\right)^{R_n p_n} \sum_{h_1=0}^{R_n p_n} \frac{(R_n p_n)!}{h_1!(R_n p_n - h_1)!} \frac{h_1!(R_n p_n - h_1)!}{(R_n p_n + 1)!} \left(\frac{2\gamma_{n,h}}{R_n}\right)^{R_n p_n - h_1} \exp\left(-\frac{(R_n p_n - h_1)\gamma_{n,h}^2}{R_n^2}\right)$$

$$\geq \left(\frac{1}{\sqrt{2\pi}}\right)^{R_n p_n} \frac{R_n p_n}{R_n p_n + 1} \left(\frac{2\gamma_{n,h}}{R_n}\right)^{R_n p_n} \exp\left(-\frac{p_n \gamma_{n,h}^2}{R_n}\right).$$

Thus,

$$\Pi(||\boldsymbol{B}_{n,h} - \boldsymbol{B}_{T,n,h}||_\infty \leq \delta_n) \geq \exp\left(-\frac{\sum_{k=1}^{p_n} ||\tilde{\boldsymbol{u}}_{T,n,h,k}||^2}{2}\right) \left(\frac{1}{\sqrt{2\pi}}\right)^{R_n p_n} \frac{R_n p_n}{R_n p_n + 1} \left(\frac{2\gamma_{n,h}}{R_n}\right)^{R_n p_n} \exp\left(-\frac{p_n \gamma_{n,h}^2}{R_n}\right)$$

Lemma A.3 *Let x^* be a real positive root of the equation $P(x) = x^D + a_{D-1}x^{D-1} + \cdots + a_1 x - a_0 = 0$ with $a_0 > 0$, $a_1, ..., a_{D-1} > 0$. Then $\frac{1}{x^*} \leq 1 + \frac{a_1}{a_0}$.*

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

Proof of Theorem 3.1

Proof Define,

$$\mathcal{D}_0(f, f_T) = \int \int f_T(\boldsymbol{Y}|x) \log(f_T(\boldsymbol{Y}|x)/f(\boldsymbol{Y}|x)) \nu_{\boldsymbol{Y}}(d\boldsymbol{Y}) \nu_x(dx),$$

where f and f_T are as defined in (6). Let $\mathcal{M}_n$ be the sequence of sets of probability densities and $\mathcal{F}_n(\epsilon_n, \mathcal{M}_n)$ be the minimum number of Hellinger balls of radius ϵ_n needed to cover $\mathcal{M}_n$. Invoking Proposition 1 in Jiang *et al.* (2007), it suffices to show that the following conditions hold for sufficiently large n to prove Theorem 3.1:
(a) $\log \mathcal{F}_n(\epsilon_n, \mathcal{M}_n) \leq n\epsilon_n^2$, (b) $\Pi(\mathcal{M}_n^c) \leq e^{-2n\epsilon_n^2}$, (c) For small enough $r_1, r_2 > 0$, $\exists\, N_{r_1,r_2}$ such that for all $n \geq N_{r_1,r_2}$, $\Pi[f : \mathcal{D}_0(f, f_T) \leq r_1\epsilon_n^2/4] \geq e^{-r_2 n\epsilon_n^2}$.

Proof of condition (b): Define a sieve of probability densities $\mathcal{M}_n$, given by

$$\mathcal{M}_n = \left\{ f^{(H_n)} = \sum_{h=1}^{H_n} \omega_h f(\boldsymbol{Y}|x, \boldsymbol{B}_{n,h}),\ \boldsymbol{B}_{n,h} = \sum_{r=1}^{R_n} \boldsymbol{u}_{n,h}^{(r)} \circ \boldsymbol{u}_{n,h}^{(r)},\ |u_{n,h,s}^{(r)}| \leq C_n,\ r = 1, ..., R_n,\ s = 1, .., p_n \right\}, \tag{12}$$

where $\sum_{h=1}^{H_n} \omega_h = 1$, $H_n \to \infty$, $n \to \infty$. Then for all large n,

$$\begin{aligned}\Pi(\mathcal{M}_n^c) &= \Pi(\cup_{h=1}^{H_n} \cup_{s=1}^{p_n} \cup_{r=1}^{R_n} \{|u_{n,h,s}^{(r)}| > C_n\}) \leq H_n R_n p_n \Pi(|u_{n,h,s}^{(r)}| > C_n) = 2H_n R_n p_n (1 - \Phi(C_n)) \\ &\leq e^{-2n\epsilon_n^2},\end{aligned}$$

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

Proof of condition (a): Define,

$$\mathcal{M}_{n,h} = \left\{ f(\boldsymbol{Y}|x, \boldsymbol{B}_{n,h}) :\ \boldsymbol{B}_{n,h} = \sum_{r=1}^{R_n} \boldsymbol{u}_{n,h}^{(r)} \circ \boldsymbol{u}_{n,h}^{(r)},\ |u_{n,h,s}^{(r)}| \leq C_n,\ r = 1, ..., R_n,\ s = 1, .., p_n \right\},$$

for $h = 1, ..., H_n$. By Theorem 2 of Genovese *et al.* (2000),

$$\mathcal{F}_n(\epsilon_n, \mathcal{M}_n) \leq H_n (2\pi e)^{H_n/2} (3/\epsilon_n)^{H_n - 1} \prod_{h=1}^{H_n} \mathcal{F}_n(\epsilon_n/3, \mathcal{M}_{n,h}).$$

Let us consider balls of the form $(u_{n,h,s}^{(r)} - \rho, u_{n,h,s}^{(r)} + \rho)_{s,r=1}^{p_n, R_n}$ with their centers $|u_{n,h,s}^{(r)}| \leq C_n$, i.e., the densities f defined through parameters $u_{n,h,s}^{(r)}$'s belonging to $\mathcal{M}_{n,h}$. There are at most $F(\rho) = (C_n/\rho + 1)^{R_n p_n}$ such balls needed to cover the parameter space $\{u_{n,h,s}^{(r)} : s = 1, .., p_n; r = 1, .., R_n, |u_{n,h,s}^{(r)}| \leq C_n\}$.

Let $\tilde{f}$ be any density in $\mathcal{M}_{n,h}$, where $\tilde{B}_{n,h,\boldsymbol{j}} = \sum_{r=1}^{R_n} v_{n,h,j_1}^{(r)} v_{n,h,j_2}^{(r)}$, with $|v_{n,h,s}^{(r)}| \leq C_n$ for all $h = 1, ..., H_n$, $r = 1, .., R_n$. There exists a density $f \in \mathcal{M}_{n,h}$ represented by parameters

$u^{(r)}_{n,h,s}$'s such that $v^{(r)}_{n,h,s} \in (u^{(r)}_{n,h,s} - \rho, u^{(r)}_{n,h,s} + \rho)$ for every r, s and h. Note that,

$$\mathcal{D}_H(f, \tilde{f}) \leq \left\{\mathcal{D}_0(f, \tilde{f})\right\}^{1/2} = \left\{\sum_{\boldsymbol{j}\in\mathcal{J}} \mathcal{D}_0(f_{\boldsymbol{j}}, \tilde{f}_{\boldsymbol{j}})\right\}^{1/2} = \left\{\sum_{\boldsymbol{j}\in\mathcal{J}} (\alpha_{n,h,\boldsymbol{j}} - \tilde{\alpha}_{n,h,\boldsymbol{j}})^2/2\right\}^{1/2}$$
$$\leq \left\{\sum_{\boldsymbol{j}\in\mathcal{J}} (B_{n,h,\boldsymbol{j}} - \tilde{B}_{n,h,\boldsymbol{j}})^2/2\right\}^{1/2},$$

where $\alpha_{n,h,\boldsymbol{j}} = xB_{n,h,\boldsymbol{j}}$ and $\tilde{\alpha}_{n,h,\boldsymbol{j}} = x\tilde{B}_{n,h,\boldsymbol{j}}$. Now note that,

$$|B_{n,h,\boldsymbol{j}} - \tilde{B}_{n,h,\boldsymbol{j}}| = |\sum_{r=1}^{R_n} u^{(r)}_{n,h,j_1} u^{(r)}_{n,h,j_2} - \sum_{r=1}^{R_n} v^{(r)}_{n,h,j_1} v^{(r)}_{n,h,j_2}|$$
$$\leq \sum_{r=1}^{R_n} \left\{|u^{(r)}_{n,h,j_1} - v^{(r)}_{n,h,j_1}||u^{(r)}_{n,h,j_1}| + |v^{(r)}_{n,h,j_1}||u^{(r)}_{n,h,j_2} - v^{(r)}_{n,h,j_2}|\right\} \leq 2R_n\rho C_n.$$

Hence,

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

Choosing $\rho = \epsilon_n/(3q_n^{1/2}R_nC_n)$, one gets $\mathcal{D}_H(f, \tilde{f}) \leq \epsilon_n/3$. Hence

$$\log \mathcal{F}_n(\epsilon_n, \mathcal{M}_n) \leq \log(H_n) + H_n \log(2\pi e)/2 + (H_n - 1)\log(3/\epsilon_n) + \sum_{h=1}^{H_n} \log \mathcal{F}_n(\epsilon_n/3, \mathcal{M}_{n,h})$$
$$\leq \log(H_n) + H_n \log(2\pi e)/2 + (H_n - 1)\log(3/\epsilon_n) + H_n \log \mathcal{F}(\rho)$$
$$\leq \log(H_n) + H_n \log(2\pi e)/2 + (H_n - 1)\log(3/\epsilon_n) + H_n R_n p_n \log\left(1 + 3q_n^{1/2}R_nC_n^2/\epsilon_n^2\right)$$
$$\leq \log(H_n) + H_n \log(2\pi e)/2 + (H_n - 1)\log(3/\epsilon_n) + H_n R_n p_n \log(6q_n^{1/2}/\epsilon_n^2) + H_n R_n p_n \log(R_nC_n^2)$$
$$\leq \log(H_n) + H_n \log(2\pi e)/2 + (H_n - 1)\log(3/\epsilon_n) + H_n R_n p_n \log(6p_n) + H_n R_n p_n \log(1/\epsilon_n^2) +$$
$$H_n R_n p_n \log(R_nC_n^2)$$
$$\leq n\epsilon_n^2, \text{ for large } n, \text{ by assumptions (i), (ii) and (iv).}$$

Proof of condition (c): Since f and f_T have the same number of mixture components, by the chain rule of entropy, $\mathcal{D}_0(f, f_T) \leq \sum_{h=1}^{H_n} \{\omega_{h,T}\mathcal{D}_0(f(\cdot|x, \boldsymbol{B}_{n,h}), f_T(\cdot|x, \boldsymbol{B}_{T,n,h})) + \mathcal{D}_0(\boldsymbol{\omega}, \boldsymbol{\omega}_T)\}$.

Note that $\mathcal{D}_0(f(\cdot|x,\boldsymbol{B}_{n,h}),f_T(\cdot|x,\boldsymbol{B}_{T,n,h}))=\sum_{\boldsymbol{j}\in\mathcal{J}}E_{\boldsymbol{x}}\left[\mathcal{D}_0(f(\cdot|x,\boldsymbol{B}_{n,h,\boldsymbol{j}}),f_T(\cdot|x,\boldsymbol{B}_{T,n,h,\boldsymbol{j}}))\right]\leq \sum_{\boldsymbol{j}\in\mathcal{J}}(B_{n,h,\boldsymbol{j}}-B_{T,n,h,\boldsymbol{j}})^2/2$, where the last inequality follows by considering that both $f(\cdot|x,\boldsymbol{B}_{n,h,\boldsymbol{j}})$ and $f_T(\cdot|x,\boldsymbol{B}_{n,h,\boldsymbol{j}})$ are Gaussian densities. Let $\delta_n=\epsilon_n/(2\sqrt{q_n})$, and define

$$\mathcal{U}_{n,1}=\cap_{h=1}^{H_n}\left\{\boldsymbol{B}_{n,h}:B_{n,h,\boldsymbol{j}}\in(B_{T,n,h,\boldsymbol{j}}-\delta_n,B_{T,n,h,\boldsymbol{j}}+\delta_n),\forall\,\boldsymbol{j}\in\mathcal{J}\right\}$$
$$\mathcal{U}_{n,2}=\left\{(\omega_1,...,\omega_{H_n}):\sum_{h=1}^{H_n}|\omega_h-\omega_{h,T}|\leq\epsilon_n/2\right\},\ \mathcal{U}_n=\mathcal{U}_{n,1}\cap\mathcal{U}_{n,2}. \tag{13}$$

Under $\mathcal{U}_n$, $\mathcal{D}_0(f,f_T)\leq\epsilon_n^2/4$ for all large n.

Now $\Pi(\{f:\mathcal{D}_0(f,f_T)\leq\epsilon_n^2/4\})\geq\Pi(\mathcal{U}_n)=\Pi(\mathcal{U}_{n,1})\Pi(\mathcal{U}_{n,2})$. By Lemma A.2, $-\log\Pi(\mathcal{U}_{n,1})=-\log\Pi(\{\boldsymbol{B}_{n,h}:B_{n,h,\boldsymbol{j}}\in(B_{T,n,h,\boldsymbol{j}}-\delta_n,B_{T,n,h,\boldsymbol{j}}+\delta_n),\forall\boldsymbol{j}\in\mathcal{J}\})=-\log\Pi(||\boldsymbol{B}_{n,h}-\boldsymbol{B}_{T,n,h}||_\infty\leq\delta_n,h=1,...,H_n)\leq\sum_{h=1}^{H_n}\sum_{k=1}^{p_n}||\boldsymbol{u}_{T,n,h,k}||^2/2+(R_nH_np_n/2)\log(2\pi)+H_n\log(1+(1/(R_np_n)))+H_nR_np_n\log(R_n)+\sum_{h=1}^{H_n}R_np_n\log(1/\gamma_{n,h})+\sum_{h=1}^{H_n}p_n\gamma_{n,h}^2/R_n$.

Since $||\boldsymbol{u}_{T,n,h,k}||\geq 0$, $\sum_{k=1}^{p_n}||\boldsymbol{u}_{T,n,h,k}||^2\leq(\sum_{k=1}^{p_n}||\boldsymbol{u}_{T,n,h,k}||)^2$ is bounded for large n, by assumption (v). By assumption (i), $H_nR_np_n\log(R_n)\prec n\epsilon_n^2$ (hence $H_nR_np_n\prec n\epsilon_n^2$). Notet that $\gamma_{n,h,\boldsymbol{j}}=\frac{-(||\boldsymbol{u}_{T,n,h,j_1}||+||\boldsymbol{u}_{T,n,h,j_1}||)+\sqrt{(||\boldsymbol{u}_{T,n,h,j_1}||+||\boldsymbol{u}_{T,n,h,j_1}||)^2+4\delta_n}}{2}\leq\sqrt{\delta_n}$, since $\delta_n>0$. This implies $\sum_{h=1}^{H_n}p_n\gamma_{n,h}^2/R_n\prec n\epsilon_n^2$, for all large n, by assumption (i). Using Lemma A.1 and A.3, $1/\gamma_{n,h}\leq(\sum_{k=1}^{p_n}||\boldsymbol{u}_{T,n,h,k}||)^2/\delta_n+1$. If $m_0=\limsup_{n\to\infty}\sum_{k=1}^{p_n}||\boldsymbol{u}_{T,n,h,k}||$, then $\sum_{h=1}^{H_n}R_np_n\log(1/\gamma_{n,h})\leq R_np_nH_n\log(m_0^2/\delta_n)=2R_np_nH_n\log(m_0)+\frac{R_np_nH_n}{2}\log(q_n)+\frac{R_np_nH_n}{2}\log(1/\epsilon_n^2)\leq 2R_np_nH_n\log(m_0)+R_np_nH_n\log(p_n)+\frac{R_np_nH_n}{2}\log(1/\epsilon_n^2)\prec n\epsilon_n^2$, by assumptions (i) and (ii). Also, $-\log(\Pi(\sum_{h=1}^{H_n}|\omega_h-\omega_{h,T}|\leq\epsilon_n/2))\prec H_n\log(2/\epsilon_n)\prec n\epsilon_n^2$, by Lemma A.2 of **?**.

All the aforementioned calculations yield $-\log\Pi(||\boldsymbol{B}_{n,h}-\boldsymbol{B}_{T,n,h}||_\infty\leq\delta_n)\leq r_2n\epsilon_n^2/4$, for any $r_2>0$ and all large n, which implies $\Pi(\{f:\mathcal{D}_0(f,f_T)\leq\epsilon_n^2/4\})\geq e^{-r_2n\epsilon_n^2/4}$ for all large n. This concludes the proof.

B 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:

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

- $\gamma^*_{s,h}|- \sim N\left(\frac{\sum_{i\in\mathcal{I}_h}z_{is}^2\mathbf{1}^T(\boldsymbol{y}_i-\sum_{h_2=1}^m\boldsymbol{\beta}^*_{h_2,h}x_{ih_2}-\mathbf{1}\sum_{h_2=1,h_2\neq s}^l\gamma^*_{h_2,h}z_{ih_2})/\sigma^2+a_\beta/b_\beta}{q\sum_{i\in\mathcal{I}_h}z_{is}^2/\sigma^2+1/b_\beta}, \frac{1}{q\sum_{i\in\mathcal{I}_h}z_{is}^2/\sigma^2+1/b_\beta}\right)$, $s=1,...,l$; $h=1,...,H$.

- $\sigma^2|- \sim IG(a_\sigma+(nq)/2, b_\sigma+\sum_{h=1}^H\sum_{i\in\mathcal{I}_h}||\boldsymbol{y}_i-\sum_{s=1}^m\boldsymbol{\beta}^*_{s,h}x_{is}-\mathbf{1}\sum_{s=1}^l\gamma^*_{s,h}z_{is}||^2/2)$

- $\boldsymbol{M}_{s,h}|- \sim IW\left[(\boldsymbol{S}+\sum_{k:\boldsymbol{u}_{s,h,k}\neq\mathbf{0}}\boldsymbol{u}_{s,h,k}\boldsymbol{u}_{s,h,k}{}^T), (\nu+\{\#k:\boldsymbol{u}_{s,h,k}\neq\mathbf{0}\})\right]$

- $\pi_{s,h,r}|- \sim Beta[(1+\lambda_{s,h,r}), (r^\eta+1-\lambda_{s,h,r})]$

- $\lambda_{s,h,r}|- \sim Ber(p_{s,h,r})$, 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 $J(\boldsymbol{\Lambda}_{s,h})=\prod_{i\in\mathcal{I}_h}N(\boldsymbol{y}_i|\gamma^*_{0,h}\mathbf{1}+\sum_{s=1}^m\boldsymbol{\beta}^*_{s,h}x_{is}+\mathbf{1}\sum_{s=1}^l\gamma^*_{s,h}z_{is},\sigma^2I)$. $J(\boldsymbol{\Lambda}_{s,h})_{(\lambda_{s,h,r}=1)}$ 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,

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

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

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

- $P(c_i=h\,|-)=\frac{\omega_hN(\boldsymbol{y}_i|\gamma^*_{0,h}\mathbf{1}+\sum_{s=1}^m\boldsymbol{\beta}^*_{s,h}x_{is}+\mathbf{1}\sum_{s=1}^l\gamma^*_{s,h}z_{is},\sigma^2I)}{\sum_{d'=1}^H\omega_{d'}N(\boldsymbol{y}_i|\gamma^*_{0,d'}\mathbf{1}+\sum_{s=1}^m\boldsymbol{\beta}^*_{s,d'}x_{is}+\mathbf{1}\sum_{s=1}^l\gamma^*_{s,d'}z_{is},\sigma^2I)}$, 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.

Acknowledgements

The second author was supported by NSF Grant DMS-1854662 and ONR Grant BAA-N00014-18-1-2741.

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.

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

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.

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

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.

Choi, T. (2008). Convergence of posterior distribution in the mixture of regressions. *Journal of Nonparametric Statistics*, **20**(4), 337–351.

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.

DeYoreo, M. and Kottas, A. (2018). Bayesian nonparametric modeling for multivariate ordinal regression. *Journal of Computational and Graphical Statistics*, **27**(1), 71–84.

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

Dunson, D. B., Pillai, N., and Park, J.-H. (2007). Bayesian density regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**(2), 163–183.

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.

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

Genovese, C. R., Wasserman, L., *et al.* (2000). Rates of convergence for the gaussian mixture sieve. *Annals of Statistics*, **28**(4), 1105–1127.

Ghosal, S., Van Der Vaart, A., *et al.* (2007). Posterior convergence rates of dirichlet mixtures at smooth densities. *The Annals of Statistics*, **35**(2), 697–723.

Guha, S. and Guhaniyogi, R. (2020). Bayesian generalized sparse symmetric tensor-on-vector regression. *Technometrics*, pages 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. (2017). Convergence rate of bayesian supervised tensor modeling with multiway shrinkage priors. *Journal of Multivariate Analysis*, **160**, 157–168.

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

Guhaniyogi, R., Qamar, S., and Dunson, D. B. (2018). Bayesian conditional density filtering. *Journal of Computational and Graphical Statistics*, **27**(3), 657–672.

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. (2009). A hierarchical eigenmodel for pooled covariance estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**(5), 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**(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.

Jeong, S. and Ghosal, S. (2020). Posterior contraction in sparse generalized linear models. *Biometrika*.

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

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.

Kiar, G., Gorgolewski, K., and Kleissas, D. (2017a). Example use case of sic with the ndmg pipeline (sic: ndmg). *GigaScience Database*.

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.* (2017b). Science in the cloud (sic): A use case in MRI connectomics. *Giga Science*, **6**(5), 1–10.

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).

Spencer, D., Guhaniyogi, R., and Prado, R. (2020). Joint bayesian estimation of voxel activation and inter-regional connectivity in fmri experiments. *Psychometrika*, pages 1–25.

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.

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

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.