Bayesian Regression with Undirected Network Predictors with an Application to Brain Connectome Data

Sharmistha Guha and Abel Rodriguez

January 24, 2018

Abstract

This article proposes a Bayesian approach to regression with a continuous scalar response and an undirected network predictor. Undirected network predictors are often expressed in terms of symmetric adjacency matrices, with rows and columns of the matrix representing the nodes, and zero entries signifying no association between two corresponding nodes. Network predictor matrices are typically vectorized prior to any analysis, thus failing to account for the important structural information in the network. This results in poor inferential and predictive performance in presence of small sample sizes. We propose a novel class of *network shrinkage priors* for the coefficient corresponding to the undirected network predictor. The proposed framework is devised to detect both nodes and edges in the network predictive of the response. Our framework is implemented using an efficient Markov Chain Monte Carlo algorithm. Empirical results in simulation studies illustrate strikingly superior inferential and predictive gains of the proposed framework in comparison with the ordinary high dimensional Bayesian shrinkage priors and penalized optimization schemes. We apply our method to a brain connectome dataset that contains information on brain networks along with a measure of creativity for multiple individuals. Here, interest lies in building a regression model of the creativity measure on the network predictor to identify important regions and connections in the brain strongly associated with creativity. To the best of our knowledge, our approach is the first principled Bayesian method that is able to detect scientifically interpretable regions and connections in the brain actively impacting the continuous response (creativity) in the presence of a small sample size.

Keywords: Brain connectome; Edge selection; High dimensional regression; Network predictors; Network shrinkage prior; Node selection.

1 Introduction

In recent years, network data has become ubiquitous in disciplines as diverse as neuroscience, genetics, finance and economics. Nonetheless, statistical models that involve network data are particularly challenging, not only because they require dimensionality reduction procedures to effectively deal with the large number of pairwise relationships, but also because flexible formulations are needed to account for the topological structure of the network.

The literature has paid heavy attention to models that aim to understand the relationship between node-level covariates and the structure of the network. A number of classic models treat the dyadic observations as the response variable, examples include random graph models (Erdos and Rényi, 1960), exponential random graph models (Frank and Strauss, 1986), social space models Hoff *et al.* (2002); Hoff (2005, 2009) and stochastic block models (Nowicki and Snijders, 2001). The goal of these models is either link prediction or to investigate *homophily*, i.e., the process of formation of social ties due to matching individual traits. Alternatively, models that investigate *influence* or *contagion* attempt to explain the node-specific covariates as a function of the network structure (e.g., see Christakis and Fowler, 2007; Fowler and Christakis, 2008; Shoham *et al.*, 2015 and references therein). Common methodological approaches in this context include simultaneous autoregressive (SAR) models (e.g., see Lin, 2010) and threshold models (e.g., see Watts and Dodds, 2009). However, ascertaining the direction of a causal relationship between network structure and link or nodal attributes, i.e. whether it pertains to homophily or contagion, is difficult (e.g., see Doreian, 2001 and Shalizi and Thomas, 2011 and references therein). Hence there has been a growing interest in joint models for the coevolution of the network structure and nodal attributes (e.g., see Fosdick and Hoff, 2015; Durante *et al.*, 2017; De la Haye *et al.*, 2010; Niezink and Snijders, 2016; Guhaniyogi and Rodriguez, 2017).

In this paper we investigate Bayesian models for network regression. Unlike the problems discussed above, in network regression we are interested in the relationship between the structure of the network and one or more global attributes of the experimental unit on which the network data is collected. As a motivating example, we consider the problem of predicting the composite creativity index of individuals on the basis of neuroimaging data reassuring the connectivity of different brain regions. The goal of these studies is twofold. First neuroscientists are interested in identifying regions of the brain that are involved in creative thinking. Secondly, it is important to determine which how the strength of connection among these influential regions affects the level of creativity of the individual. More specifically, we construct a novel Bayesian *network shrinkage prior* that combines ideas from spectral decomposition methods and spike-and-slab priors to generate a model that respects the structure of the predictors. The model produces accurate predictions, allows us to identify both modes and links that have influence on the response, and yield

well-calibrated interval estimates for the model parameters.

A common approach to network regression is to use a few summary measures from the network in the context of a flexible regression or classification approach (see, for example, Bullmore and Sporns, 2009 and references therein). Clearly, the success of this approach is highly dependent on selecting the right summaries to include. Furthermore, it cannot identify the impact of specific nodes on the response, which is of clear interest in our setting. Alternatively, a number of authors have proceeded to vectorize the network predictor (originally obtained in the form of a symmetric matrix). Subsequently, the continuous response would be regressed on the high dimensional collection of edge weights (e.g., see Richiardi *et al.*, 2011 and Craddock *et al.*, 2009). This approach can take advantage of the recent developments in high dimensional regression, consisting of both penalized optimization (Tibshirani, 1996), and Bayesian shrinkage (Park and Casella, 2008; Carvalho *et al.*, 2010; Armagan *et al.*, 2013) perspectives. However, this treat the links of the network as if they were exchangeable, ignoring the fact that coefficients that involving common nodes can be expect to be correlated a priori . Ignoring this correlation often leads to poor predictive performance and can potentially impact model selection.

Recently, Relión *et al.* (2017) proposed a penalized optimization scheme that not only enables classification of networks, but also identifies important nodes and edges. Although this model seems to perform well for prediction problem, uncertainty quantification is difficult because standard bootstrap methods are not consistent for Lasso-type methods (e.g., see Kyung *et al.*, 2010 and Chatterjee and Lahiri, 2010). Modifications of the bootstrap that produce well-calibrated confidence intervals in the context of standard Lasso regression have been proposed (e.g., see Chatterjee and Lahiri, 2011), but it is not clear whether they extend to the more complex penalties introduced in Relión *et al.* (2017). Recent developments on tensor regression (e.g., see Zhou *et al.*, 2013; Guhaniyogi *et al.*, 2017) are also relevant to our work. However, these approaches tend to focus mainly on prediction and are not designed to detect important nodes impacting the response.

The rest of the article evolves as follows. Section 2 proposes the novel network shrinkage prior and discusses posterior computation for the proposed model. Empirical investigations with various simulation studies are presented in Section 3 while Section 4 analyzes the brain connectome dataset. We provide results on *region of interest* (ROI) and *edge* selection and find them to be scientifically consistent with previous studies. Finally, Section 5 concludes the article with an eye towards future work.

2 Model Formulation

2.1 Notations

Let $y_i \in \mathbb{R}$ and $\boldsymbol{A}_i$ represent the observed scalar response and the corresponding weighted undirected network for the ith sample, $i = 1, ..., n$ respectively. The weights corresponding to the edges belong to $\mathbb{R}$ and all graphs share the same labels on their nodes. For example, in our brain connectome application discussed subsequently, y_i corresponds to a phenotype, while $\boldsymbol{A}_i$ encodes the network connections between different regions of the brain for the ith individual. Let the network corresponding to any individual consist of V nodes. Mathematically, it amounts to $\boldsymbol{A}_i$ being a $V \times V$ matrix, with the (k, l)th entry of $\boldsymbol{A}_i$ denoted by $a_{i,k,l} \in \mathbb{R}$. We focus on networks that contain no self relationship, i.e. $a_{i,k,k} \equiv 0$, and are undirected ($a_{i,k,l} = a_{i,l,k}$). The brain connectome application considered here naturally justifies these assumptions. Although we present our model specific to these settings, it will be evident that the proposed model can be easily extended to directed networks with self-relations. Throughout this article, we denote the Frobenius inner product between two $V \times V$ matrices $\boldsymbol{C}$ and $\boldsymbol{D}$ by $\langle \boldsymbol{C}, \boldsymbol{D} \rangle_F = Trace(\boldsymbol{D}'\boldsymbol{C})$. Frobenius inner product is the natural inner product on the space of matrices and is a generalization of the dot product from vector to matrix spaces. The square root of the Frobenius inner product of a matrix with itself yields the Frobenius norm. Mathematically, it is denoted by $||\boldsymbol{C}||_F = \sqrt{\langle \boldsymbol{C}, \boldsymbol{C} \rangle_F}$. Additionally, we define $||\boldsymbol{B}||_1$ as the sum of the absolute values of all the elements of matrix $\boldsymbol{B}$. Also, define $|| \cdot ||_2$ as the l_2 norm of a vector and $\boldsymbol{B}_{(k)}$ as the kth row of $\boldsymbol{B}$.

2.2 Bayesian Network Regression Model

We propose the high dimensional regression model of the response y_i for the i-th individual on the undirected network predictor $\boldsymbol{A}_i$ as

$$y_i = \mu + \langle \boldsymbol{A}_i, \boldsymbol{B} \rangle_F + \epsilon_i, \;\; \epsilon_i \overset{iid}{\sim} N(0, \tau^2), \tag{1}$$

where $\boldsymbol{B}$ is the network coefficient matrix of dimension $V \times V$ whose (k, l)th element is given by $\beta_{k,l}$. Similar to the network predictor, the network coefficient matrix $\boldsymbol{B}$ is also assumed to be symmetric with zero diagonal entries. τ^2 is the variance of the idiosyncratic error.

Since self relationship is absent and $\boldsymbol{A}_i$ is symmetric, $\langle \boldsymbol{A}_i, \boldsymbol{B} \rangle_F = 2 \sum_{1 \le k < l \le V} a_{i,k,l} \beta_{k,l}$. Then, denoting $\gamma_{k,l} = 2\beta_{k,l}$, (1) can be rewritten as

$$y_i = \mu + \sum_{1 \le k < l \le V} a_{i,k,l} \gamma_{k,l} + \epsilon_i, \; \epsilon_i \sim N(0, \tau^2), \tag{2}$$

Equation (2) connects the network regression model with the linear regression framework with $a_{i,k,l}$'s as predictors and $\gamma_{k,l}$'s as the corresponding coefficients.

While in ordinary linear regression, the predictor coefficients are indexed by the natural numbers $\mathcal{N}$, Model (2) indexes the predictor coefficients by their positions in the matrix $\boldsymbol{B}$. This is done in order to keep a tab not only on the edge itself but also on the nodes connecting the edges. As mentioned earlier, we are interested in identifying nodes and edges which contribute to the regression. Additionally, we would also like to accurately estimate the coefficients $\gamma_{k,l}$ and subsequently make accurate predictions. The next section describes a novel network shrinkage prior to achieve these goals.

2.3 Developing the Network Shrinkage Prior

2.3.1 Vector Shrinkage Prior

High dimensional regression with vector predictors has recently been of interest in Bayesian statistics. An overwhelming literature in Bayesian statistics in the last decade has focused on shrinkage priors which shrink coefficients corresponding to unimportant variables to zero while minimizing the shrinkage of coefficients corresponding to influential variables. Many of these shrinkage prior distributions can be expressed as a scale mixture of normal distributions, commonly referred to as *global-local (GL) scale mixtures* (Polson and Scott (2010)), that enable fast computation employing simple conjugate Gibbs sampling. More precisely, in the context of model (2), a global-local scale mixture prior would take the form

$$\gamma_{k,l} \sim N(0, s_{k,l}\tau^2), \;\; s_{k,l} \sim g_1, \;\; \tau^2 \sim g_2, \quad 1 \leq k < l \leq V,$$

where $(\gamma_{1,2}, ..., \gamma_{V-1,V})$ are the regression coefficients in (2) with $q = V(V-1)/2$ predictors. Note that $s_{1,2}, ..., s_{V-1,V}$ are local scale parameters controlling the shrinkage of the coefficients, while τ^2 is the global scale parameter. Different choices of g_1 and g_2 lead to different classes of Bayesian shrinkage priors which have appeared in the literature. For example, the Bayesian Lasso (Park and Casella (2008)) prior takes g_1 as exponential and g_2 as the Jeffreys prior, the Horseshoe prior (Carvalho *et al.* (2010)) takes both g_1 and g_2 as half-Cauchy distributions and the Generalized Double Pareto Shrinkage prior (Armagan *et al.* (2013)) takes g_1 as exponential and g_2 as the Gamma distribution.

The direct application of this global-local prior in the context of (2) is unappealing. To elucidate further, note that if node k contributes minimally to the response, one would expect to have smaller estimates for all coefficients $\gamma_{k,l}$, $l > k$ and $\gamma_{l',k}$, $l' < k$ corresponding to edges connected to node k. The ordinary GL shrinkage prior distribution given as above does not necessarily conform to such an important restriction. In what follows, we build a network

shrinkage prior upon the existing literature that respects this constraint as elucidated in the next section.

2.3.2 Network Shrinkage Prior

We propose a shrinkage prior on the coefficients $\gamma_{k,l}$ and refer to it as the *Bayesian Network Shrinkage prior*. The prior borrows ideas from low-order spectral representations of matrices. Let $\boldsymbol{u}_1, ..., \boldsymbol{u}_V \in \mathbb{R}^R$ be a collection of R-dimensional latent variables, one for each node, such that $\boldsymbol{u}_k$ corresponds to node k. We draw each $\gamma_{k,l}$ conditionally independent from a density that can be represented as a location and scale mixture of normals. More precisely,

$$\gamma_{k,l}|s_{k,l}, \boldsymbol{u}_k, \boldsymbol{u}_l, \tau^2 \sim N(\boldsymbol{u}_k'\boldsymbol{\Lambda}\boldsymbol{u}_l, \tau^2 s_{k,l}), \tag{3}$$

where $s_{k,l}$ is the scale parameter corresponding to each $\gamma_{k,l}$ and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_R)$ is an $R \times R$ diagonal matrix. $\lambda_r \in \{0, 1\}$'s are introduced to assess the effect of the rth dimension of $\boldsymbol{u}_k$ on the mean of $\gamma_{k,l}$. In particular, $\lambda_r = 0$ implies that the rth dimension of the latent variable $\boldsymbol{u}_k$ is not informative for any k. Note that if $s_{k,l} = 0$, this prior will imply that $\boldsymbol{\Gamma} = 2\boldsymbol{B} = \boldsymbol{U}'\boldsymbol{\Lambda}\boldsymbol{U}$, where $\boldsymbol{U}$ is an $R \times V$ matrix whose kth column corresponds to $\boldsymbol{u}_k$ and $\boldsymbol{\Gamma} = ((\gamma_{k,l}))_{k,l=1}^V$. Since $R << V$, the mean structure of $\boldsymbol{\Gamma}$ assumes a low-rank matrix decomposition.

In order to learn how many components of $\boldsymbol{u}_k$ are informative for (3), we assign a hierarchical prior

$$\lambda_r \sim Ber(\pi_r)\ ,\ \ \pi_r \sim Beta(1, r^\eta),\ \eta > 1.$$

The choice of hyper-parameters of the beta distribution is crucial in order to impart increasing shrinkage on λ_r as r grows. In particular, $E[\lambda_r] = 1/(1 + r^\eta) \to 0$, as $r \to \infty$, so that the prior favors choice of smaller number of active components in $\boldsymbol{u}_k$'s impacting the response. Additionally, the hyper-parameter of the distribution of λ_r safeguards the prior on λ_r from flattening out even at large r. In particular, $\sum_{r=1}^R E[\lambda_r] = \sum_{r=1}^R 1/(1 + r^\eta)$, so that $\sum_{r=1}^R E[\lambda_r]$ converges as $R \to \infty$. Note that $\sum_{r=1}^R \lambda_r$ is the number of dimensions of $\boldsymbol{u}_k$ contributing to predict the response. We refer to $\sum_{r=1}^R \lambda_r$ as R_{eff}, the *effective dimensionality* of the latent variables.

The mean structure of $\gamma_{k,l}$ is constructed to take into account the interaction between the kth and the lth nodes. In particular, the node specific latent variables $\boldsymbol{u}_1, ..., \boldsymbol{u}_V$ can be thought of as points in an R-dimensional sphere. Drawing intuition from Hoff (2005), we might imagine that the interaction between the kth and lth nodes has a positive, negative or neutral impact on the response depending on whether $\boldsymbol{u}_k$ and $\boldsymbol{u}_l$ are in the same direction, opposite direction or orthogonal to each other respectively. In other words, whether the

angle between $\boldsymbol{u}_k$ and $\boldsymbol{u}_l$ is acute, obtuse or right, i.e. $\boldsymbol{u}_k'\boldsymbol{\Lambda}\boldsymbol{u}_l > 0$, $\boldsymbol{u}_k'\boldsymbol{\Lambda}\boldsymbol{u}_l < 0$ or $\boldsymbol{u}_k'\boldsymbol{\Lambda}\boldsymbol{u}_l = 0$ respectively. The conditional mean of $\gamma_{k,l}$ in (3) is constructed to capture such latent network information in the predictor.

In order to directly infer on node selection (i.e. to determine if a node is inactive in explaining the response), we assign the *spike-and-slab* (Ishwaran and Rao (2005)) mixture distribution prior on the latent factor $\boldsymbol{u}_k$ as below

$$\boldsymbol{u}_k \sim \begin{cases} N(\boldsymbol{0}, \boldsymbol{M}), & \text{if } \xi_k = 1 \\ \delta_{\boldsymbol{0}}, & \text{if } \xi_k = 0 \end{cases}, \quad \xi_k \sim Ber(\Delta), \tag{4}$$

where $\delta_{\boldsymbol{0}}$ is the Dirac-delta function at $\boldsymbol{0}$ and $\boldsymbol{M}$ is a covariance matrix of order $R \times R$. The parameter Δ corresponds to the probability of the nonzero mixture component. Note that if the kth node of the network predictor is inactive in predicting the response, then a-posteriori ξ_k should provide high probability to 0. Thus, based on the posterior probability of ξ_k, it will be possible to identify unimportant nodes in the network regression. The rest of the hierarchy is accomplished by assigning prior distributions on the $s_{k,l}$'s, τ^2 and $\boldsymbol{M}$ as follows:

$$s_{k,l} \sim Exp(\lambda^2/2), \;\; \tau^2 \sim \pi(\tau^2) \propto \frac{1}{\tau^2}, \;\; \boldsymbol{M} \sim IW(\boldsymbol{S}, \nu),$$
$$\lambda^2 \sim Gamma(r_\lambda, \delta), \;\; \Delta \sim Beta(a_\Delta, b_\Delta),$$

where $\boldsymbol{S}$ is an $R \times R$ positive definite scale matrix. $IW(\boldsymbol{S}, \nu)$ denotes an Inverse-Wishart distribution with scale matrix $\boldsymbol{S}$ and degrees of freedom ν. Finally, we choose a non-informative flat prior on μ. Appendix B shows the propriety of the marginal posterior distributions for all parameters.

Note that, if $\boldsymbol{u}_k'\boldsymbol{\Lambda}\boldsymbol{u}_l = 0$, the marginal prior distribution of $\gamma_{k,l}$ integrating out all the latent variables turns out to be the double exponential distribution which is connected to the Bayesian Lasso prior. When $\boldsymbol{u}_k'\boldsymbol{\Lambda}\boldsymbol{u}_l = 0$, the marginal prior distribution of $\gamma_{k,l}$ appears to be a location mixture of double exponential prior distributions, the mixing distribution being a function of the network. Owing to this fact, we coin the proposed prior distribution as the *Bayesian Network Lasso* prior.

2.4 Posterior Computation

Although summaries of the posterior distribution cannot be computed in closed form, full conditional distributions for all parameters are available and correspond to standard families. Thus posterior computation of parameters proceed through Gibbs sampling. Details of all the full conditionals are presented in Appendix A.

In order to identify whether the kth node is important in terms of predicting the response,

we rely on the post burn-in L samples $\xi_k^{(1)}, ..., \xi_k^{(L)}$ of ξ_k. Node k is recognized to be influential if $\frac{1}{L}\sum_{l=1}^{L}\xi_k^{(l)} > 0.5$. Next, we turn our attention to inferring on the dimension of the latent variable $\boldsymbol{u}_k$. To this end, we empirically estimate the full posterior distribution of R_{eff}. In particular, an estimate of $P(R_{eff} = r \,|\, Data)$ is given by $\frac{1}{L}\sum_{l=1}^{L} I(\sum_{m=1}^{R}\lambda_m^{(l)} = r)$, where $I(A)$ for an event A is 1 if the event A happens and 0 otherwise and $\lambda_m^{(1)}, ..., \lambda_m^{(L)}$ are the L post burn-in MCMC samples of λ_m. The estimated posterior distribution of R_{eff} is presented in the form of a dot chart.

3 Simulation Studies

This section comprehensively contrasts both the inferential and the predictive performances of our proposed approach with a number of competitors in various simulation settings. We refer to our proposed approach as the *Bayesian Network Regression* (BNR). As competitors, we consider both frequentist penalized optimization methods as well as Bayesian shrinkage priors for high-dimensional regression.

First we use the generic variable selection and shrinkage methods that ignore the relational nature of the predictors. More specifically, we use Lasso (Tibshirani (1996)) as a popular penalized optimization scheme, and the Bayesian Lasso (Park and Casella (2008)) and Horseshoe priors (Carvalho *et al.* (2010)) as Bayesian shrinkage regression methods. The Horseshoe is considered to be the state-of-the-art Bayesian shrinkage prior and is known to perform well, both in sparse and not-so-sparse regression settings. In order to implement Lasso, Bayesian Lasso (henceforth interchangeably referred to as BLasso) and the Horseshoe (henceforth interchangeably referred to as BHS), we employ the `glmnet` (Friedman *et al.* (2010)) and `monomvn` (Gramacy and Gramacy (2013)) packages in `R`, respectively. All these methods treat edges between nodes as "bags of predictors" to run high dimensional regression, thereby ignoring the structure of the network predictor. A thorough comparison with these methods will indicate the relative advantage of exploiting the structure of the network predictor.

Additionally, we compare our method to a frequentist approach that develops network regression in the presence of a network predictor and scalar response (Relión *et al.* (2017)). To be precise, we adapt Relión *et al.* (2017) to a *continuous response* context and propose to estimate the network regression coefficient matrix $\boldsymbol{B}$ from

$$\hat{\boldsymbol{B}} = \arg \min_{\boldsymbol{B}\in\mathbb{R}, \boldsymbol{B}=\boldsymbol{B}', diag(\boldsymbol{B})=\boldsymbol{0}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu - \langle \boldsymbol{A}_i, \boldsymbol{B}\rangle_F)^2 + \frac{\varphi}{2}||\boldsymbol{B}||_F^2 + \varsigma\left(\sum_{k=1}^{V}||\boldsymbol{B}_{(k)}||_2 + \rho||\boldsymbol{B}||_1\right)\right\}, \tag{5}$$

where φ, ρ, ς are tuning parameters. The best possible choice of the tuning parameter triplet $(\varphi, \rho, \varsigma)$ is made using cross validation over a grid of possible values. Relión *et al.* (2017) argue that the penalty in (5) incorporates the network information of the predictor, thereby yielding superior inference to any ordinary penalized optimization scheme. Hence comparison with (5) will potentially highlight the advantages of a carefully structured Bayesian network shrinkage prior over the penalized optimization scheme incorporating network information. It is also worth mentioning that in the absence of any open source code, we implement the algorithm in Relión *et al.* (2017). All Bayesian competitors are allowed to draw $50,000$ MCMC samples out of which the first $30,000$ are discarded as burn-ins. All posterior inference is carried out based on the rest $20,000$ MCMC samples after suitably thinning the post burn-in chain.

3.1 Predictor and Response Data Generation

In all simulation studies, the undirected symmetric network predictor $\boldsymbol{A}_i$ for the ith sample is simulated by drawing $a_{i,k,l} \overset{iid}{\sim} N(0,1)$ for $k < l$ and setting $a_{i,k,l} = a_{i,l,k}, a_{i,k,k} = 0$ for all $k, l \in \{1, ..., V\}$. The response y_i is generated according to the network regression model

$$y_i = \mu_0 + \langle \boldsymbol{A}_i, \boldsymbol{B}_0 \rangle_F + \epsilon_i;\ \epsilon_i \sim N(0, \tau_0^2), \tag{6}$$

with τ_0^2 as the true noise variance. In a similar vein as section 2.2, define $\gamma_{k,l,0} = 2\beta_{k,l,0}$, which will later be used to define the mean squared error. In the interest of better presentation, all of our simulations use $V = 20$ nodes and $n = 70$ samples.

To study all competitors under various data generation schemes, we simulate the true network predictor coefficient $\boldsymbol{B}_0$ under the three simulation scenarios, referred to as *Simulation 1*, *Simulation 2* and *Simulation 3*.

Simulation 1

In *Simulation 1*, we draw V latent variables $\boldsymbol{w}_k$, each of dimension R_{gen}, from a mixture distribution given by

$$\boldsymbol{w}_k \sim \pi_w N_{R_{gen}}(\boldsymbol{w}_{mean}, \boldsymbol{w}_{sd}^2) + (1 - \pi_w)\boldsymbol{\delta_0};\ k \in \{1, ..., V\}, \tag{7}$$

where $\boldsymbol{\delta_0}$ is the Dirac-delta function and π_w is the probability of any $\boldsymbol{w}_k$ being nonzero. The (k,l)th element (and the (l,k)th element) of the network predictor coefficient $\boldsymbol{B}_0$, denoted by $\beta_{k,l,0}$ (and $\beta_{l,k,0}$), is given by $\beta_{k,l,0} = \beta_{l,k,0} = \frac{\boldsymbol{w}_k' \boldsymbol{w}_l}{2}$ (Hoff (2005)). Note that if $\boldsymbol{w}_k$ is zero, any edge connecting the kth node has no contribution to the regression mean function in (6), i.e. the kth node becomes inactive in predicting the response. According to (7), $(1-\pi_w)$ is the probability of a node being inactive. Hereafter, $(1 - \pi_w)$ is referred to as the *sparsity* parameter in the context of the data generation mechanism under *Simulation 1*. R_{gen} is the

true dimension of $\boldsymbol{w}_k$.

Simulation 2

To simulate $\boldsymbol{B}_0$ under *Simulation 2*, we draw V latent variables $\boldsymbol{w}_k$, each of dimension R_{gen}, from a mixture distribution given by

$$\boldsymbol{w}_k \sim \pi_{w^*} N_{R_{gen}}(\boldsymbol{w}_{mean}, \boldsymbol{w}_{sd}^2) + (1 - \pi_{w^*})\boldsymbol{\delta_0};\ k \in \{1, ..., V\}. \tag{8}$$

If $\boldsymbol{w}_k$ is simulated as $\mathbf{0}$, we set $\beta_{k,l,0} = \beta_{l,k,0} = 0$ for any l. Thus if $\boldsymbol{w}_k = \mathbf{0}$, the kth node is set to be inactive in predicting the response according to (6). If $\boldsymbol{w}_k$ and $\boldsymbol{w}_l$ are both nonzero, the edge coefficient $\beta_{k,l,0}$ connecting the kth and the lth nodes ($k < l$) is simulated from N(0.8,1). Respecting the symmetry condition, we set $\beta_{l,k,0} = \beta_{k,l,0}$. The edge between any two nodes k and l contributes in predicting the response if either $\boldsymbol{w}_k$ or $\boldsymbol{w}_l$ is nonzero. Hence in the context of *Simulation 2*, $1 - \pi_{w^*}$ is referred to as the *sparsity* parameter.

Simulation 3

To simulate $\boldsymbol{B}_0$ under *Simulation 3*, we draw V latent variables $\boldsymbol{w}_k$, each of dimension R_{gen}, from a mixture distribution given by

$$\boldsymbol{w}_k \sim \pi_{w^{**}} N_{R_{gen}}(\boldsymbol{w}_{mean}, \boldsymbol{w}_{sd}^2) + (1 - \pi_{w^{**}})\boldsymbol{\delta_0};\ k \in \{1, ..., V\}. \tag{9}$$

If $\boldsymbol{w}_k$ is simulated as $\mathbf{0}$, we set $\beta_{k,l,0} = \beta_{l,k,0} = 0$ for any l, i.e. the kth node is set to be inactive for the purpose of regression. The edge coefficient $\beta_{k,l,0}$ connecting any two nodes k and l with $\boldsymbol{w}_k$ and $\boldsymbol{w}_l$ both nonzero ($k < l$), is simulated from a mixture distribution given by

$$\beta_{k,l,0} \sim \pi_{w^{***}} N_{R_{gen}}(0.8, 1) + (1 - \pi_{w^{***}})\delta_0;\ k, l \in \{1, ..., V\}. \tag{10}$$

Again, we set $\beta_{k,l,0} = \beta_{l,k,0}$ respecting the symmetry condition. Contrary to *Simulation 2*, *Simulation 3* allows the possibility of an edge between the kth and the lth nodes having no impact on the response even when both $\boldsymbol{w}_k$ and $\boldsymbol{w}_l$ are nonzero. In the context of *Simulation 3*, $(1-\pi_{w^{**}})$ and $(1-\pi_{w^{***}})$ are referred to as the *node sparsity* and *edge sparsity* parameters.

It is worth mentioning that in *Simulation 1*, the simulated coefficient $\beta_{k,l,0}$ corresponding to the (k,l)th edge $a_{i,k,l}$ represents a bi-linear interaction between the latent variables corresponding to the kth and the lth nodes. Thus *Simulation 1* generates $\boldsymbol{B}_0$ respecting the network structure in $\boldsymbol{A}_i$ and is the most interesting simulation scenario to investigate. For a comprehensive picture for *Simulation 1*, we consider 9 different cases as summarized in Table 1. In each of these cases, the network predictor and the response are generated by changing the sparsity π_w and the true dimension R_{gen} of the latent variables $\boldsymbol{w}_k$'s. The table also presents the maximum dimension R of the latent variables $\boldsymbol{u}_k$ for the fitted network re-

gression model (2). Note that various cases also allow model mis-specification with unequal choices of R and R_{gen}.

Both *Simulation 2* and *Simulation 3* partially respect the network structure while considering no impact on the regression function in (6) from any edge that is connected to any inactive node. However, both in simulations 2 and 3, the coefficient corresponding to the (k, l)th edge $a_{i,k,l}$ is drawn from a normal distribution with no interaction specific to nodes k and l. Thus, the data generation schemes for simulations 2 and 3 will presumably offer no favor in terms of predictive ability or estimation of edge specific parameters to the network regression model over the ordinary high dimensional regression models that treat edges as bags of predictors and run a penalized optimization or shrinkage method. We present two cases for both simulations 2 and 3, recorded in Table 2. In all three simulation cases, apart from investigating the model as a tool for inference and prediction, it is of interest to observe the posterior distributions of λ_r's to judge if the model effectively learns the dimensionality R_{gen} of the latent variables $\boldsymbol{w}_k$. Noise variance τ_0^2 is fixed at 1 for all scenarios. $\boldsymbol{w}_{mean}$ and $\boldsymbol{w}_{sd}^2$ are set as $0.8 \times \mathbf{1}_{Rgen}$ and $\boldsymbol{I}_{R_{gen} \times R_{gen}}$ for all simulations.

Cases	R_{gen}	R	Sparsity
Case - 1	2	2	0.5
Case - 2	2	3	0.6
Case - 3	2	5	0.3
Case - 4	2	5	0.4
Case - 5	3	5	0.5
Case - 6	4	5	0.4
Case - 7	2	5	0.5
Case - 8	2	4	0.7
Case - 9	3	5	0.7

Table 1: Table presents different cases under Simulation 1. The true dimension R_{gen} is the dimension of vector object $\boldsymbol{w}_k$ using which data has been generated. The maximum dimension R is the dimension of vector object $\boldsymbol{u}_k$ using which the model has been fit. *Sparsity* refers to the fraction of generated $\boldsymbol{w}_k = \mathbf{0}$, i.e. $(1 - \pi_w)$.

3.2 Results

In all simulation results shown in this section, the BNR model is fitted with the choices of the hyper-parameters given by $\boldsymbol{S} = \boldsymbol{I}$, $\nu = 10$, $a_\Delta = 1$, $b_\Delta = 1$, $r_\lambda = 1$ and $\delta = 1$. Our extensive simulation studies reveal that both inference and prediction are robust with various choices of the hyper-parameters.

Node Selection

Figure 1 shows a matrix whose rows correspond to different cases in *Simulation 1* and

Simulation 2				Simulation 3				
Cases	R_{gen}	R	Sparsity	Cases	R_{gen}	R	Node Sparsity	Edge Sparsity
Case - 1	3	5	0.7	Case - 1	3	5	0.7	0.5
Case - 2	3	5	0.2	Case - 2	3	5	0.2	0.5

Table 2: Table presents different cases for Simulations 2 and 3. The true dimension R_{gen} is the dimension of vector object $\boldsymbol{w}_k$ using which data has been generated. The maximum dimension R is the dimension of vector object $\boldsymbol{u}_k$ using which the model has been fit. While Simulation 2 only has a sparsity parameter π_{w^*}, Simulation 3 has a node sparsity and an edge sparsity parameter.

columns correspond to the nodes of the network. The dark and clear cells correspond to the truly active and inactive nodes respectively. The posterior probability of the kth node being detected as active, i.e. $P(\xi_k = 1 \,|\, Data)$ has been overlaid for all $k \in \{1, ...20\}$ in all 9 simulation cases. The plot suggests overwhelmingly accurate detection of nodes influencing the response. We provide similar plots on node detection in Figure 2 for various cases in *Simulation 2* and *Simulation 3*. Both figures show overwhelming detection of active nodes both in simulations 2 and 3 with a very few false positives. We emphasize the fact that BNR is designed to detect important nodes while BLasso, HS or Lasso (or any other ordinary high dimensional regression technique) do not allow node selection in the present context. Relión *et al.* (2017) performs sub-optimally in terms of node selection and is not shown here.

Estimating the network coefficient

Another important aspect of the parametric inference lies in the performance in terms of estimating the network predictor coefficient. Table 3, 4, 5 present the MSE of all the competitors in simulations 1, 2 and 3 respectively. Given that both the fitted network regression coefficient $\boldsymbol{B}$ and the true coefficient $\boldsymbol{B}_0$ are symmetric, the MSE for any competitor is calculated in each dataset as $\frac{1}{q}\sum_{k<l}(\hat{\gamma}_{k,l} - \gamma_{k,l,0})^2$, where $\hat{\gamma}_{k,l}$ is the point estimate of $\gamma_{k,l} = 2\beta_{k,l}$ and $\gamma_{k,l,0}$ is the true value of the coefficient. For Bayesian models (such as the proposed model), $\hat{\gamma}_{k,l}$ is taken to be the posterior mean of $\gamma_{k,l}$.

From Table 3, it is evident that the proposed Bayesian network regression (BNR) significantly outperforms all its Bayesian and frequentist competitors in *Simulation 1* when the sparsity parameter is low to moderate (cases 1-7). While BNR is expected to perform much better than BLasso, Horseshoe and Lasso due to incorporation of network information, it is important to note that the carefully chosen local-global shrinkage with a well formulated hierarchical mean structure seems to possess more detection and estimation power than Relión *et al.* (2017). When the sparsity parameter is high in *Simulation 1* (cases 8-9), our simulation scheme sets an overwhelming proportion of $\beta_{k,l,0}$'s as zero. As a result, only a small subnetwork of $\boldsymbol{A}_i$ affects the response y_i, and hence BNR's performance becomes closely

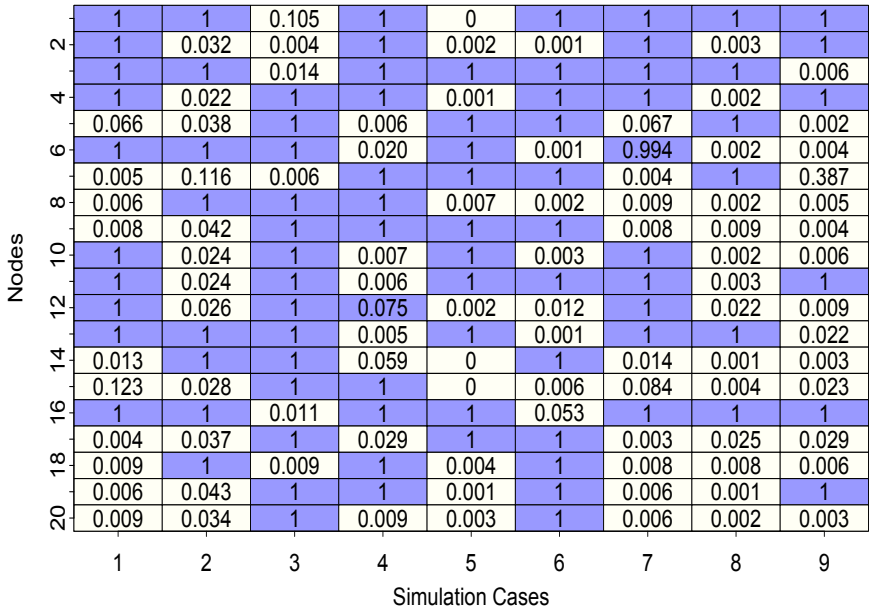
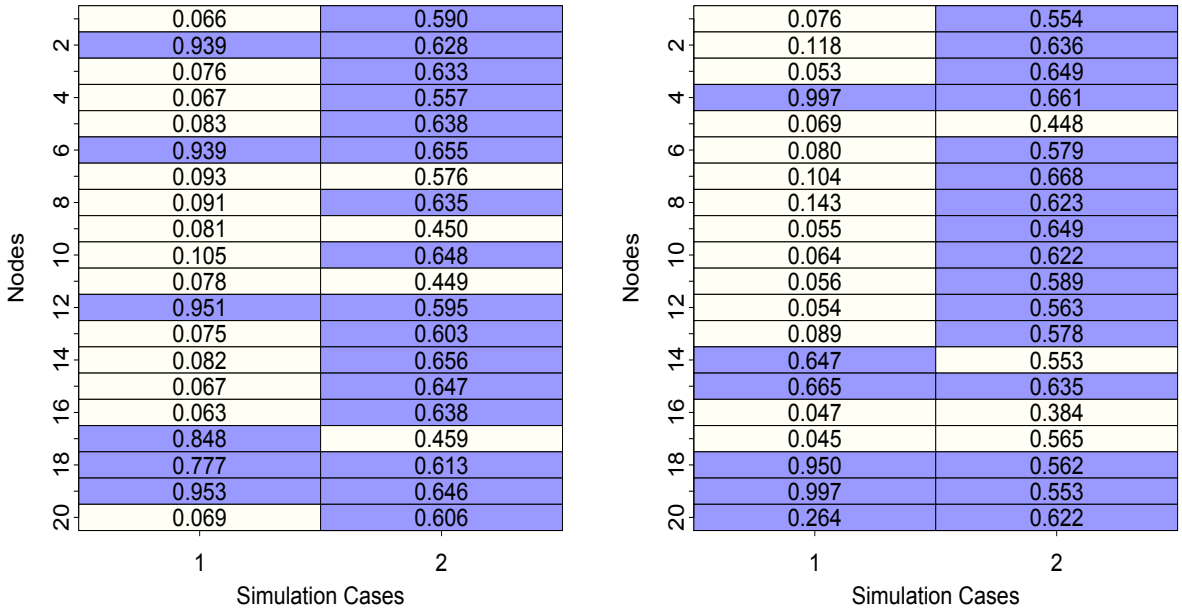


Figure 1: True activity status of a network node (clear blackground denotes *inactive* and dark background denotes *active*). Note that there are 20 rows (corresponding to 20 nodes) and 9 columns (corresponding to 9 different cases. The model-detected posterior probability of being active has been super-imposed onto the corresponding node.

comparable to Horseshoe and BLasso in terms of estimating $\boldsymbol{B}_0$ from (6). As argued earlier, generation of $\beta_{k,l,0}$ does not fully incorporate the interaction between the kth and the lth nodes in *Simulation 2* and *Simulation 3*. Thus in *Simulation 2* and *Simulation 3*, BNR is not expected to perform significantly better than its competitors. Indeed, Table 4 and 5 demonstrates comparable performance of all competitors with BNR slightly outperforming in the low-sparsity cases which are less conducive to ordinary high dimensional regression methods.

Inference on the effective dimensionality

Next, the attention turns to inferring on the posterior expected value of the effective dimensionality of $\boldsymbol{u}_k$. Figure 3 presents posterior probabilities of effective dimensionality in all 9 cases in *Simulation 1*. The filled bullets indicate the true value of the effective dimensionality. All 9 figures indicate that the true dimensionality of the latent variable $\boldsymbol{u}_k$ is effectively captured by the models. Mostly, the performance seems to have been impacted very negligibly by the change in sparsity or discrepancy between R and R_{gen}. Only in cases 8 and 9, in presence of less network information, the posterior distribution of R_{eff} appears to be bimodal. We also investigate similar figures for simulations 2 and 3 and arrive at an identical conclusion. In the interest of space, we omit these figures from the main text.



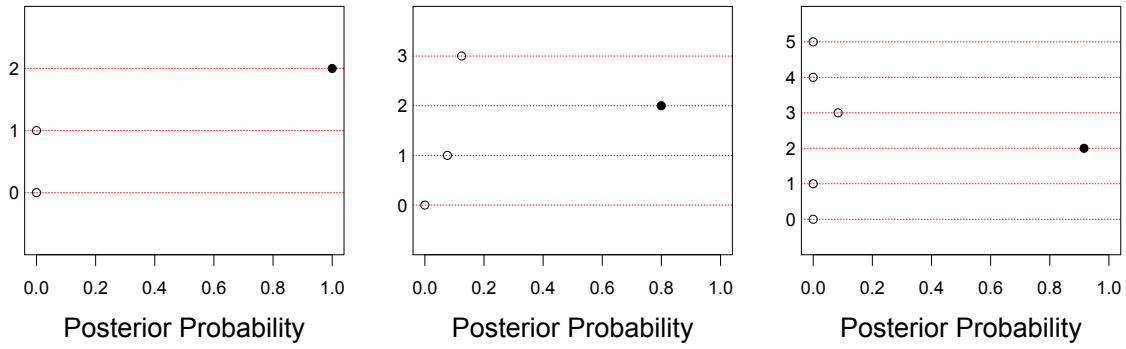
(a) Simulation 2

(b) Simulation 3

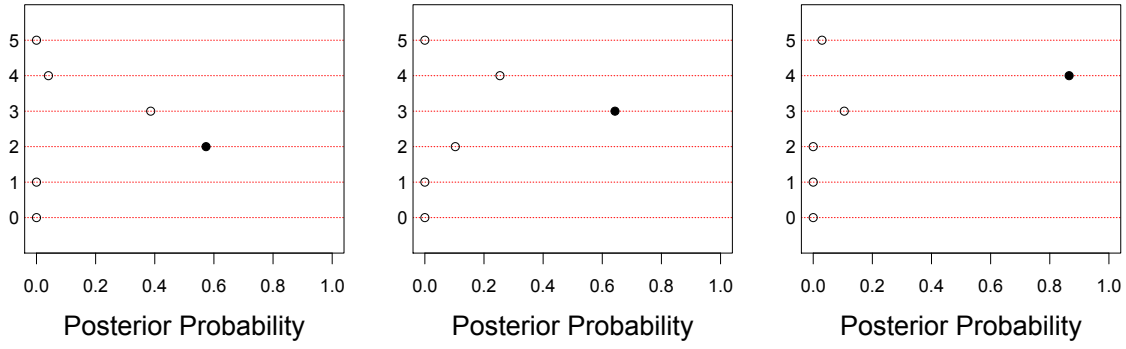
Figure 2: True activity status of a network node (clear blackground denotes *inactive* and dark background denotes *active*). Note that there are 20 rows (corresponding to 20 nodes) and 2 columns corresponding to 2 different cases both in Simulations 1 and 2. The model-detected posterior probability of being active has been super-imposed onto the corresponding node.

				MSE				
Cases	R_{gen}	R	Sparsity	**BNR**	Lasso	Relión(2017)	BLasso	Horseshoe
Case - 1	2	2	0.5	**0.009**	0.438	0.524	0.472	0.395
Case - 2	2	3	0.6	**0.007**	0.660	0.929	0.863	0.012
Case - 3	2	5	0.3	**0.006**	1.295	1.117	1.060	1.070
Case - 4	2	5	0.4	**0.006**	0.371	0.493	0.699	0.298
Case - 5	3	5	0.5	**0.009**	1.344	1.629	1.638	1.381
Case - 6	4	5	0.4	**0.006**	3.054	2.601	2.680	3.284
Case - 7	2	4	0.5	**0.009**	0.438	0.524	0.472	0.395
Case - 8	2	4	0.7	**0.005**	0.015	0.251	0.007	0.008
Case - 9	3	5	0.7	**0.004**	0.029	0.071	0.019	0.007

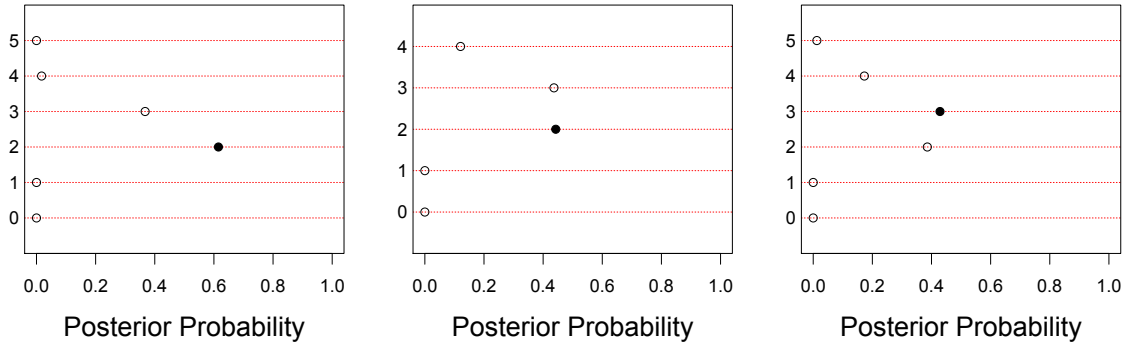
Table 3: Performance of Bayesian Network Regression (BNR) vis-a-vis competitors for cases in *Simulation 1*. Parametric inference in terms of point estimation of edge coefficients has been captured through the Mean Squared Error (MSE). The minimum MSE among competitors for any case is made bold.



(a) Case 1 (b) Case 2 (c) Case 3



(d) Case 4 (e) Case 5 (f) Case 6



(g) Case 7 (h) Case 8 (i) Case 9

Figure 3: Plots showing posterior probability distribution of effective dimensionality in all 9 cases in *Simulation 1*. Filled bullets indicate the true value of effective dimensionality.

				MSE				
Cases	R_{gen}	R	Sparsity	**BNR**	Lasso	Relión(2017)	BLasso	Horseshoe
Case - 1	3	5	0.7	0.011	0.013	0.036	0.010	**0.008**
Case - 2	3	5	0.2	**0.629**	0.843	0.859	0.836	0.948

Table 4: Performance of Bayesian Network Regression (BNR) vis-a-vis competitors for cases in *Simulation 2*. Parametric inference in terms of point estimation of edge coefficients has been captured through the Mean Squared Error (MSE). The minimum MSE among competitors for any case is made bold.

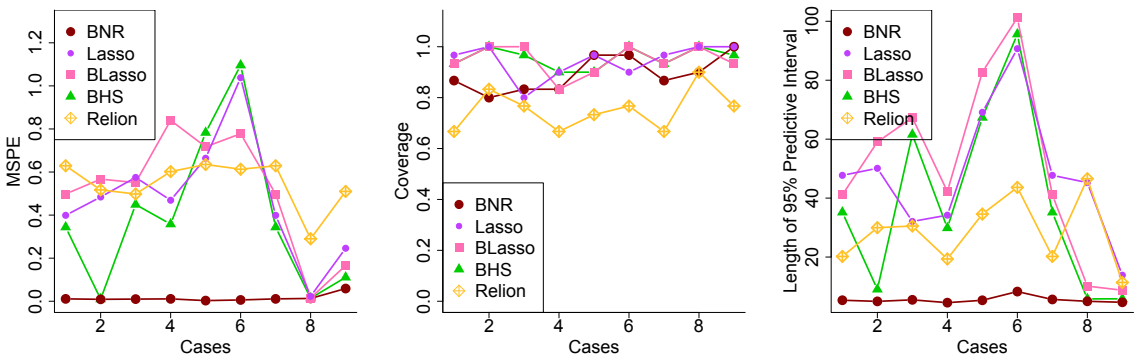
					MSE				
Cases	R_{gen}	R	Node Sparsity	Edge Sparsity	**BNR**	Lasso	Relión(2017)	BLasso	Horseshoe
Case - 1	3	5	0.7	0.5	0.004	0.006	0.017	0.004	**0.003**
Case - 2	3	5	0.2	0.5	**0.457**	0.636	0.617	0.659	0.629

Table 5: Performance of Bayesian Network Regression (BNR) vis-a-vis competitors for cases in *Simulation 3*. Parametric inference in terms of point estimation of edge coefficients has been captured through the Mean Squared Error (MSE). The minimum MSE among competitors for any case is made bold.

3.3 Predictive Inference

For the purpose of assessing predictive inference of the competitors, $n_{pred} = 30$ samples are generated from (6). We compare the predictive ability of competitors based on the point prediction and characterization of predictive uncertainties. To assess point prediction, we employ the mean squared prediction error (MSPE) which is obtained as the average squared distance between the point prediction and the true responses for all the competitors. As measures of predictive uncertainty, we provide coverage and length of 95% predictive intervals. For frequentist competitors, 95% predictive intervals are obtained by using predictive point estimates plus and minus 1.96 times standard errors. Figure 4 provides all three measures for all competitors in the 9 cases for *Simulation 1*.

It is quite evident from Figure 4(a) that BNR remarkably outperforms other competitors in terms of point prediction. Among the competitors, Horseshoe does a reasonably good job in cases with a higher degree of sparsity. This can be explained by taking into account the fact that the data generation procedure in *Simulation 1* ensures predictor coefficients to take reasonably large positive and negative values, along with an overwhelming number set to zero. Clearly, with a small sample size, all local and global parameters for the ordinary vector shrinkage priors are shrunk to zero to provide a good estimation of zero coefficients. While doing so, they heavily miss out on estimating coefficients which significantly deviate from



(a) MSPE (b) Coverage of 95% PI (c) Length of 95% PI

Figure 4: Figures from left to right show MSPE, coverage and length of 95% predictive intervals for all competitors.

zero. Thus, the ordinary vector shrinkage priors do a modest job in terms of prediction. Lasso also fails to provide satisfactory performance for similar reasons. It is important to note that all competitors provide close to nominal coverage. However, BNR yields so with predictive intervals $8-10$ times narrower than other competitors. The precise characterization of uncertainty in the BNR is due to fact that the relational nature of the predictor is properly accounted for. On the contrary, the ordinary high dimensional regression models do not account for the relational nature of the predictor and thus inflate the predictive interval. Overall, results from predictive inference from *Simulation 1* demonstrate strikingly superior performance of BNR.

The predictive performance of all competitors corresponding to simulations 2 and 3 are given in Table 6 and 7 respectively. Since the data generation scheme in simulations 2 and 3 minimally incorporates network structure, BNR does not shine much over its competitors in the high sparsity case. As expected, in presence of low node sparsity, performance of Lasso, BLasso and HS deteriorate and BNR turn out to be the best performer. Thus, when the data generation procedure incorporates minimal network structure, BNR shows comparable or little better performance than its competitors. Substantial gain using BNR is evident when the network structure is prominent as in *Simulation 1*.

4 Application to Human Brain Network Data

This section illustrates the inferential and predictive ability of Bayesian network regression in the context of a diffusion tensor imaging (DTI) dataset. Along with the brain network data, the dataset of interest contains a measure of *creativity* (Kiar *et al.* (2016); Kiar *et al.* (2017a);

				MSPE				
Cases	R_{gen}	R	Sparsity	**BNR**	Lasso	Relión(2017)	BLasso	Horseshoe
Case - 1	3	5	0.7	0.079	0.100	0.371	0.076	**0.061**
Case - 2	3	5	0.2	**0.432**	0.726	0.859	0.629	0.725
				Coverage of 95% PI				
Case - 1	3	5	0.7	1.00	1.00	0.867	1.00	0.967
Case - 2	3	5	0.2	0.94	0.73	0.56	0.96	0.87
				Length of 95% PI				
Case - 1	3	5	0.7	8.97	18.70	10.23	8.25	6.40
Case - 2	3	5	0.2	42.18	32.81	23.54	45.69	42.91

Table 6: MSPE, coverage and length of 95% predictive intervals (PIs) of Bayesian Network Regression (BNR) vis-a-vis competitors for cases in Simulation 2. Lowest MSPE for any case is made bold.

					MSPE				
Cases	R_{gen}	R	Node Sparsity	Edge Sparsity	**BNR**	Lasso	Relión(2017)	BLasso	Horseshoe
Case - 1	3	5	0.7	0.5	0.119	0.151	0.425	0.125	**0.096**
Case - 2	3	5	0.2	0.5	**0.451**	0.549	0.699	0.692	0.566
					Coverage of 95% PI				
Case - 1	3	5	0.7	0.5	0.93	1.00	0.86	0.96	0.96
Case - 2	3	5	0.2	0.5	1.00	0.83	0.70	1.00	1.00
					Length of 95% PI				
Case - 1	3	5	0.7	0.5	6.18	14.19	8.35	6.44	5.91
Case - 2	3	5	0.2	0.5	41.69	27.98	17.84	51.70	49.12

Table 7: MSPE, coverage and length of 95% predictive intervals (PIs) of Bayesian Network Regression (BNR) vis-a-vis competitors for cases in Simulation 3. Lowest MSPE for any case is made bold.

Kiar *et al.* (2017b)) for several subjects, known as the Composite Creativity Index (CCI). The scientific goal in this setting pertains to understanding the relationship between brain connectivity and the composite creativity index (CCI). Specifically, our main interest lies in proposing a regression relationship between the brain network and CCI, and subsequently predicting the CCI of a subject from his/her brain network, which serves as the predictor. Additionally, it is desirable to identify brain regions (nodes in the brain network) that are involved with creativity, as well as significant connections between different brain regions. Our analysis involves 37 subjects.

Human creativity has been at the crux of the evolution of the human civilization, and has been the topic of research in several disciplines including, of course, neuroscience. Though

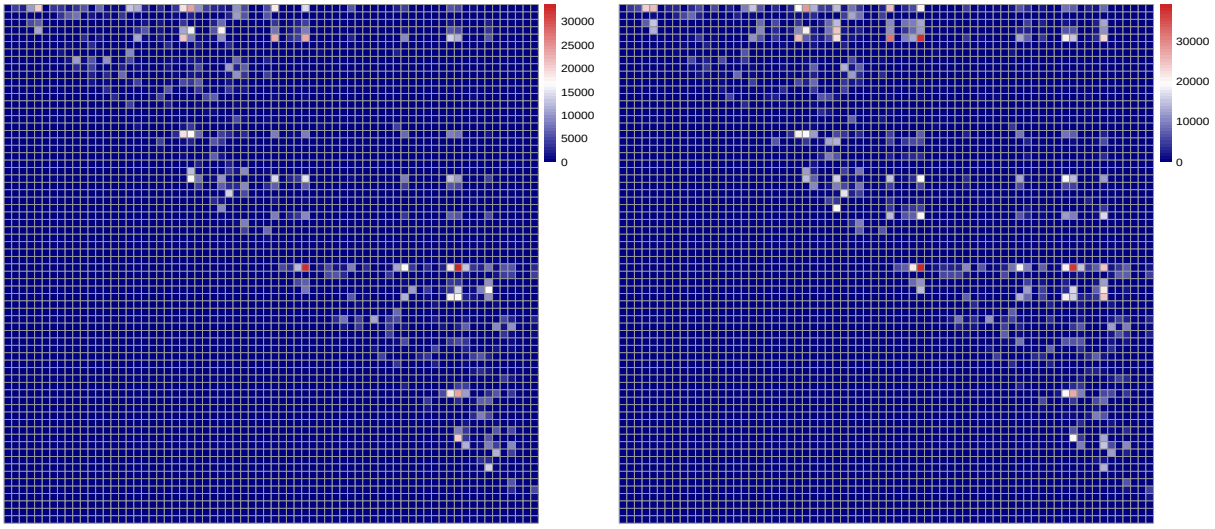
creativity can be defined in numerous ways, one could envision a creative idea as one that is unusual as well as effective in a given social context (Flaherty (2005)). Neuroscientists generally concur that a coalescence of several cognitive processes determines the creative process, which often involves a *divergence of ideas* to conceivable solutions for a given problem. To measure the creativity of an individual, Jung *et al.* (2010) propose the CCI, which is formulated by linking measures of divergent thinking and creative achievement to cortical thickness of young (23.7 ± 4.2 years), healthy subjects. Three independent judges grade the creative products of a subject from which the "composite creativity index" (CCI) is derived. CCI serves as the response in our study.

The brain network information for subjects is gathered using the diffusion tensor imaging technique. Diffusion tensor imaging (DTI) is a magnetic resonance imaging technique that enables measurement of the restricted diffusion of water in tissue in order to produce neural tract images. In the context of DTI, the human brain is divided according to the Desikan atlas (Desikan *et al.* (2006)) that identifies 34 cortical regions of interest (ROIs) both in the left and right hemispheres of the human brain, implying 68 cortical ROIs in all. A 'brain network' for each subject is represented by a symmetric adjacency matrix whose rows and columns correspond to different ROIs and entries correspond to estimates of the number of 'fibers' connecting pairs of brain regions. Thus, for each individual, representing the brain network, is a weighted adjacency matrix of dimension 68×68, with the (k, l)th off-diagonal entry in the adjacency matrix being the estimated number of fibers connecting the kth and the lth brain regions. Figure 5 shows maps of the brain network for two representative individuals in the sample.

4.1 Findings from BNR

We focus on identifying influential ROIs in the brain network using the node selection strategy described in the simulation studies. For the purpose of this data analysis, the Bayesian network regression model is fitted with $R = 5$ which is found to be sufficient for this study. Recall that the kth node is identified as *active* if $P(\xi_k = 1 \,|\, Data)$ exceeds 0.5. This principle, when applied to the real data discussed above, identifies 36 ROIs out of 68 as *active*. Of these 36 ROIs, 16 belong to the left portion of the brain (or the left hemisphere) and 20 belong to the right hemisphere. Table 9 shows the brain regions of interest (ROIs) in the Desikan atlas detected as being actively associated with the CCI.

A large number of the 36 active nodes detected by our method are part of the *frontal* (15) and *temporal* (8) cortices in both hemispheres. The frontal cortex has been scientifically associated with divergent thinking and problem solving ability, in addition to motor function, spontaneity, memory, language, initiation, judgement, impulse control, and social behavior (Stuss *et al.* (1985)). Some of the other functions directly related to the frontal cortex



(a) Representative Network Adjacency Matrix 1 (b) Representative Network Adjacency Matrix 2

Figure 5: Figure shows maps of the brain network (weighted adjacency matrices) for two representative individuals in the sample. Since the (k, l)th off-diagonal entry in any adjacency matrix corresponds to the number of *fibers* connecting the kth and the lth ROIs, the adjacency matrices are symmetric. Hence the figure only shows the upper triangular portion.

seem to be "behavioral spontaneity", interpreting environmental feedback and risk taking (Razumnikova (2007); Miller and Milner (1985) ; Kolb and Milner (1981)). On the other hand, Finkelstein *et al.* (1991) report *de novo* artistic expression to be associated with the temporal and frontal regions. Our method also finds a strong relationship between creativity and the *right parahippocampal gyrus* and *right inferior parietal lobule*, regions found active by a few earlier scientific studies, see e.g., Chavez *et al.* (2004).

As a reference point to our analysis, we compare our findings with Jung *et al.* (2010). It is worth mentioning that Jung *et al.* (2010) does not employ a sophisticated model based Bayesian analysis for identifying important nodes with uncertainties. Hence, we do not expect a complete overlap with their results. Our analysis finds a number of overlaps with the regions that Jung *et al.* (2010) identify as significantly associated with the creativity process, namely the *middle frontal gyrus*, the *left cingulate cortex*, the *left orbitofrontal* region, the *left lingual* region, the *right fusiform*, the *left cuneus* and the *right superior parietal lobule*. They also find the *inferior parietal*, *superior parietal* lobules and the *right posterior singulate* regions to be very significantly associated with creativity. Our model detects these nodes as active. Although there is significant intersection between the findings of Jung *et al.* (2010) and our method, there are a couple of regions that we detect as active and they do not, and vice versa. For e.g., our model detects the *precuneus* and the *supramarginal* regions in both the hemispheres to be significantly related to CCI, while they do not. On the other hand,

they identify the *right angular* region to be significant while we do not.

Along with the influential ROIs, another important scientific question remains identifying the statistically significant edges or connections between the 68 ROIs. We consider the edge between two ROIs k and l to have a statistically significant impact on CCI if the 95% credible interval of the posterior distribution of its corresponding coefficient $\gamma_{k,l}$ does not contain 0. Under this measure, our model identifies 576 significant $\gamma_{k,l}$'s connecting 30 ROIs. Figure 6 plots significant inter-connections detected among brain regions of interest (ROIs), where the brain can be viewed from different angles. Red dots show the active ROIs and blue lines show significant connections between them. It is to be noted that *all* 30 of the aforementioned nodes are a subset of the group of 36 nodes that have been detected as *active*. This hints at effective detection. Figure 7 plots these influential interconnections, where a white cell represents an edge predictive of the response with the corresponding row ROI and column ROI. Since this is an undirected network, the matrix is symmetric and we only show connections in the upper triangular region.

Finally, our interest turns to the predictive ability of the Bayesian network regression model. To this end, Table 8 reports the mean squared prediction error (MSPE) between observed and predicted responses, length and coverage of 95% predictive intervals. Here, the average is computed over 10 cross-validated folds. As reference, we also present MSPE, length and coverage values for Lasso, BLasso and Relión *et al.* (2017). Note that Lasso and BLasso completely ignore the network structure in the predictor. Thus the edge predictors become highly correlated, with the correlation structure governed by the network structure in the predictor. It is well known that high dimensional regression models perform poorly in terms of predictive performance when the sample size is small and predictors are highly correlated. In the present context, the poor MSPE of Lasso and BLasso can be attributed to this phenomenon. The table clearly shows excellent point prediction of our proposed approach even under small sample size and low signal to noise ratio. Additionally, BNR provides close to nominal coverage for 95% predictive intervals with a much narrower predictive interval. Hence, uncertainty quantification from BNR also turns out to be precise.

	BNR	Lasso	BLasso	Relión(2017)
MSPE	0.69	0.98	1.84	0.98
Coverage of 95% PI	0.93	0.97	0.97	0.97
Length of 95% PI	2.72	3.88	3.40	3.89

Table 8: Predictive performance of competitors in terms of mean squared prediction error (MSPE), coverage and length of 95% predictive intervals, obtained through 10-Fold Cross Validation in the context of real data. Note that since the response has been standardized, an MSPE value greater than or around 1 will denote an inconsequential analysis.

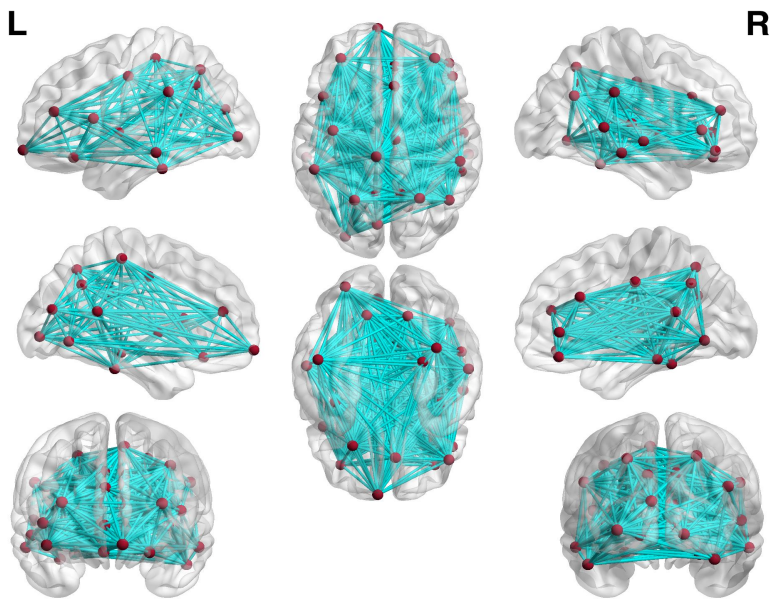


Figure 6: Significant inter-connections detected among brain regions of interest (ROIs) in the Desikan atlas. Red dots show the *active* ROIs and blue lines show significant connections between them.

5 Conclusion and Future Work

This article proposes a novel and pioneering Bayesian framework to address a regression problem with a continuous response and network-valued predictors, respecting the underlying network structure. Our contribution lies in carefully constructing a novel class of network shrinkage priors corresponding to the network predictor, simultaneously recognizing the latent network structure in the predictor variable. Empirical results from simulation studies display remarkably superior performance of our method, both in terms of inference as well as prediction. Our framework is employed to analyze a brain connectome data on composite creativity index along with the brain network of multiple individuals. It is able to identify important regions in the brain and important brain connectivity patterns which have profound influence on the creativity of a person.

A number of future directions emerge from this work. First, our framework finds natural extension to regression problems with a binary response and *any* network predictor, whether binary or weighted. Such a framework would be useful in various classification problems involving network predictors, e.g. in classifying diseased patients from normal people in neuroimaging studies. Another important direction appears to be the development of a regression framework with the network as the response regressed on a few scalar/vector predictors. Some of these constitute our current work.

Left Hemisphere Lobes					
Temporal	**Cingulate**	**Frontal**	**Occipital**	**Parietal**	**Insula**
inferior temporal gyrus	isthmus cingulate cortex	lateral orbitofrontal	cuneus	precuneus	insula
middle temporal gyrus		paracentral	lateral occipital gyrus	superior parietal lobule	
		pars opercularis	lingual	supramarginal gyrus	
		precentral			
		rostral middle frontal gyrus			
		frontal pole			

Right Hemisphere Lobes					
Temporal	**Cingulate**	**Frontal**	**Occipital**	**Parietal**	**Insula**
bank of the superior temporal sulcus	caudal anterior cingulate	medial orbitofrontal	lingual	inferior parietal lobule	insula
fusiform	isthmus cingulate cortex	pars orbitalis		precuneus	
middle temporal gyrus	posterior cingulate cortex	pars triangularis		superior parietal lobule	
parahippocampal	rostral anterior cingulate cortex	rostral middle frontal gyrus		supramarginal gyrus	
superior temporal gyrus					
transverse temporal					

Table 9: Brain regions (ROIs) detected as actively associated with the composite creativity index by BNR.

Appendix A

This section provides details of posterior computation for all the parameters in the Bayesian network regression with a continuous response.

Let $\boldsymbol{x}_i = (a_{i,1,2}, a_{i,1,3}, ..., a_{i,1,V}, a_{i,2,3}, a_{i,2,4}, ..., a_{i,2,V}, ..., a_{i,V-1,V})'$ be of dimension $q \times 1$, where $q = \frac{V \times (V-1)}{2}$. Assume $\boldsymbol{y} = (y_1, ..., y_n)' \in \mathbb{R}^n$ and $\boldsymbol{X} = (\boldsymbol{x}_1 : \cdots : \boldsymbol{x}_n)'$ is an $n \times q$ matrix. Further, assume $\boldsymbol{W} = (\boldsymbol{u}_1' \boldsymbol{\Lambda} \boldsymbol{u}_2, ..., \boldsymbol{u}_1' \boldsymbol{\Lambda} \boldsymbol{u}_V, ..., \boldsymbol{u}_{V-1}' \boldsymbol{\Lambda} \boldsymbol{u}_V)'$, $\boldsymbol{D} = diag(s_{1,2}, ..., s_{V-1,V})$ and $\boldsymbol{\gamma} = (\gamma_{1,2}, ..., \gamma_{V-1,V})'$. Thus, with n data points, the hierarchical model with the Bayesian Network Lasso prior can be written as

$$\boldsymbol{y} \sim \mathrm{N}(\mu + \boldsymbol{X}\boldsymbol{\gamma}, \tau^2 \boldsymbol{I})$$

$$\boldsymbol{\gamma} \sim \mathrm{N}(\boldsymbol{W}, \tau^2 \boldsymbol{D}),\ log(\tau^2) \sim flat(),\ \boldsymbol{u}_k | \xi_k = 1 \sim N(\boldsymbol{u}_k \,|\, \boldsymbol{0}, \boldsymbol{M}),\ \boldsymbol{u}_k | \xi_k = 0 \sim \delta_{\boldsymbol{0}},\ \mu \sim flat()$$

$$s_{k,l} \sim Exp(\lambda^2/2), \quad \lambda^2 \sim Gamma(r_\lambda, \delta),\ \boldsymbol{M} \sim IW(\boldsymbol{S}, \nu),\ \Delta \sim Beta(a_\Delta, b_\Delta), \xi_k \sim Ber(\Delta)$$

$$\lambda_r \sim Ber(\pi_r),\ \pi_r \sim Beta(1, r^\eta),\ \eta > 1.$$

The hierarchical model specified above leads to straightforward Gibbs sampling with full conditionals obtained as following:

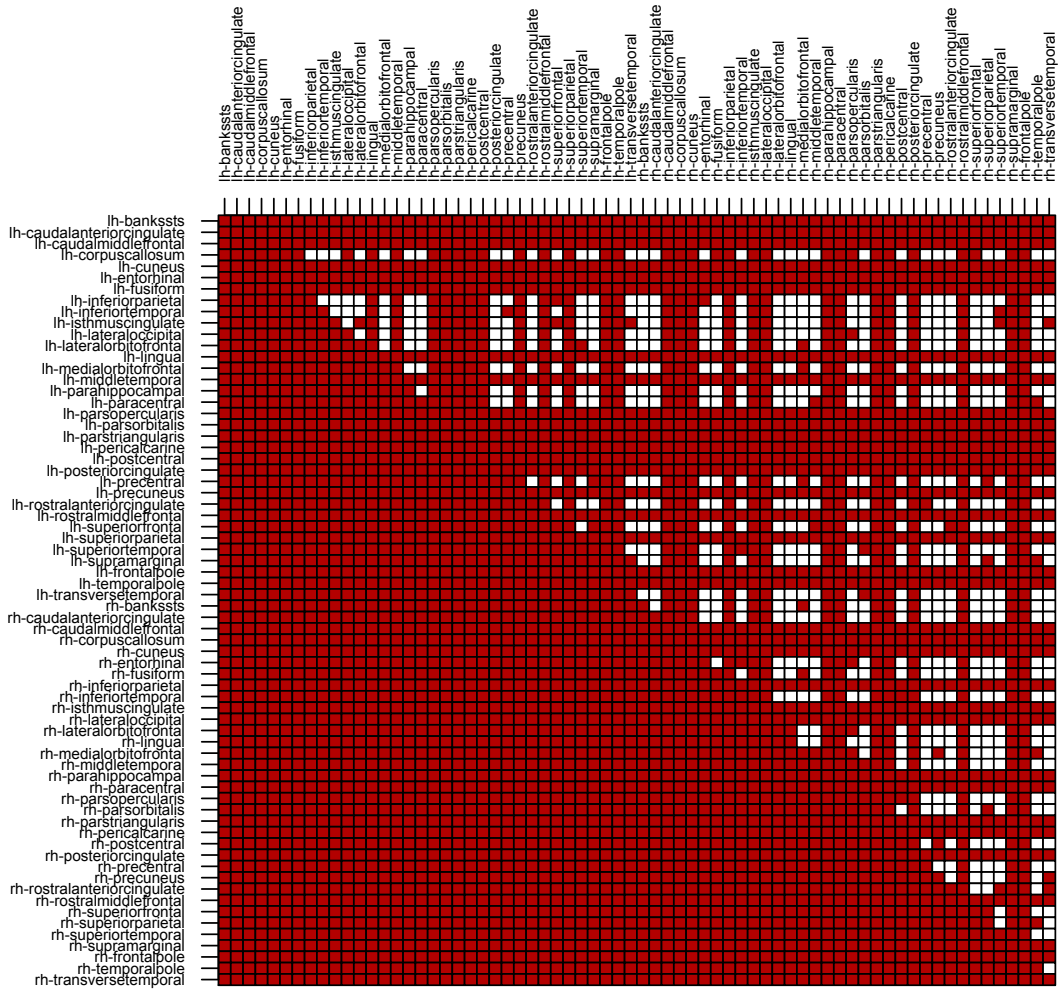


Figure 7: Significant inter-connections detected among brain regions of interest (ROIs) in the Desikan atlas. White cells show significant nodal associations among ROIs denoted by the corresponding rows and columns. Prefix 'lh-' and 'rh-' in the ROI names denote their positions in the left and right hemispheres of the brain respectively. For full names of the ROIs specified on the axes, please consult the widely available Desikan Atlas.

- $\mu \,|\, - \sim N\left(\frac{\mathbf{1}'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\gamma})}{n}, \frac{\tau^2}{n}\right)$
- $\boldsymbol{\gamma} \,|\, - \sim N(\boldsymbol{\mu}_{\boldsymbol{\gamma}\,|\,\cdot}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}\,|\,\cdot})$, where $\boldsymbol{\mu}_{\boldsymbol{\gamma}\,|\,\cdot} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}^{-1})^{-1}(\boldsymbol{X}'(\boldsymbol{y} - \mu\mathbf{1}) + \boldsymbol{D}^{-1}\boldsymbol{W})$ and $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\,|\,\cdot} = \tau^2(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}^{-1})^{-1}$
- $\tau^2 \,|\, - \sim IG\left[(\frac{n}{2} + \frac{V(V-1)}{4}), \frac{(\boldsymbol{y}-\mu\mathbf{1}-\boldsymbol{X}\boldsymbol{\gamma})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\gamma})+(\boldsymbol{\gamma}-\boldsymbol{W})'\boldsymbol{D}^{-1}(\boldsymbol{\gamma}-\boldsymbol{W})}{2}\right]$

- $s_{k,l} \mid - \sim GIG\left[\frac{1}{2}, \frac{(\gamma_{k,l} - \boldsymbol{u}_k' \boldsymbol{\Lambda} \boldsymbol{u}_l)^2}{\tau^2}, \lambda^2\right]$, where GIG denotes the generalized inverse Gaussian distribution.
- $\lambda^2 \mid - \sim Gamma\left[\left(r_\lambda + \frac{V(V-1)}{2}\right), \left(\delta + \sum_{k<l} \frac{s_{k,l}}{2}\right)\right]$
- $\boldsymbol{u}_k \mid - \sim w_{\boldsymbol{u}_k}\, \delta_0(\boldsymbol{u}_k) + (1 - w_{\boldsymbol{u}_k})\, N(\boldsymbol{u}_k \mid \boldsymbol{m}_{\boldsymbol{u}_k}, \boldsymbol{\Sigma}_{\boldsymbol{u}_k})$, where $\boldsymbol{U}_k^* = (\boldsymbol{u}_1 : \cdots : \boldsymbol{u}_{k-1} : \boldsymbol{u}_{k+1} : \cdots : \boldsymbol{u}_V)' \boldsymbol{\Lambda}$, $\boldsymbol{H}_k = diag(s_{1,k}, ..., s_{k-1,k}, s_{k,k+1}, ..., s_{k,V})$, $\boldsymbol{\gamma}_k = (\gamma_{1,k}, ..., \gamma_{k-1,k}, \gamma_{k,k+1}, ..., \gamma_{k,V})$, and

$$\boldsymbol{\Sigma}_{\boldsymbol{u}_k} = \left(\boldsymbol{U}_h^{*'} \boldsymbol{H}_k^{-1} \boldsymbol{U}_k^* / \tau^2 + \boldsymbol{M}^{-1}\right)^{-1}, \;\; \boldsymbol{m}_{\boldsymbol{u}_k} = \boldsymbol{\Sigma}_{\boldsymbol{u}_k} \boldsymbol{U}_k^{*'} \boldsymbol{H}_k^{-1} \boldsymbol{\gamma}_k / \tau^2$$
$$w_{\boldsymbol{u}_k} = \frac{(1-\pi) N(\boldsymbol{\gamma}_k \mid \boldsymbol{0}, \tau^2 \boldsymbol{H}_k)}{(1-\pi) N(\boldsymbol{\gamma}_k \mid \boldsymbol{0}, \tau^2 \boldsymbol{H}_k) + \pi N(\boldsymbol{\gamma}_k \mid \boldsymbol{0}, \tau^2 \boldsymbol{H}_k + \boldsymbol{U}_k^* \boldsymbol{M} \boldsymbol{U}_k^{*'})}$$

- $\xi_k | - \sim Ber(1 - w_{\boldsymbol{u}_k})$
- $\Delta \mid - \sim Beta\left[(a_\Delta + \sum_{k=1}^V \xi_k), (b_\Delta + \sum_{k=1}^V (1 - \xi_k))\right]$.
- $\boldsymbol{M} \mid - \sim IW[(\boldsymbol{S} + \sum_{k:\boldsymbol{u}_k \neq \boldsymbol{0}} \boldsymbol{u}_k \boldsymbol{\Lambda} \boldsymbol{u}_k'), (\nu + \{\#k : \boldsymbol{u}_k \neq \boldsymbol{0}\})]$.
- $\lambda_r \mid - \sim Ber(p_{\lambda_r})$, where $p_{\lambda_r} = \frac{\pi_r N(\boldsymbol{\gamma} \mid \boldsymbol{W}_1, \tau^2 \boldsymbol{D})}{\pi_r N(\boldsymbol{\gamma} \mid \boldsymbol{W}_1, \tau^2 \boldsymbol{D}) + (1-\pi_r) N(\boldsymbol{\gamma} \mid \boldsymbol{W}_0, \tau^2 \boldsymbol{D})}$. Here $\boldsymbol{W}_1 = (\boldsymbol{u}_1' \boldsymbol{\Lambda}_1 \boldsymbol{u}_2, ..., \boldsymbol{u}_1' \boldsymbol{\Lambda}_1 \boldsymbol{u}_V, ..., \boldsymbol{u}_{V-1}' \boldsymbol{\Lambda}_1 \boldsymbol{u}_V)'$, $\boldsymbol{W}_0 = (\boldsymbol{u}_1' \boldsymbol{\Lambda}_0 \boldsymbol{u}_2, ..., \boldsymbol{u}_1' \boldsymbol{\Lambda}_0 \boldsymbol{u}_V, ..., \boldsymbol{u}_{V-1}' \boldsymbol{\Lambda}_0 \boldsymbol{u}_V)'$, $\boldsymbol{\Lambda}_1 = diag(\lambda_1, .., \lambda_{r-1}, 1, \lambda_{r+1}, .., \lambda_R)$, $\boldsymbol{\Lambda}_0 = diag(\lambda_1, .., \lambda_{r-1}, 0, \lambda_{r+1}, .., \lambda_R)$, for $r = 1, .., R$.
- $\pi_r \mid - \sim Beta(\lambda_r + 1, 1 - \lambda_r + r^\eta)$, for $r = 1, .., R$.

Appendix B

This section shows the posterior propriety of the parameters in the BNR model. Without loss of generality, we set $\mu = 0$ while proving the posterior propriety. To begin with, we state a number of useful lemmas.

Lemma 5.1 *If $\boldsymbol{C}$ is an $h \times h$ non-negative definite matrix, then $|\boldsymbol{C} + \boldsymbol{I}| \geq 1$.*

Proof The eigenvalues of $(\boldsymbol{C} + \boldsymbol{I})$ are given by $\varphi_1 + 1, ..., \varphi_m + 1$, where $\varphi_1, ..., \varphi_m$ are eigenvalues of $\boldsymbol{C}$. Since $\boldsymbol{C}$ is non-negative definite, $\varphi_1 \geq 0, ..., \varphi_m \geq 0$. The result follows from the fact that $|\boldsymbol{C} + \boldsymbol{I}| = \prod_{m=1}^h (\varphi_m + 1)$ is the product of eigenvalues.

Lemma 5.2 *If $\boldsymbol{C}$ is an $h \times h$ diagonal matrix with diagonal entries $c_1, ..., c_h$ all greater than 0. Suppose $\boldsymbol{A}$ is an $n \times h$ matrix with the largest eigenvalue of $\boldsymbol{A}\boldsymbol{A}'$ is given by $\mu_{\boldsymbol{A}\boldsymbol{A}'}$. Then $\boldsymbol{A}\boldsymbol{C}\boldsymbol{A}' + \boldsymbol{I} \leq \left(\mu_{\boldsymbol{A}\boldsymbol{A}'} \sum_{l=1}^h c_l + 1\right) \boldsymbol{I}$, where $\boldsymbol{H}_1 \leq \boldsymbol{H}_2$ implies $\boldsymbol{H}_2 - \boldsymbol{H}_1$ is a positive definite matrix.*

Proof Since $c_1, ..., c_h > 0$, $\boldsymbol{ACA'} \leq (\sum_{l=1}^{h} c_l)\boldsymbol{AA'}$. Consider the spectral decomposition of the matrix $\boldsymbol{AA'}$. Let the eigen-decomposition of $\boldsymbol{AA'} = \boldsymbol{\Lambda H \Lambda'}$, where $\boldsymbol{\Lambda}$ is the matrix of eigenvectors and $\boldsymbol{H}$ is a diagonal matrix with diagonal entries $\mu_1, ..., \mu_n$. Since each $\mu_i \leq \mu_{\boldsymbol{AA'}}$, $\boldsymbol{AA'} \leq \mu_{\boldsymbol{AA'}}\boldsymbol{\Lambda\Lambda'} = \mu_{\boldsymbol{AA'}}\boldsymbol{I}$. Thus, $\boldsymbol{ACA'} \leq (\sum_{l=1}^{h} c_l)\mu_{\boldsymbol{AA'}}\boldsymbol{I}$. Hence $\boldsymbol{ACA'} + \boldsymbol{I} \leq \left(\mu_{\boldsymbol{AA'}} \sum_{l=1}^{h} c_l + 1\right)\boldsymbol{I}$.

Lemma 5.3 *Suppose $\boldsymbol{z}$ is an $h \times 1$ vector and $\boldsymbol{A}$ is an $h \times h$ positive definite matrix. Let $\boldsymbol{B}$ be another $h \times h$ positive definite matrix such that $\boldsymbol{A} \geq \boldsymbol{B}$ (where $\boldsymbol{A} \geq \boldsymbol{B}$ implies $\boldsymbol{A} - \boldsymbol{B}$ is non-negative definite). Then $\boldsymbol{z'A^{-1}z} \leq \boldsymbol{z'B^{-1}z}$.*

Proof $\boldsymbol{A} \geq \boldsymbol{B}$ implies $\boldsymbol{A^{-1}} \leq \boldsymbol{B^{-1}}$. Then $\boldsymbol{z'A^{-1}z} \leq \boldsymbol{z'B^{-1}z}$.

Note that the posterior distribution of the parameters is given by:

$$p(\boldsymbol{\gamma}, \boldsymbol{\Lambda}, \tau^2, \boldsymbol{u}_1, .., \boldsymbol{u}_V, \xi_1, .., \xi_V, \lambda^2, \Delta, \{s_{k,l}\}_{k<l}, \pi_1, ..., \pi_R, \boldsymbol{M} \,|\, \boldsymbol{y}, \boldsymbol{X}) \propto \mathrm{N}(\boldsymbol{y} \,|\, \boldsymbol{X\gamma}, \tau^2\boldsymbol{I}) \times \mathrm{N}(\boldsymbol{\gamma} \,|\, \boldsymbol{W}, \tau^2\boldsymbol{D}) \times$$
$$\frac{1}{\tau^2} \times \prod_{k=1}^{V} [\xi_k N(\boldsymbol{u}_k \,|\, \boldsymbol{0}, \boldsymbol{M}) + (1 - \xi_k)\delta_{\boldsymbol{0}}] \times \prod_{k<l} Exp(s_{k,l} \,|\, \lambda^2/2) \times Gamma(\lambda^2 \,|\, r_\lambda, \delta) \times$$
$$IW(\boldsymbol{M} \,|\, \boldsymbol{S}, \nu) \times Beta(\Delta \,|\, a_\Delta, b_\Delta) \times Ber(\lambda_r \,|\, \pi_r) \times Beta(\pi_r \,|\, 1, r^\eta) \times \prod_{k=1}^{V} Ber(\xi_k \,|\, \Delta).$$

Integrating over $\xi_1, ..., \xi_V$,

$$p(\boldsymbol{\gamma}, \boldsymbol{\Lambda}, \tau^2, \boldsymbol{u}_1, .., \boldsymbol{u}_V, \lambda^2, \Delta, \{s_{k,l}\}_{k<l}, \pi_1, ..., \pi_R, \boldsymbol{M} \,|\, \boldsymbol{y}, \boldsymbol{X}) \propto \mathrm{N}(\boldsymbol{y} \,|\, \boldsymbol{X\gamma}, \tau^2\boldsymbol{I}) \times \mathrm{N}(\boldsymbol{\gamma} \,|\, \boldsymbol{W}, \tau^2\boldsymbol{D}) \times$$
$$\frac{1}{\tau^2} \times \prod_{k=1}^{V} [\Delta N(\boldsymbol{u}_k \,|\, \boldsymbol{0}, \boldsymbol{M}) + (1 - \Delta)\delta_{\boldsymbol{0}}] \times \prod_{k<l} Exp(s_{k,l} \,|\, \lambda^2/2) \times Gamma(\lambda^2 \,|\, r_\lambda, \delta) \times$$
$$IW(\boldsymbol{M} \,|\, \boldsymbol{S}, \nu) \times Beta(\Delta \,|\, a_\Delta, b_\Delta) \times Ber(\lambda_r \,|\, \pi_r) \times Beta(\pi_r \,|\, 1, r^\eta).$$

The prior specifications on λ_r, π_r, Δ enable these quantities to be bounded within a finite interval. Thus in showing the posterior propriety of parameters with unbounded range, it is enough to treat these quantities as constant. We treat them as fixed henceforth.

Integrating over $\boldsymbol{\gamma}$, we obtain,

$$p(\boldsymbol{u}_1, .., \boldsymbol{u}_V, \tau^2, \lambda^2, \{s_{k,l}\}_{k<l}, \boldsymbol{M} \,|\, \boldsymbol{y}, \boldsymbol{X}) \propto \frac{1}{(\tau^2)^{n/2} |\boldsymbol{XDX'} + \boldsymbol{I}|^{1/2}} \times$$
$$\exp\left\{-\frac{(\boldsymbol{y} - \boldsymbol{XW})'(\boldsymbol{XDX'} + \boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{XW})}{2\tau^2}\right\} \times \prod_{k=1}^{V} [\Delta N(\boldsymbol{u}_k \,|\, \boldsymbol{0}, \boldsymbol{M}) + (1 - \Delta)\delta_{\boldsymbol{0}}] \times$$
$$\prod_{k<l} Exp(s_{k,l} \,|\, \lambda^2/2) \times Gamma(\lambda^2 \,|\, r_\lambda, \delta) \times IW(\boldsymbol{M} \,|\, \boldsymbol{S}, \nu).$$

Next, we integrate w.r.t. λ^2 to obtain

$$p(\boldsymbol{u}_1,..,\boldsymbol{u}_V,\tau^2,\{s_{k,l}\}_{k<l},\boldsymbol{M}\,|\,\boldsymbol{y},\boldsymbol{X}) \propto \frac{1}{(\tau^2)^{n/2+1}|\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}'+\boldsymbol{I}|^{1/2}}\times$$
$$\exp\left\{-\frac{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{W})'(\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}'+\boldsymbol{I})^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{W})}{2\tau^2}\right\}\times\prod_{k=1}^{V}[\Delta N(\boldsymbol{u}_k\,|\,\boldsymbol{0},\boldsymbol{M})+(1-\Delta)\delta_{\boldsymbol{0}}]\times$$
$$\frac{1}{(\delta+\sum_{k<l}s_{k,l})^{q+r_\lambda}}\times IW(\boldsymbol{M}\,|\,\boldsymbol{S},\nu). \tag{11}$$

Note the fact that $\boldsymbol{D}$ is a diagonal matrix with all positive diagonal entries. Thus $\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}'$ is non-negative definite and by using Lemma 5.1, $\frac{1}{|\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}'+\boldsymbol{I}|^{1/2}}<1$. Further, using Lemma 5.2

$$\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}'+\boldsymbol{I}\leq\boldsymbol{X}\boldsymbol{X}'\sum_{k<l}s_{k,l}+\boldsymbol{I}\leq\left(\mu_{\boldsymbol{X}\boldsymbol{X}'}\sum_{k<l}s_{k,l}+1\right)\boldsymbol{I},$$

where $\boldsymbol{A}\leq\boldsymbol{B}$ implies $\boldsymbol{A}-\boldsymbol{B}$ is a non-negative definite matrix and $\mu_{\boldsymbol{X}\boldsymbol{X}'}$ is the largest eigenvalue of $\boldsymbol{X}\boldsymbol{X}'$. Using Lemma 5.3, the above inequality implies

$$(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{W})'(\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}'+\boldsymbol{I})^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{W})\geq\frac{||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{W}||^2}{\mu_{\boldsymbol{X}\boldsymbol{X}'}\sum_{k<l}s_{k,l}+1}.$$

Let

$$\tilde{p}(\boldsymbol{u}_1,..,\boldsymbol{u}_V,\tau^2,\{s_{k,l}\}_{k<l},\boldsymbol{M}) = \frac{1}{(\tau^2)^{n/2+1}|\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}'+\boldsymbol{I}|^{1/2}}\times$$
$$\exp\left\{-\frac{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{W})'(\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}'+\boldsymbol{I})^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{W})}{2\tau^2}\right\}\times\prod_{k=1}^{V}N(\boldsymbol{u}_k\,|\,\boldsymbol{0},\boldsymbol{M})\times$$
$$\frac{1}{(\delta+\sum_{k<l}s_{k,l})^{q+r_\lambda}}\times IW(\boldsymbol{M}\,|\,\boldsymbol{S},\nu). \tag{12}$$

With little algebra it can be shown that

$$p(\boldsymbol{u}_1,..,\boldsymbol{u}_V,\tau^2,\{s_{k,l}\}_{k<l},\boldsymbol{M}\,|\,\boldsymbol{y},\boldsymbol{X})$$
$$= constant\times\sum_{1\leq j_1,...,j_l\leq V,0\leq l\leq V}\Delta^l(1-\Delta)^{V-l}\tilde{p}(\boldsymbol{u}_{j_1},..,\boldsymbol{u}_{j_l},\boldsymbol{u}_{j_{l+1}}=0,..,\boldsymbol{u}_{j_V}=0,\tau^2,\{s_{k,l}\}_{k<l},\boldsymbol{M}).$$

Therefore, the integral of (11) w.r.t all parameters is finite if and only if $\int\tilde{p}(\boldsymbol{u}_1,..,\boldsymbol{u}_V,\tau^2,\{s_{k,l}\}_{k<l},\boldsymbol{M})<\infty$. Henceforth, we will proceed to show that this integral is finite.

With little algebra, we have that $\int IW(\boldsymbol{M}|\nu,\boldsymbol{S})\prod_{k=1}^{V}N(\boldsymbol{u}_k\,|\,\boldsymbol{0},\boldsymbol{M})d\boldsymbol{M}\propto\frac{1}{|\boldsymbol{S}+\sum_{k=1}^{V}\boldsymbol{u}_k\boldsymbol{u}_k'|^{(\nu+V)/2}}$.

Hence,

$$\tilde{p}(\boldsymbol{u}_1,..,\boldsymbol{u}_V,\tau^2,\{s_{k,l}\}_{k<l}) \leq \text{constant} \times \frac{1}{|\boldsymbol{S}+\sum_{k=1}^{V}\boldsymbol{u}_k\boldsymbol{u}_k'|^{(\nu+V)/2}}\frac{1}{(\tau^2)^{n/2+1}}\times$$
$$\exp\left\{-\frac{||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{W}||^2}{2\tau^2(\mu_{\boldsymbol{X}\boldsymbol{X}'}\sum_{k<l}s_{k,l}+1)}\right\}\times\frac{1}{(\delta+\sum\limits_{k<l}s_{k,l})^{q+r_\lambda}}\frac{1}{|\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}'+\boldsymbol{I}|^{1/2}}.$$

Define $\mathcal{A}=\{(\boldsymbol{u}_1,...,\boldsymbol{u}_V):||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{W}||^2>1\}$. Then

$$\int\tilde{p}(\boldsymbol{u}_1,..,\boldsymbol{u}_V,\tau^2,\{s_{k,l}\}_{k<l})d\boldsymbol{u}_1\cdots d\boldsymbol{u}_V d\tau^2 d\prod_{k<l}s_{k,l}=\int\limits_{\mathcal{A}}\tilde{p}(\boldsymbol{u}_1,..,\boldsymbol{u}_V,\tau^2,\{s_{k,l}\}_{k<l})d\boldsymbol{u}_1\cdots d\boldsymbol{u}_V d\tau^2 d\prod_{k<l}s_{k,l}+$$
$$\int\limits_{\mathcal{A}^c}\tilde{p}(\boldsymbol{u}_1,..,\boldsymbol{u}_V,\tau^2,\{s_{k,l}\}_{k<l})d\boldsymbol{u}_1\cdots d\boldsymbol{u}_V d\tau^2 d\prod_{k<l}s_{k,l}. \tag{13}$$

Now,

$$\int\limits_{\mathcal{A}}\tilde{p}(\boldsymbol{u}_1,..,\boldsymbol{u}_V,\tau^2,\{s_{k,l}\}_{k<l})d\tau^2 d\prod_{k<l}s_{k,l}d\boldsymbol{u}_1\cdots d\boldsymbol{u}_V \leq \text{constant}\int\frac{1}{|\boldsymbol{S}+\sum_{k=1}^{V}\boldsymbol{u}_k\boldsymbol{u}_k'|^{(\nu+V)/2}}\times$$
$$\frac{1}{(\tau^2)^{n/2+1}}\exp\left\{-\frac{1}{2\tau^2(\mu_{\boldsymbol{X}\boldsymbol{X}'}\sum_{k<l}s_{k,l}+1)}\right\}\times\frac{1}{(\delta+\sum\limits_{k<l}s_{k,l})^{q+r_\lambda}}\frac{1}{|\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}'+\boldsymbol{I}|^{1/2}}$$
$$\leq \text{constant}\left\{\int\frac{1}{|\boldsymbol{S}+\sum_{k=1}^{V}\boldsymbol{u}_k\boldsymbol{u}_k'|^{(\nu+V)/2}}d\boldsymbol{u}_1\cdots d\boldsymbol{u}_V\right\}\times\left\{\int\frac{(\mu_{\boldsymbol{X}\boldsymbol{X}'}\sum_{k<l}s_{k,l}+1)^{n/2}}{|\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}'+\boldsymbol{I}|^{1/2}(\delta+\sum\limits_{k<l}s_{k,l})^{q+r_\lambda}}d\prod_{k<l}s_{k,l}\right\}.$$

Use the fact that $q>n$ in all our applications to conclude that $\int\frac{(\mu_{\boldsymbol{X}\boldsymbol{X}'}\sum_{k<l}s_{k,l}+1)^{n/2}}{(\delta+\sum\limits_{k<l}s_{k,l})^{q+r_\lambda}}d\prod_{k<l}s_{k,l}=c<\infty$. Similarly,

$$\int\frac{1}{|\boldsymbol{S}+\sum_{k=1}^{V}\boldsymbol{u}_k\boldsymbol{u}_k'|^{(\nu+V)/2}}d\boldsymbol{u}_1\cdots d\boldsymbol{u}_V\leq\int\frac{1}{\prod_{k=1}^{V}|\boldsymbol{S}+\boldsymbol{u}_k\boldsymbol{u}_k'|^{\nu/V+1/2}}d\boldsymbol{u}_1\cdots d\boldsymbol{u}_V$$
$$\prod_{k=1}^{V}\left(\int\frac{1}{|\boldsymbol{S}+\boldsymbol{u}_k\boldsymbol{u}_k'|^{2\nu/V+1}}d\boldsymbol{u}_k\right)^{1/2},$$

where the first inequality follows from the fact that $|\boldsymbol{S}+\sum_{k=1}^{V}\boldsymbol{u}_k\boldsymbol{u}_k'|\geq|\boldsymbol{S}+\boldsymbol{u}_k\boldsymbol{u}_k'|$ for all k. The second inequality is a direct application of the Cauchy-Schwartz inequality. By the ratio test of integrals, this integral is finite if $\int\frac{1}{[(1+u_{k,1})^2\cdots(1+u_{k,R})^2]^{2\nu/V+1}}d\boldsymbol{u}_k$ is finite. Now use the fact that $\int\frac{1}{x^{1+c}}dx<\infty$ for any $c>0$ to argue that $\int\frac{1}{[(1+u_{k,1})^2\cdots(1+u_{k,R})^2]^{2\nu/V+1}}d\boldsymbol{u}_k$ is finite. Hence, $\int\limits_{\mathcal{A}}\tilde{p}(\boldsymbol{u}_1,..,\boldsymbol{u}_V,\tau^2,\{s_{k,l}\}_{k<l})d\boldsymbol{u}_1\cdots d\boldsymbol{u}_V d\tau^2 d\prod\limits_{k<l}s_{k,l}\leq\infty$.

Now consider the expression $\int\limits_{\mathcal{A}} \tilde{p}(\boldsymbol{u}_1, .., \boldsymbol{u}_V, \tau^2, \{s_{k,l}\}_{k<l})$. It is easy to see that $\{(\boldsymbol{u}_1, ..., \boldsymbol{u}_V) : ||\boldsymbol{y} - \boldsymbol{XW}||^2 \leq 1\}$ is a bounded set, so that the bounded function $\exp\left\{-\frac{||\boldsymbol{y}-\boldsymbol{XW}||^2}{2\tau^2(\mu_{\boldsymbol{XX'}}\sum_{k<l} s_{k,l}+1)}\right\}$ achieves the maximum value at $\boldsymbol{W} = \boldsymbol{W}^*$. Thus,

$$\begin{aligned}
&\int\limits_{\mathcal{A}} \tilde{p}(\boldsymbol{u}_1, .., \boldsymbol{u}_V, \tau^2, \{s_{k,l}\}_{k<l}) \leq \text{constant} \int \frac{1}{|\boldsymbol{S} + \sum_{k=1}^{V} \boldsymbol{u}_k \boldsymbol{u}_k'|^{(\nu+V)/2}} \frac{1}{(\tau^2)^{n/2+1}} \times \\
&\qquad \exp\left\{-\frac{||y - X\boldsymbol{W}^*||^2}{2\tau^2(\mu_{\boldsymbol{XX'}}\sum_{k<l} s_{k,l} + 1)}\right\} \times \frac{1}{(\delta + \sum\limits_{k<l} s_{k,l})^{q+r_\lambda}} \frac{1}{|\boldsymbol{XDX'} + \boldsymbol{I}|^{1/2}} \\
&\leq \frac{\text{constant}}{||y - X\boldsymbol{W}^*||^n} \left\{\int \frac{1}{|\boldsymbol{S} + \sum_{k=1}^{V} \boldsymbol{u}_k \boldsymbol{u}_k'|^{(\nu+V)/2}} d\boldsymbol{u}_1 \cdots d\boldsymbol{u}_V\right\} \times \\
&\qquad \left\{\int \frac{(\mu_{\boldsymbol{XX'}}\sum_{k<l} s_{k,l} + 1)^{n/2}}{|\boldsymbol{XDX'} + \boldsymbol{I}|^{1/2}(\delta + \sum\limits_{k<l} s_{k,l})^{q+r_\lambda}} d\prod_{k<l} s_{k,l}\right\} \\
&< \infty,
\end{aligned}$$

where the last step follows from earlier discussions.

References

Armagan, A., Dunson, D. B., and Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica*, **23**(1), 119–143.

Bullmore, E. and Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews. Neuroscience*, **10**(3), 186–198.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, **97**(2), 465–480.

Chatterjee, A. and Lahiri, S. (2010). Asymptotic properties of the residual bootstrap for lasso estimators. *Proceedings of the American Mathematical Society*, **138**(12), 4497–4509.

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

Chavez, R., Graff-Guerrero, A., Garcia-Reyna, J., Vaugier, V., and Cruz-Fuentes, C. (2004). Neurobiology of creativity: Preliminary results from a brain activation study. *Salud Mental*, **27**(3), 38–46.

Christakis, N. A. and Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *n engl j med*, **2007**(357), 370–379.

Craddock, R. C., Holtzheimer, P. E., Hu, X. P., and Mayberg, H. S. (2009). Disease state prediction from resting state functional connectivity. *Magnetic Resonance in Medicine*, **62**(6), 1619–1628.

De la Haye, K., Robins, G., Mohr, P., and Wilson, C. (2010). Obesity-related behaviors in adolescent friendship networks. *Social Networks*, **32**(3), 161–167.

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.

Doreian, P. (2001). Causality in social network analysis. *Sociological Methods & Research*, **30**(1), 81–114.

Durante, D., Dunson, D. B., *et al.* (2017). Bayesian inference and testing of group differences in brain networks. *Bayesian Analysis*.

Erdos, P. and Rényi, A. (1960). On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, **5**(1), 17–60.

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

Flaherty, A. W. (2005). Frontotemporal and dopaminergic control of idea generation and creative drive. *Journal of Comparative Neurology*, **493**(1), 147–153.

Fosdick, B. K. and Hoff, P. D. (2015). Testing and modeling dependencies between a network and nodal attributes. *Journal of the American Statistical Association*, **110**(511), 1047–1056.

Fowler, J. H. and Christakis, N. A. (2008). Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. *British Medical Journal*, **337**, a2338.

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

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**(1), 1–22.

Gramacy, R. B. and Gramacy, M. R. B. (2013). `R` package `monomvn`.

Guhaniyogi, R. and Rodriguez, A. (2017). Joint modeling of longitudinal relational data and exogenous variables. *https://www.soe.ucsc.edu/sites/default/files/technical-reports/UCSC-SOE-17-17.pdf*.

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

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.

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

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

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.

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

Kyung, M., Gill, J., Ghosh, M., Casella, G., *et al.* (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, **5**(2), 369–411.

Lin, X. (2010). Identifying peer effects in student academic achievement by spatial autoregressive models with group unobservables. *Journal of Labor Economics*, **28**(4), 825–860.

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.

Niezink, N. M. K. and Snijders, T. A. B. (2016). Co-evolution of social networks and continuous actor attributes.

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.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, **103**(482), 681–686.

Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian Statistics*, **9**, 501–538.

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., and Taylor, S. F. (2017). Network classification with applications to brain connectomics. *arXiv preprint arXiv:1701.08140*.

Richiardi, J., Eryilmaz, H., Schwartz, S., Vuilleumier, P., and Van De Ville, D. (2011). Decoding brain states from fMRI connectivity graphs. *Neuroimage*, **56**(2), 616–626.

Shalizi, C. R. and Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological methods & research*, **40**(2), 211–239.

Shoham, D. A., Hammond, R., Rahmandad, H., Wang, Y., and Hovmand, P. (2015). Modeling social norms and social influence in obesity. *Current Epidemiology Reports*, **2**(1), 71–79.

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.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), 267–288.

Watts, D. J. and Dodds, P. (2009). Threshold models of social influence. *The Oxford Handbook of Analytical Sociology*, pages 475–497.

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