Bayesian Analysis (2023) **18**, Number 4, pp. 1131–1160

High-Dimensional Bayesian Network Classification with Network Global-Local Shrinkage Priors

Sharmistha Guha[*] and Abel Rodriguez[†]

Abstract. This article proposes a novel Bayesian binary classification framework for networks with labeled nodes. Our approach is motivated by applications in brain connectome studies, where the overarching goal is to identify both regions of interest (ROIs) in the brain and connections between ROIs that influence how study subjects are classified. We propose a novel binary logistic regression framework with the network as the predictor, and model the associated network coefficient using a novel class of *global-local* network shrinkage priors. We perform a theoretical analysis of a member of this class of priors (which we call the Network Lasso Prior) and show asymptotically correct classification of networks even when the number of network edges grows faster than the sample size. Two representative members from this class of priors, the Network Lasso prior and the Network Horseshoe prior, are implemented using an efficient Markov Chain Monte Carlo algorithm, and empirically evaluated through simulation studies and the analysis of a real brain connectome dataset.

Keywords: brain connectome, high-dimensional binary regression, global-local shrinkage prior, node selection, network predictor, posterior consistency.

1 Introduction

Statistical models for the analysis of individual networks have received substantial attention in the literature. 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; Guhaniyogi and Rodriguez, 2020; Sosa and Rodríguez, 2021) and stochastic block models (Nowicki and Snijders, 2001; Rodriguez, 2012). However, there are many important applications in which network data is available for every individual in the sample, and interest lies in using such networks as predictors to explain an outcome of interest, rather than understanding the underlying process that drives network formation. Section 6 in this article presents one such example from a brain connectome study. In this study, brain connectome networks are available for multiple individuals who are classified as subjects with high or low IQ (Intelligence Quotient). To construct the networks, the human brain has been divided according to the Desikan Atlas (Desikan et al., 2006) that identifies 34 cortical regions of interest (ROIs), both in the left and the 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

[*] Department of Statistics, Texas A&M University, College Station, TX 77843, sharmistha@tamu.edu

[†] Department of Statistics, University of Washington, Seattle, WA 98195, abelrod@uw.edu

https://doi.org/10.1214/23-BA1378

whose rows and columns correspond to the different ROIs (shared among networks for all individuals) and whose entries correspond to estimates of the number of *fibers* connecting pairs of brain regions. The scientific goals in this setting are (1) to develop a predictive rule for classifying a new subject as having low or high IQ based on his/her observed brain network, and (2) to identify influential brain regions (nodes in the brain network) as well as significant connections between different brain regions (links in the network) that are predictive of IQ.

Much of the early literature on network and graph classification was motivated by the problem of classification of chemical compounds, where a graph represents a compound's molecular structure. In such analyses, discriminative patterns in a graph were identified and used as features for training a standard classification method (e.g., see Srinivasan et al., 1996, Helma et al., 2001, Deshpande et al., 2005 and Fei and Huan, 2010). Similar approaches, based on summary measures such as average degree, clustering coefficient and average path length, have also been employed in the context of neuroscience applications (e.g., see Bullmore and Sporns, 2009, Olde Dubbelink et al., 2013, Daianu et al., 2013 and references therein). Alternative approaches use kernels defined on graph spaces to construct similarity metrics between two networks (e.g., see Vishwanathan et al., 2010), which in turn can be used to build, for example, nearest-neighbor classifiers. While all these methods scale well with the size of the network, it is our experience that the choice of features/kernels (which is typically ad-hoc) has a dramatic influence on the results. Furthermore, these type of methods typically do not allow for a full exploration of which nodes/edges are influential on the responses. Other relevant references in this area include Vogelstein et al. (2013), who propose to look for a minimal set of nodes which best explains the difference between two groups of networks, and Durante et al. (2018), who propose a high-dimensional Bayesian tensor factorization model for a population of networks that allows to test for local edge differences between two groups of subjects. Both of these approaches tend to focus mainly on classification and are not designed to detect important nodes and edges impacting the response.

An alternate approach to constructing classifiers based on network data is to vectorize the network predictor and treat the edge weights as a long vector of predictors (e.g., see Richiardi et al., 2011, Craddock et al., 2009 and Zhang et al., 2012) in a binary regression setting. This approach can take advantage of recent developments in high-dimensional regression, either penalized likelihood estimation (e.g., see Tibshirani, 1996) or Bayesian shrinkage (e.g., see Park and Casella, 2008, Carvalho et al., 2010, Armagan et al., 2013a, and Du and Ghosal, 2018). However, such an approach treats the links of the network as exchangeable, ignoring the fact that coefficients involving common nodes can be expected to be correlated a-priori. In other words, it does not capture the fact that we expect, a priori, higher correlation among coefficients that share one node in common. Being agnostic to the structure of the network predictor, the ordinary shrinkage priors are not well adapted to node-level (as opposed to edge-level) inference.

This article develops a high-dimensional Bayesian network classifier that can identify both influential nodes and specific edges impacting classification. To achieve this goal, we formulate a high-dimensional logistic regression model with the binary response regressed on the network predictor. While the model can be represented as a linear function of the network data, we carefully add structure to the network coefficient and design prior distributions to exploit the network topology and draw inference on influential network nodes. More concretely, the coefficients associated with the network predictor are assigned a prior from the class of *network global-local shrinkage priors*. In the context of linear regression, Guha and Rodriguez (2021) develop a special case of the network shrinkage prior, known as the *Network Lasso prior*. This article generalizes the earlier approach to develop a broader class of network shrinkage priors. In particular, this article introduces another member of this broad class of priors, called the *Network Horseshoe prior*, and presents a comparative study of the empirical performance of these two priors in various simulation settings and in the brain network data. Another related approach has been proposed by Relión et al. (2019), who develop a method that relies on a combination of regular and group Lasso penalties within a logistic regression framework to carry out network classification. One key advantage of our Bayesian approach over this penalized likelihood method is our ability to fully quantify the uncertainty associated with our estimates.

A major contribution of this article is a careful study of the asymptotic properties of the resulting binary network classification framework. Theory for posterior contraction for high-dimensional regression models has gained traction over the last 10 years. For example, Castillo et al. (2012) and Belitser and Nurushev (2015) have established posterior concentration and variable selection properties for certain point-mass priors in linear regression models. The latter article also establishes asymptotically nominal coverage of Bayesian credible sets (see also Castillo et al., 2015 and Martin et al., 2017). In the same thread, Jeong and Ghosal (2021) develop posterior contraction properties for sparse generalized linear models with variable selection priors on coefficients. In contrast, the literature on posterior contraction properties for high-dimensional Bayesian shrinkage priors is less well developed. Armagan et al. (2013b) were the first to show posterior consistency in the ordinary linear regression model with shrinkage priors for low-dimensional settings under the assumption that the number of covariates *does not* exceed the number of observations. Using direct calculations, Van Der Pas et al. (2014) have shown that the posterior based on the ordinary horseshoe prior concentrates at the optimal rate for normal-mean problems. More recently, Song and Liang (2017) consider a general class of global-local continuous shrinkage priors and obtain posterior contraction rates in ordinary high-dimensional linear regression models. Bai and Ghosh (2018), and later Zhang and Ghosh (2019), have developed posterior contraction theory for global-local shrinkage priors in the context of multivariate regression frameworks. An insightful review on properties of the posterior distribution under global-local shrinkage priors is available in Bhadra et al. (2019). Notably, the posterior contraction literature for global-local shrinkage priors on binary logistic regression is limited, with a few exceptions, such as Wei and Ghosal (2020). In this article we develop posterior contraction theory for the Network Lasso prior. Our development requires considerably different techniques and results compared to Wei and Ghosal (2020). In fact, developing the theory for network classification with the Network Lasso prior proposed in this article faces two major challenges. First, our prior has a low-rank structure in the prior mean of the edge coefficients, based on node specific latent variables (Guha and Rodriguez, 2021), as well as an additional shrinkage structure through the variance. Second, we introduce a variable selection prior on the node-specific latent variables. The combination of these two structures adds substantial theoretical challenges over and above Wei and Ghosal (2020), so that different proof techniques are required (see Section 1 of the Supplementary Material (Guha and Rodriguez, 2023)). Second, we aim to prove a challenging but practically desirable result of asymptotically correct classification when the number of edges in the network predictor grows at a super-linear rate as a function of the sample size. Both these features present obstacles which we overcome in this work. The theoretical results provide insights on how the number of nodes in the network predictor, the dimensions of node-specific latent variables, and the structure and sparsity in the true network predictor coefficients can vary with the sample size n to achieve asymptotically correct classification.

The remainder of the manuscript is organized as follows: Section 2 develops the model and the prior distributions. Section 3 discusses theoretical developments justifying the asymptotically desirable prediction from the proposed model. Section 4 details posterior computation. Results from various simulation experiments and a brain connectome data analysis have been presented in Sections 5 and 6, respectively. Finally, Section 7 concludes the article with a brief discussion of the proposed methodology.

2 Model Framework

2.1 Notation

For $i = 1, \ldots, n$, let $\boldsymbol{A}_i \in \mathbb{R}^{V \times V}$ denote the weighted undirected network predictor with V nodes, and $y_i \in \{0, 1\}$ be the binary response corresponding to the i-th individual. The entry $a_{i,k,l} \in \mathbb{R}$ of $\boldsymbol{A}_i$ indicates the strength of association between the k-th and l-th nodes of the network. Our framework allows entries of $\boldsymbol{A}_i$ to be continuous, binary or count. In this paper, we focus on networks that have no self relationships, i.e., $a_{i,k,l} \equiv 0$ when $k = l$, and are undirected ($a_{i,k,l} = a_{i,l,k}$), but generalizations to directed networks are straightforward.

2.2 Model Formulation

In the context of network classification, we propose a high-dimensional logistic regression model of the binary response $y_i \in \{0, 1\}$ on the undirected network predictor $\boldsymbol{A}_i$ as

$$y_i \sim Ber\left(G(\psi_i)\right), \qquad \psi_i = \mu + \langle \boldsymbol{A}_i, \boldsymbol{\Gamma} \rangle_F, \tag{2.1}$$

where $G : \mathsf{R} \to [0, 1]$ is a link function and $\boldsymbol{\Gamma}$ is a $V \times V$ symmetric network coefficient matrix whose (k, l)-th element is given by $\gamma_{k,l}/2$, with $\gamma_{k,k} = 0$, for all $k = 1, \ldots, V$. The notation $\langle \boldsymbol{A}_i, \boldsymbol{\Gamma} \rangle_F$ denotes the Frobenius inner product between the two matrices $\boldsymbol{A}_i$ and $\boldsymbol{\Gamma}$, defined as the sum of the element-wise product between the two matrices. For conciseness, we focus on logistic models in this paper, where $G(z) = \left[1 + \exp\{-z\}\right]^{-1}$, but our prior formulation can be easily adapted to other link functions.

Model (2.1) can be expressed in the form of a generalized linear model. To be more specific, $\langle \boldsymbol{A}_i, \boldsymbol{\Gamma} \rangle_F = \sum\limits_{1 \le k < l \le V} a_{i,k,l} \gamma_{k,l}$, due to the symmetry and zero diagonal entries in $\boldsymbol{A}_i$ and $\boldsymbol{\Gamma}$, so that

$$\psi_i = \mu + \sum_{1\leq k<l\leq V} a_{i,k,l}\gamma_{k,l}.$$

Furthermore, note that, if $\boldsymbol{x}_i = (a_{i,1,2}, \ldots, a_{i,(V-1),V})' \in \mathbb{R}^{V(V-1)/2}$ is the collection of all upper triangular elements of $\boldsymbol{A}_i$, and $\boldsymbol{\gamma} = (\gamma_{1,2}, \ldots, \gamma_{(V-1),V})' \in \mathbb{R}^{V(V-1)/2}$ is the vector of corresponding upper triangular elements of $2\boldsymbol{\Gamma}$, then the linear predictor can be written as $\psi_i = \mu + \boldsymbol{x}_i'\boldsymbol{\gamma}$.

Specifying an ordinary high-dimensional shrinkage prior on $\boldsymbol{\gamma}$, such as the ones described in Carvalho et al. (2010); Armagan et al. (2013a); Park and Casella (2008), is unsatisfactory in our context since (a) it loses information on network nodes and makes inference on influential network nodes (with uncertainty) difficult; and (b) it does not capture the fact that we expect, a priori, higher correlation among coefficients that share one node in common. These shortcomings motivate the development of the priors in the next Section.

2.3 Network Global-Local Shrinkage Priors

Let $\boldsymbol{u}_1, .., \boldsymbol{u}_V$ be a collection of R-dimensional latent variables, one for each of the V nodes in the network. In this article, we consider a $N(0, 1)$ prior for the common intercept parameter μ and propose a general class of network global-local shrinkage priors on the edge coefficients $\gamma_{k,l}$'s given by,

$$\gamma_{k,l} \mid s_{k,l}, \sigma^2 \sim N(\boldsymbol{u}_k'\boldsymbol{\Lambda}\boldsymbol{u}_l, \sigma^2 s_{k,l}^2), \qquad \sigma \sim H_1(\cdot), \qquad s_{k,l} \sim H_2(\cdot), \tag{2.2}$$

where $\boldsymbol{\Lambda} = diag(\lambda_1, \ldots, \lambda_R)$ is an $R\times R$ diagonal matrix, with the r-th diagonal entry λ_r. In this framework, $\mathrm{E}(\boldsymbol{\Gamma}) = \boldsymbol{U}'\boldsymbol{\Lambda}\boldsymbol{U}$, where $\boldsymbol{U}$ is a matrix whose k-th column corresponds to $\boldsymbol{u}_k$. This representation, related to the eigenvalue decomposition of $\boldsymbol{\Gamma}$, provides an embedding into an $R \leq V$ dimensional latent space. This kind of embedding has been widely used to construct sparse models for network data that can capture common properties such as transitivity (Hoff, 2005). Furthermore, the multiplicative structure associated with the prior variance allows for both global (controlled by σ^2) and local (controlled by the $s_{k,l}$'s) shrinkage effects.

The formulation in (2.2) leads to a wide variety of network shrinkage priors by choosing different forms for $H_1(\cdot)$ and $H_2(\cdot)$. For example, setting $H_1(\sigma)$ as a point mass at 1 and $H_2(s_{k,l}^2)$ as an exponential distribution, $s_{k,l}^2 \sim Exp(\theta/2)$ with $\theta \sim Gamma(\zeta, \iota)$, corresponds to the *Network Lasso prior* discussed in Guha and Rodriguez (2021). Alternatively, setting both H_1 and H_2 to be half-Cauchy distributions, leads to what we call the *Network Horseshoe prior* (see Carvalho et al., 2010).

In order to identify which nodes are actively related to the response, we assign a zero-inflated Gaussian prior on the latent factors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_V$

$$\boldsymbol{u}_k \mid \xi_k, \boldsymbol{M} \sim \xi_k N(\boldsymbol{0}, \boldsymbol{M}) + (1-\xi_k)\delta_{\boldsymbol{0}}, \qquad \boldsymbol{M} \sim IW(\nu, \boldsymbol{I}), \tag{2.3}$$

where $\xi_k \mid \Delta \sim Ber(\Delta)$ and $\Delta \sim Beta(a_\Delta, b_\Delta)$. Here, $\delta_{\mathbf{0}}$ is the Dirac-delta function at $\mathbf{0}$, $\boldsymbol{M}$ is a $R \times R$ covariance matrix and $IW(\nu, \boldsymbol{I})$ denotes an inverse-Wishart distribution with parameters ν and $\boldsymbol{I}$. The parameter Δ corresponds to the prior probability of the nonzero $\boldsymbol{u}_k$. Note that, if the k-th node of the network predictor is inactive in predicting the response, then a-posteriori ξ_k should assign high probability to 0. Thus, we can use the posterior probability of the event $\{\xi_k = 0\}$ to identify influential nodes. Assigning a prior distribution to Δ ensures multiplicity correction in the simultaneous selection of multiple $\boldsymbol{u}_k$'s (Scott and Berger, 2010).

An important aspect of these models is the selection of the model rank R. In order to learn the effective dimension of the latent space in which the matrix of coefficients is being embedded, we employ a hierarchical prior of the form

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

Note that, if $\lambda_r = 0$, the r-th component of the vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_V$ has no effect on the value of the regression coefficients. Hence, $R_{eff} = \sum_{r=1}^{R} \lambda_r \leq R$ can be interpreted as the *effective dimensionality* of the model. In this context, the choice of hyper-parameters of the beta distribution is crucial. In particular, note that $E[\lambda_r] = 1/(1 + r^\eta) \to 0$ as $r \to \infty$ and that $\sum_{r=1}^{R} var(\lambda_r) = \sum_{r=1}^{R} \frac{r^\eta}{(1+r^\eta)^2(2+r^\eta)} < \infty$ as $R \to \infty$. The first property provides (weak) identifiability of the different latent dimensions, while the second ensures that $\lim_{R\to\infty} var(\boldsymbol{u}_k) < \infty$ as long as the prior for the $\boldsymbol{u}_k$'s has a finite second moment. For our empirical investigations in Sections 5 and 6, we set $\eta = 1.1$.

Under the previous formulation, we can think about selecting R as similar to selecting the truncation level of a large dimensional model. As long as R is "large enough", the results should be robust to our choice (and further increasing R would lead to negligible changes). Natural analogies are the truncation of a Dirichlet process mixture model which results in an (approximate) finite mixture model, and the truncation of the stick-breaking construction of the Indian Buffet process (Teh et al., 2007). In our illustrations, we perform sensitivity analyses to determine an optimal value of R that balances computational efficiency and inferential accuracy, noting that further increases in R beyond the optimal value do not lead to any significant change in model performance. Along with R, sensitivity analyses regarding the choice of hyper-parameters $\iota, \zeta, a_\Delta, b_\Delta$ and ν are also performed and recorded in the simulation studies.

3 Posterior Contraction of the Binary Network Classification Model

This section establishes convergence results for (2.1) under the Network Lasso shrinkage prior given by $\gamma_{k,l} \mid s_{k,l} \sim N(\boldsymbol{u}_k' \boldsymbol{\Lambda} \boldsymbol{u}_l, s_{k,l}^2)$, $s_{k,l}^2 \sim Exp(\theta_n/2)$. For the theoretical study, a common practice is to fix θ_n as a function of n (Armagan et al., 2013a). Our theoretical investigations fix this function, with the exact expression given in Condition (6) in Section 3.2.

In our analysis, we consider an asymptotic setting in which the number of nodes in the network predictor, V_n, grows with the sample size n. This framework attempts to capture the fact that the number of coefficients, $q_n = \frac{V_n(V_n-1)}{2}$, will typically be much larger than n. This creates theoretical challenges, which are related to (but distinct from) those faced in showing posterior consistency for high-dimensional continuous (Armagan et al., 2013a) and binary regressions (Wei and Ghosal, 2020).

3.1 True Data Distribution and Assumption for the True Network Coefficient

Let $\boldsymbol{y}_n = (y_1, \ldots, y_n)'$ be the vector of observations. Using the subscript 0 to indicate the true parameter values, the data generating model is assumed to be

$$y_i \sim Ber\left(\frac{\exp\{\psi_{i,0}\}}{1+\exp\{\psi_{i,0}\}}\right), \qquad \psi_{i,0} = \mu_0 + \langle \boldsymbol{A}_i, \boldsymbol{\Gamma}_0\rangle_F, \tag{3.1}$$

where μ_0 is the true intercept in the network predictor and $\boldsymbol{\Gamma}_0$ is the true symmetric network coefficient matrix. We assume that $\boldsymbol{\Gamma}_0$ can be represented by the sum of a symmetric low-dimensional matrix and a symmetric sparse matrix. Thus, we assume that $\gamma_{k,l,0} = \boldsymbol{u}_{k,0}'\boldsymbol{\Lambda}\boldsymbol{u}_{l,0} + \vartheta_{k,l,0}$, where $\boldsymbol{u}_{k,0}$ is a R_0 dimensional vector, $k = 1, \ldots, V_n$. Also, $\boldsymbol{\vartheta}_0$ is the vector of all $\vartheta_{k,l,0}$, $k < l$, and we denote the number of nonzero elements of $\boldsymbol{\vartheta}_0$ by $h_{n,0}$, i.e., $||\boldsymbol{\vartheta}_0||_0 = h_{n,0}$.

For any $\epsilon > 0$, define $\mathcal{A}_n = \left\{(\mu, \boldsymbol{\Gamma}) : \frac{1}{n}\sum_{i=1}^{n} |p_{\mu,\boldsymbol{\Gamma}}(y_i = 1) - p_{\mu_0,\boldsymbol{\Gamma}_0}(y_i = 1)| \leq \epsilon\right\}$ as a neighborhood around the true density. This neighborhood construction has been used in Ghosal et al. (2006) in the context of density estimation in nonparametric binary regression with Gaussian processes. Further, let $\pi_n(\cdot)$ and $\Pi_n(\cdot)$ be the prior and posterior densities of $(\mu, \boldsymbol{\Gamma})$ based on $\boldsymbol{y}_n$, such that

$$\Pi_n(\mathcal{A}_n^c) = \frac{\int_{\mathcal{A}_n^c} p_{\mu,\boldsymbol{\Gamma}}(\boldsymbol{y}_n)\pi_n(\mu, \boldsymbol{\Gamma})d\mu d\boldsymbol{\Gamma}}{\int p_{\mu,\boldsymbol{\Gamma}}(\boldsymbol{y}_n)\pi_n(\mu, \boldsymbol{\Gamma})d\mu d\boldsymbol{\Gamma}},$$

where $p_{\mu,\boldsymbol{\Gamma}}(\boldsymbol{y}_n)$ denotes the likelihood of the n-dimensional response vector $\boldsymbol{y}_n$.

3.2 Main Results

To show the posterior contraction results, we generally follow Wei and Ghosal (2020) and Armagan et al. (2013a), but with substantial modifications required due to the nature of our proposed network lasso prior distribution. In proving the results, we make a couple of simplifications. First, it is assumed that the dimension R_n of $\boldsymbol{u}_k$ is the same as $R_{0,n}$, the dimension of $\boldsymbol{u}_{k,0}$, and that both of them increase with n. Consequently, the *effective dimensionality* of the latent space does not need to be estimated and $\boldsymbol{\Lambda} = \boldsymbol{I}$ is a non-random matrix. Second, we assume that $\boldsymbol{M}$ is non-random and $\boldsymbol{M} = \boldsymbol{I}$. Finally, we set hyper-parameters $a_\Delta = b_\Delta = 1$ in proving our theoretical results. We emphasize that these assumptions are not essential for the contraction result to be true, and are only introduced in order to simplify the derivations.

For two sequences $\{C_{1,n}\}_{n\geq 1}$ and $\{C_{2,n}\}_{n\geq 1}$, $C_{1,n} = o(C_{2,n})$ if $C_{1,n}/C_{2,n} \to 0$, as $n \to \infty$. The following theorem shows contraction of the posterior asymptotically under mild sufficient conditions on V_n and $h_{n,0}$. The proof of the theorem is provided in Section 1 of the Supplementary Material (Guha and Rodriguez, 2023).

Theorem 3.1. *Assume*

1. $\sup_{r=1,..,R_n;k=1,..,V_n} |u_{k,r,0}| < M_u < \infty$;
2. $R_n V_n = o\left(\frac{n}{\log(n)}\right)$; *$V_n \to \infty$ as $n \to \infty$.*
3. $\|\boldsymbol{A}_i\|_\infty$ *is bounded for all* $i = 1,..,n$, *w.l.o.g., assume* $||\boldsymbol{A}_i||_\infty \leq 1$.
4. $h_{n,0}\log(q_n) = o(n)$
5. $\|\boldsymbol{\vartheta}_0\|_\infty$ *is bounded; w.l.g.* $\|\boldsymbol{\vartheta}_0\|_\infty < 1$
6. $\theta_n = \frac{C}{q_n n^{\rho/2}\log(n)}$ *for some $C > 0$ and some $\rho \in (1, 2)$.*

Under assumptions (1)-(6) for the Network Lasso prior on $\boldsymbol{\Gamma}$, $\Pi_n(\mathcal{A}_n^c) \to 0$ *as* $n \to \infty$, *for any* $\epsilon > 0$.

Conditions (1), (3) and (5) are technical conditions ensuring that each of the entries in the true network coefficient and the network predictor are bounded. Condition (2) puts an upper bound on the growth of the number of network nodes and dimensions of node specific latent variables vis-a-vis the sample size to achieve asymptotically correct classification of networks. Similarly, (4) puts a restriction on the number of nonzero elements of $\boldsymbol{\vartheta}_0$ with respect to n.

The proof bears some connections with the proofs of results in Wei and Ghosal (2020), but is significantly different from it, mainly due to the introduction of network shrinkage priors. Notably, the proof of Theorem 3.1 is built upon Lemma A.1 and Lemma A.2 (see Section 1 of Supplementary Material (Guha and Rodriguez, 2023)), where Lemma A.2 is a new result specifically needed to address the network lasso prior structure in our framework. Additionally, as part of the proof of Theorem 3.1, we needed to establish that the posterior probability $-\log\Pi(||\boldsymbol{W} - \boldsymbol{W}_0||_\infty < \frac{\eta_1}{2n^{\rho/2}})$ grows sublinearly with the sample size n, for a constant $\eta_1 > 0$, where $\boldsymbol{W} = (\boldsymbol{u}_1'\boldsymbol{u}_2, \ldots, \boldsymbol{u}_{V_n-1}'\boldsymbol{u}_{V_n})'$, $\boldsymbol{W}_0 = (\boldsymbol{u}_{1,0}'\boldsymbol{u}_{2,0}, \ldots, \boldsymbol{u}_{V_n-1,0}'\boldsymbol{u}_{V_n,0})'$, $\boldsymbol{u}_k = (u_{k,1}, \ldots, u_{k,R_n})'$ and $\boldsymbol{u}_k = (u_{k,1,0}, \ldots, u_{k,R_n,0})'$, for $k = 1,..,V_n$. This result is also novel and specific to our construction of the Bayesian network lasso shrinkage prior. While the other parts in the proof of Theorem 3.1 exploit known techniques, they have been tailored specific to the result. The next few sections show empirical performance of the proposed class of priors.

4 Posterior Computation

We have implemented Markov chain Monte Carlo (MCMC) samplers for posterior inference for both the Network Lasso and Network Horseshoe shrinkage priors on $\boldsymbol{\Gamma}$. We use the result discussed in Theorem 1 of Polson et al. (2013) to obtain

$$\begin{aligned} p(y_i|\boldsymbol{A}_i,\boldsymbol{\Gamma}) &= \frac{\exp\{y_i(\mu+\langle\boldsymbol{A}_i,\boldsymbol{\Gamma}\rangle_F)\}}{1+\exp\{\mu+\langle\boldsymbol{A}_i,\boldsymbol{\Gamma}\rangle_F\}} \\ &= \exp\{(y_i-0.5)(\mu+\langle\boldsymbol{A}_i,\boldsymbol{\Gamma}\rangle_F)\}\int\exp(-\omega_i(\mu+\langle\boldsymbol{A}_i,\boldsymbol{\Gamma}\rangle_F)^2/2)p(\omega_i)d\omega_i, \end{aligned}$$

where $p(\omega_i)$ is the density for PG(1,0) distribution. We exploit this result and rely on data augmentation approach as outlined in Polson et al. (2013) to introduce variables $\omega_1,\ldots,\omega_n$ in the likelihood and write the likelihood function associated with our model as

$$\begin{aligned} p(\boldsymbol{y}\,|\,\boldsymbol{A}_1,..,\boldsymbol{A}_n,\boldsymbol{\Gamma},\boldsymbol{\omega}) &\propto \prod_{i=1}^{n} p(y_i\,|\,\boldsymbol{A}_i,\boldsymbol{\Gamma},\omega_i) \\ &\propto \prod_{i=1}^{n}\exp\left\{(y_i-0.5)(\mu+\langle\boldsymbol{A}_i,\boldsymbol{\Gamma}\rangle_F)-\omega_i(\mu+\langle\boldsymbol{A}_i,\boldsymbol{\Gamma}\rangle_F)^2/2\right\} \\ &\propto \prod_{i=1}^{n}\exp\left\{-\frac{\omega_i}{2}\left[\frac{(y_i-0.5)}{\omega_i}-(\mu+\langle\boldsymbol{A}_i,\boldsymbol{\Gamma}\rangle_F)\right]^2\right\}. \end{aligned}$$

Hence, while the original conditional posterior distributions for the parameters are not available in closed forms, the augmented full conditional distributions mostly belong to standard families. Sections 2 and 3 in the supplementary material provide details of the algorithms for the Network Lasso and Network Horseshoe priors on $\boldsymbol{\gamma}$, respectively.

Let $\boldsymbol{\Gamma}^{(1)},\ldots,\boldsymbol{\Gamma}^{(L)}$ and $\mu^{(1)},\ldots,\mu^{(L)}$ be the L MCMC samples for $\boldsymbol{\Gamma}$ and μ, respectively, obtained after suitable burn-in and thinning. To classify a newly observed network $\boldsymbol{A}_*$ as a member of one of the two groups, we compute

$$S^{(l)} = \frac{\exp(\mu^{(1)}+\langle\boldsymbol{A}_*,\boldsymbol{\Gamma}^{(l)}\rangle)}{1+\exp(\mu^{(1)}+\langle\boldsymbol{A}_*,\boldsymbol{\Gamma}^{(l)}\rangle)}$$

for $l=1,\ldots,L$. Then $\boldsymbol{A}_*$ is classified as a member of group 'low' or 'high' if $\frac{1}{L}\sum_{l=1}^{L}S^{(l)}$ is less than or greater than a selected cut-off value t_c, respectively. Similarly, node k is recognized to be influential in the classification process if $\frac{1}{L}\sum_{l=1}^{L}\xi_k^{(l)}>t_n$ for a prespecified threshold t_n, where $\xi_k^{(1)},\ldots,\xi_k^{(L)}$ are the L post burn-in MCMC samples of ξ_k. In the same spirit, we employ the algorithm described in Section 4 of the supplementary material to postprocess the posterior samples and identify the influential edges impacting the response. The algorithm takes care of multiplicity correction by controlling the false discovery rate (FDR). Finally, we obtain an estimate of the posterior distribution of the effective dimensionality, $Pr(R_{eff}=r\,|\,Data)\approx\frac{1}{L}\sum_{l=1}^{L}I(\sum_{m=1}^{R}\lambda_m^{(l)}=r)$, where $I(A)$ for an event A is 1 if the event A happens and 0 otherwise, and $\lambda_m^{(1)},\ldots,\lambda_m^{(L)}$ are the L post burn-in MCMC samples of λ_m.

5 Simulation Studies

This section evaluates the inferential and classification ability of two of our proposed network classification priors, the Bayesian Network Lasso classifier (BNLC) and the Bayesian Network Horseshoe classifier (BNHC), vis-a-vis a number of competitors using synthetic networks generated under various simulation settings. In each simulation, we assess the ability of the various approaches to correctly identify influential nodes and edges, to accurately estimate edge coefficients, and to classify a network with precise characterization of uncertainties.

5.1 Simulation Setup

For all of our simulations, data is generated from a logistic regression model of the form

$$y_i \mid \boldsymbol{A}_i, \boldsymbol{\Gamma}_0 \sim Ber\left(\frac{\exp(\mu_0 + \langle \boldsymbol{A}_i, \boldsymbol{\Gamma}_0 \rangle_F)}{1 + \exp(\mu_0 + \langle \boldsymbol{A}_i, \boldsymbol{\Gamma}_0 \rangle_F)}\right), \tag{5.1}$$

where $\boldsymbol{\Gamma}_0$ is a symmetric matrix with zero diagonal entries. We fix the value of the intercept μ_0 at 2 in all simulation scenarios, and then consider different mechanisms for constructing the matrix covariates $\boldsymbol{A}_1, \dots, \boldsymbol{A}_n$ and the matrix of coefficients $\boldsymbol{\Gamma}_0$. In all of our experiments, we work with $V = 25$ nodes and $n = 250$ samples.

We study two different schemes for generating the network $\boldsymbol{A}_i$, referred to as *Simulation 1* and *Simulation 2*, respectively. In *Simulation 1*, the values associated with the network edges are simulated from a standard normal distribution. In contrast, in *Simulation 2*, the nodes in each network are organized into communities so that nodes in the same community tend to have stronger connections than nodes belonging to different communities (i.e., the networks are generated from a blockmodel). This pattern closely mimics real brain connectome networks (Bullmore and Sporns, 2009). More specifically, in *Simulation 2* we assign each node a community label, $f_k \in \{1, 2, 3\}$, $k = 1, \dots, V$. The node assignments are the same for all networks in the population, and the size of the communities are approximately the same. Given the community labels, the (k, k')th element of $\boldsymbol{A}_i$ is simulated from $N(m_{f_k, f_{k'}}, \sigma_0^2)$, where $m_{k,l} = 0.5$ when $k = l$. When $k \neq l$, i.e., when the concerned edges connect nodes belonging to different communities, we sample a fixed number of edge locations randomly and simulate the values from $N(0, 1)$, assigning the values at the remaining locations to be 0. In all cases, we set $\sigma_0^2 = 1$.

On the other hand, the true matrix of coefficients $\boldsymbol{\Gamma}_0$ is constructed as the sum of a low-rank matrix $\boldsymbol{\Gamma}_{0,1}$, which provides the majority of the structure, and a sparse contamination matrix $\boldsymbol{\vartheta}_0$. To construct the matrix $\boldsymbol{\Gamma}_{0,1}$, we draw V latent variables $\boldsymbol{u}_{k,0}$, each of dimension R_0, from a mixture distribution given by

$$\boldsymbol{u}_{k,0} \sim \varrho_0 N\left(0.5\, \mathbf{1}_{R_0}, \boldsymbol{I}_{R_0}\right) + (1 - \varrho_0)\delta_{\mathbf{0}}, \qquad k \in \{1, \dots, V\}, \tag{5.2}$$

where $\mathbf{1}_{R_0}$ is a vector of ones of length R_0, $\boldsymbol{I}_{R_0}$ is an identity matrix of size $R_0 \times R_0$, and ϱ_0 is the probability that any $\boldsymbol{u}_{k,0}$ is nonzero. Then, we define the (k, l)-th element of $\mathbf{\Gamma}_{0,1}$ as $\frac{\boldsymbol{u}_{k,0}'\boldsymbol{u}_{l,0}}{2}$, $k < l$ and as 0 if $k = l$. We refer to $(1 - \varrho_0)$ as the *node sparsity* parameter in the context of the data generation mechanism.

On the other hand, the matrix $\boldsymbol{\vartheta}_0$ is constructed by randomly selecting a proportion $\varrho_{0,2}$ of entries to be non-zero. We refer to $(1 - \varrho_{0,2})$ as the *residual edge sparsity*. We consider three different strategies to generate the non-zero elements of $\boldsymbol{\vartheta}_0$. Under Strategy 1, the non-zero entries of $\boldsymbol{\vartheta}_0$ are simulated from a normal distribution with mean 1 and variance 0.1. Under Strategy 2, they are simulated from normal distribution with mean 0.5 and variance 0.1. Finally, under Strategy 3, the non-zero entries are fixed at 0.5.

For each of *Simulation 1* and *Simulation 2* we consider four different experiments that combine different levels of node sparsity, edge sparsity, true and maximum latent dimensions R_0 and R, as well as different strategies for generating $\boldsymbol{\vartheta}_0$ (see Tables 1 and 2). In particular, note that the various experiments allow model mis-specification with unequal choices of R and R_0. As competitors, we use generic variable selection and shrinkage methods that ignore the network structure in the predictor and treat edges as a long predictor vector to fit high-dimensional binary regression with a logit link function. In particular, we compare the performance of our model with a binary logistic regression with the Lasso penalty (Tibshirani, 1996) on the coefficients. With a slight abuse of terminology, we call it Lasso hereon. We also compare our approach to ordinary high-dimensional binary logistic regression with Bayesian Lasso (BLasso in short) prior (Park and Casella, 2008), and Bayesian Horseshoe (BHS in short) prior (Carvalho et al., 2010) on coefficients, which are popular Bayesian shrinkage regression methods. We used the `glmnet` package in `R` (Friedman et al., 2010) to implement the frequentist Lasso for binary logistic regression, while we have written our own codes for BLasso and BHS. A comparison with these methods will indicate any relative advantage of exploiting the structure of the network predictor. We have also compared our methods to a frequentist approach that develops network classification in the presence of a network predictor and a binary response (Relión et al., 2019). However, we find that all the competitors outperform this approach, and hence we have not included the results for the same.

Cases	R_0	R	Node	Residual Edge	Strategy
			Sparsity $(1 - \varrho_0)$	Sparsity $(1 - \varrho_{0,2})$	
Case - 1	2	2	0.5	0.95	Strategy 1
Case - 2	3	5	0.6	0.95	Strategy 1
Case - 3	2	5	0.5	0.90	Strategy 2
Case - 4	2	5	0.4	0.90	Strategy 3

Table 1: Table presents different cases for ***Simulation 1***. The true dimension R_0 is the dimension of vector object $\boldsymbol{u}_{k,0}$ 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 fitted. Node sparsity and residual edge sparsity are described in the text.

For all Bayesian models we generate $50,000$ MCMC samples, out of which the first $30,000$ are discarded as burn-ins. Convergence is assessed by comparing different simulated sequences of representative parameters starting at different initial values (Gelman

Cases	R_0	R	Node Sparsity $(1-\varrho_0)$	Residual Edge Sparsity $(1-\varrho_{0,2})$	Strategy
Case - 1	2	2	0.5	0.95	Strategy 1
Case - 2	2	4	0.5	0.95	Strategy 1
Case - 3	2	3	0.7	0.95	Strategy 1
Case - 4	2	5	0.4	0.90	Strategy 3

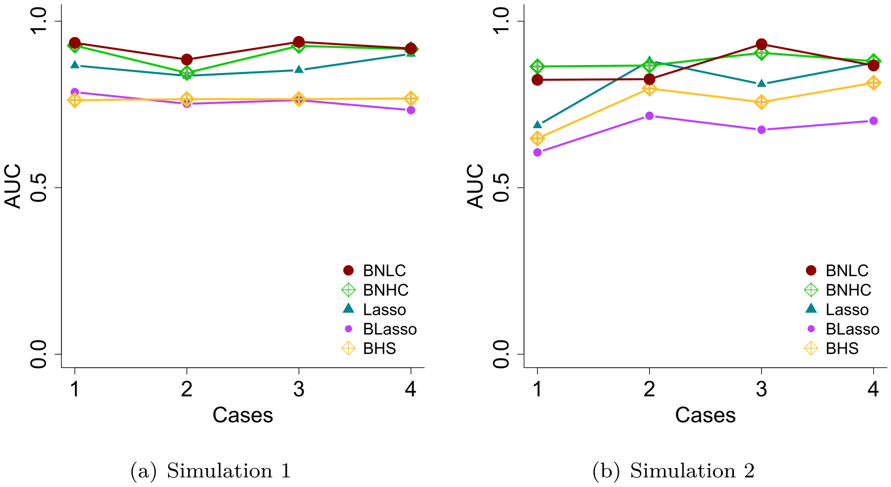
Table 2: Table presents different cases for ***Simulation 2***. The true dimension R_0 is the dimension of vector object $\boldsymbol{u}_{k,0}$ using which data has been generated. The maximum dimension R is the dimension of vector object $\boldsymbol{u}_k$ using which the model has been fitted. Node sparsity and residual edge sparsity are described in the text.

et al., 2014b). We monitor the auto-correlation plots and effective sample sizes of the log likelihood function. In our analysis, we set $\nu = 20$ and $a_\Delta = b_\Delta = 1$. For BNLC, there are two additional hyper-parameters ι and ζ, both of which are set to 1. Note that the choice of $a_\Delta = b_\Delta = 1$ ensures that the prior on models is such that we have a uniform distribution on the number of active nodes, and conditional on the size of the model, a uniform distribution on all possible models of that size. The choice of $\nu = 20$ ensures that the prior distribution on $\boldsymbol{M}$ is concentrated around a scaled identity matrix. Since the model is invariant to rotations of the latent positions, we want the prior on $\boldsymbol{u}_k$'s to also be invariant under rotation. This requires that we center $\boldsymbol{M}$ around a matrix that is proportional to the identity. Our choice of ι and ζ set the prior mean of $s_{k,l}$ at 0.5, which is the suggested prior mean for the local parameters proposed in Park and Casella (2008). Sensitivity to the choice of hyper-parameters is discussed in Section 5.6 for simulation studies and in Section 6.2 for the real data analysis.

5.2 Classification Accuracy and Estimation of Edge Coefficients

To evaluate the out-of-sample predictive performance of the different models, Figure 1 presents the area under the receiver operating characteristic curves (AUCs) obtained by using different classification thresholds t_c (recall the discussion in Section 4). Under *Simulation 1*, BNLC consistently outperforms all other models, with BNHC a very close second. The performance is quite good for models, with AUCs above 0.9 in all four cases. On the other hand, under *Simulation 2* the order appears to switch and BNHC seems to be the best performing model, closely followed by BNLC. AUC values are also high in this case, but uniformly lower than in *Simulation 1*. This suggests that the presence of structure in the network predictor can affect the accuracy of the classification. Among the methods that ignore the network structure associated with the predictor, BLasso tends to have the worse performance, particularly under *Simulation 2*.

In addition to classification rates, we also compute mean squared errors (MSE) associated with the point estimates of the edge coefficients under each model (see Tables 3 and 4). For the Bayesian approaches, point estimates are computed using the posterior means of the edge coefficients. In all cases, BNLC yields point estimates with the lowest MSE. BNHC is the second best-performing method under this metric in almost all cases, closely followed by BLasso.



(a) Simulation 1

(b) Simulation 2

Figure 1: Figure shows classification performance in the form of Area under Curve (AUC) of Receiver Operating Characteristic (ROC) curve for all cases in Simulations 1 and 2.

	MSE				
Cases	BNLC	BNHC	Lasso	BLasso	BHS
Case - 1	**0.164**	0.683	1.197	0.980	1.160
Case - 2	**2.349**	3.568	3.943	3.502	3.993
Case - 3	**0.106**	0.467	0.906	0.695	0.856
Case - 4	**0.166**	0.200	0.485	0.329	0.415

Table 3: Performance of BNLC and BNHC 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.

	MSE				
Cases	BNLC	BNHC	Lasso	BLasso	BHS
Case - 1	**0.279**	0.418	0.807	0.712	0.739
Case - 2	**0.225**	0.396	0.772	0.695	0.728
Case - 3	**0.134**	0.549	0.906	0.748	0.883
Case - 4	**0.066**	0.106	0.167	0.137	0.141

Table 4: Performance of BNLC and BNHC 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.

5.3 Estimation of the Effective Dimensionality

Figures 2 and 3 present posterior distribution of R_{eff}, the effective dimensionality of the latent space, for BNLC and BNHC in *Simulations 1* and *2*, respectively. In all eight experiments, the mode of the posterior distribution corresponds to the true dimension of the latent space. Furthermore, compared to BNLC, note that the posterior distribution of R_{eff} under BNHC tends to concentrate more sharply around the true value.

5.4 Identification of Influential Nodes

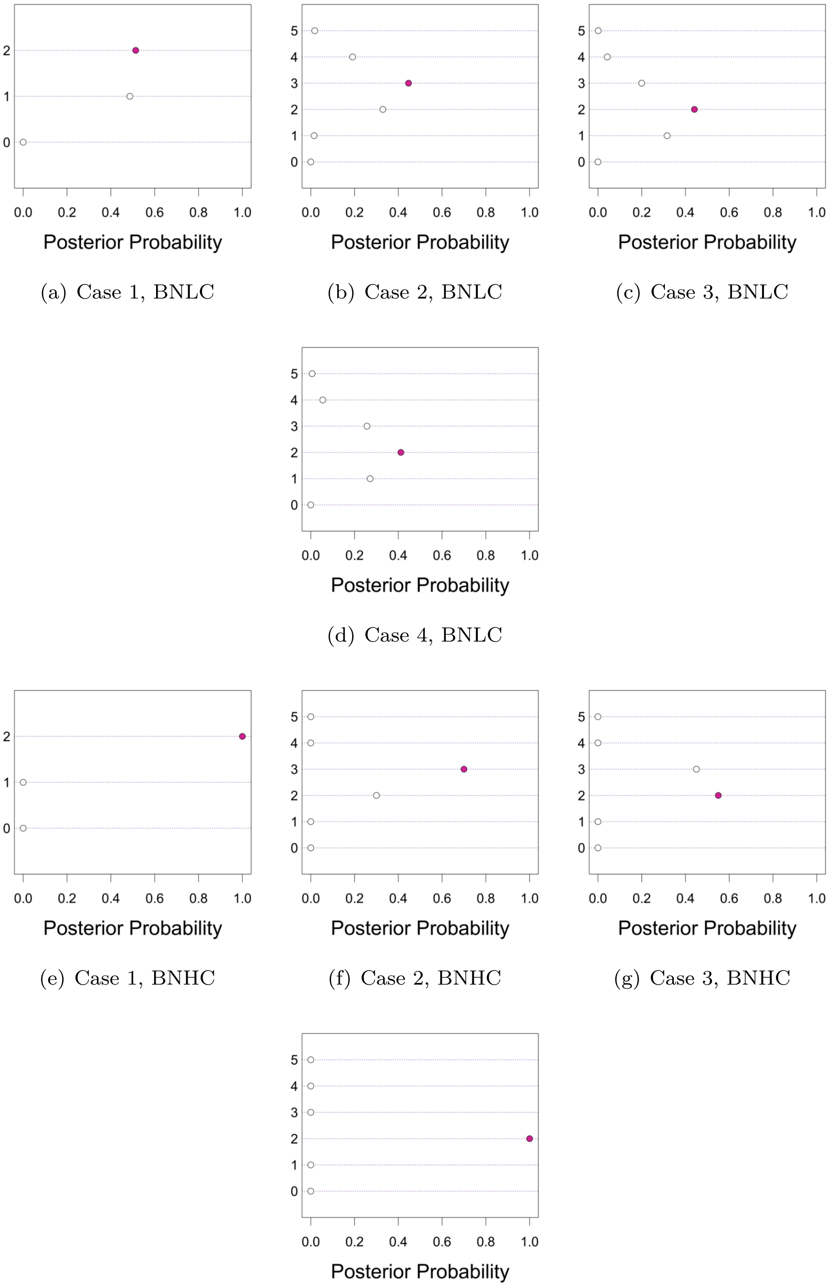
Figures 4 and 5 show the posterior probability of the k-th node being detected as influential, i.e., $Pr(\xi_k = 1 \mid Data)$, under BNLC and BNHC for each node and each case within *Simulations 1* and *2*, respectively. Generally speaking, in *Simulation 1*, BNLC tends to outperform BNHC, with the two methods having comparable true positive rates, but BNHC having a much higher false positive rate (at the standard threshold $t_n = 0.5$). BNHC behaves particularly poorly in case 2, where it identifies 4 false positive nodes and one false negative node (against a perfect separation for BNLC), and in case 3, where it identifies all but one node as influential.

The pattern is similar for *Simulation 2*, with both methods having comparable true positive rates but with BNHC having consistently higher false positive rates and again performing particularly poorly in case 3. However, there are also some notable differences. For example, the true positive rates tend to be lower in *Simulation 2* than in *Simulation 1* for both methods. Similarly, the false positive rate for BNHC seems to be lower in *Simulation 2* than in *Simulation 1*. Finally, in *Simulation 2* both models struggle with case 3 (which was not the case in *Simulation 1*), although they do it in slightly different ways. BNLC shows relatively high rates for both false positives and false negatives (16% in both cases), while BNHC shows a very high false positive rate (32%) but a low false negative rate (4%).

5.5 Identification of Influential Edges

We use the algorithm described in Section 4 of the Supplementary Material (Guha and Rodriguez, 2023) to estimate (local) false discovery rates (FDR) for each of the edge coefficients under the various shrinkage priors. These, in turn, can be used to identify influential edges in the network while controlling for the overall FDR of the procedure. In this section, we attempt to control FDR so that it does not exceed 0.05.

Table 5 shows the realized FDR after applying the procedure to each simulated dataset, as well as true positive rates (TPR) and false positive rates (FPR). The best performing prior is BNLC, where our procedure seems to be well calibrated (realized FDR rates seem to be roughly consistent with the desired nominal rate of 0.05), FPRs tend to be quite low, and TPRs range between 0.46 and 0.72. On the other hand, BNHC tends to yield higher TPRs (ranging between 0.59 and 0.86) in most cases, but at the cost of uniformly higher FPRs, which in turn lead to FDR above the nominal value. The other three variable selection procedures (Lasso, Bayesian Lasso and Horseshoe)



(a) Case 1, BNLC

(b) Case 2, BNLC

(c) Case 3, BNLC

(d) Case 4, BNLC

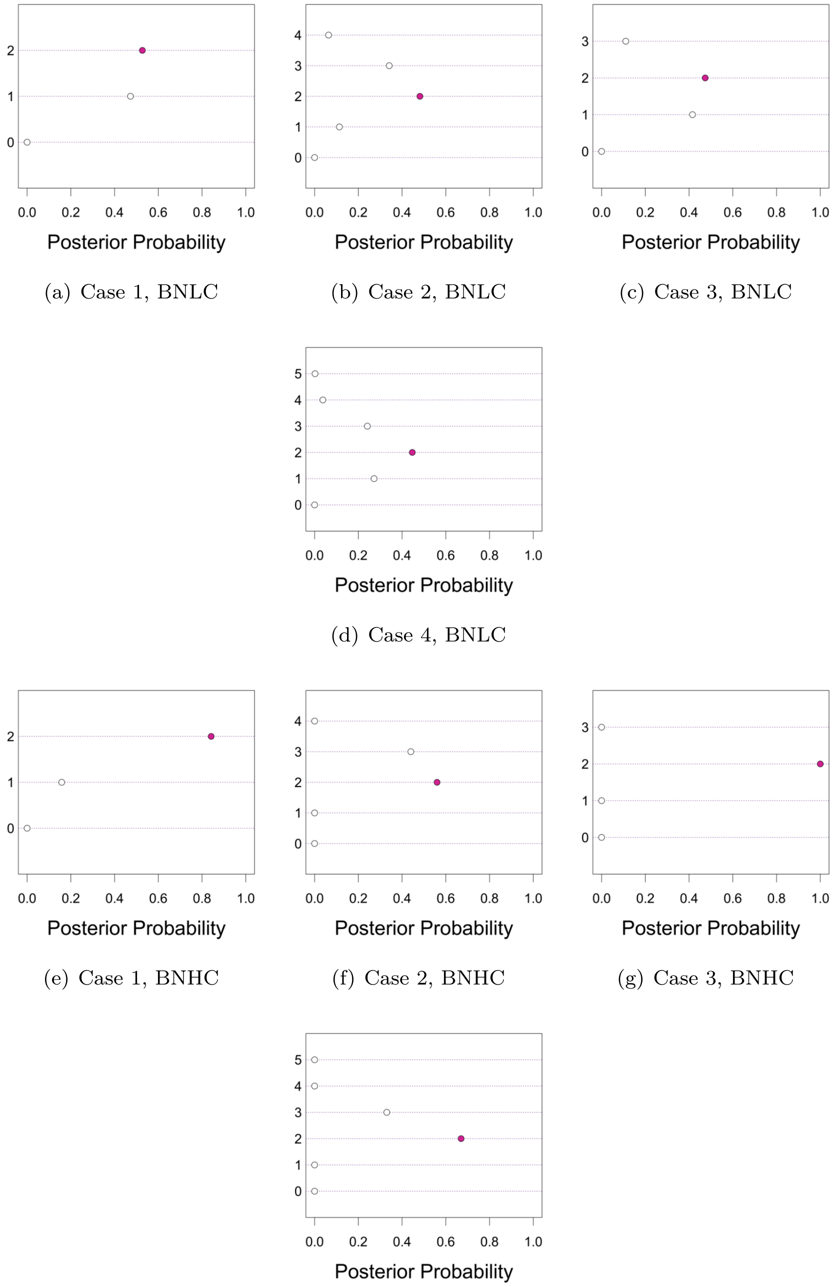
(e) Case 1, BNHC

(f) Case 2, BNHC

(g) Case 3, BNHC

(h) Case 4, BNHC

Figure 2: Plots showing posterior probability distribution of effective dimensionality for BNLC and BNHC models in all 4 cases in Simulation 1. Filled bullets indicate the true value of effective dimensionality.



(a) Case 1, BNLC

(b) Case 2, BNLC

(c) Case 3, BNLC

(d) Case 4, BNLC

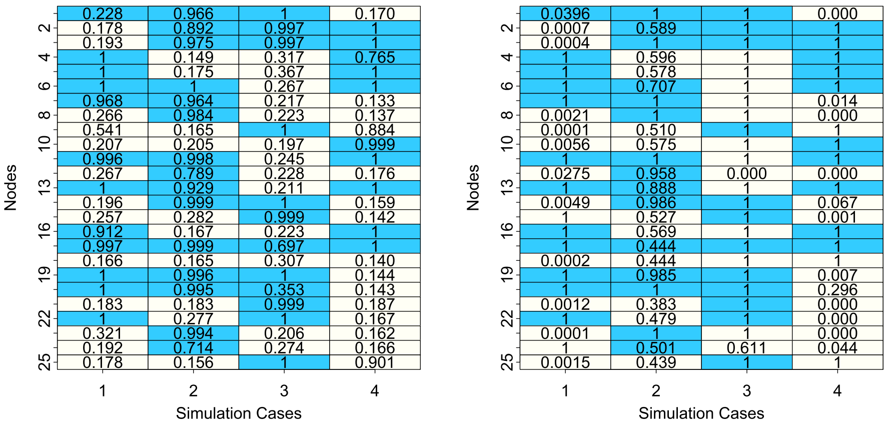
(e) Case 1, BNHC

(f) Case 2, BNHC

(g) Case 3, BNHC

(h) Case 4, BNHC

Figure 3: Plots showing posterior probability distribution of effective dimensionality for BNLC and BNHC models in all 4 cases in Simulation 2. Filled bullets indicate the true value of effective dimensionality.



(a) BNLC

(b) BNHC

Figure 4: Simulation 1: Clear background denotes *uninfluential* and dark background denotes *influential* nodes in the truth for BNLC and BNHC models. Note that there are 25 rows (corresponding to 25 nodes) and 4 columns corresponding to 4 different cases in Simulation 1. The model-detected posterior probability of being influential has been super-imposed onto the corresponding node.

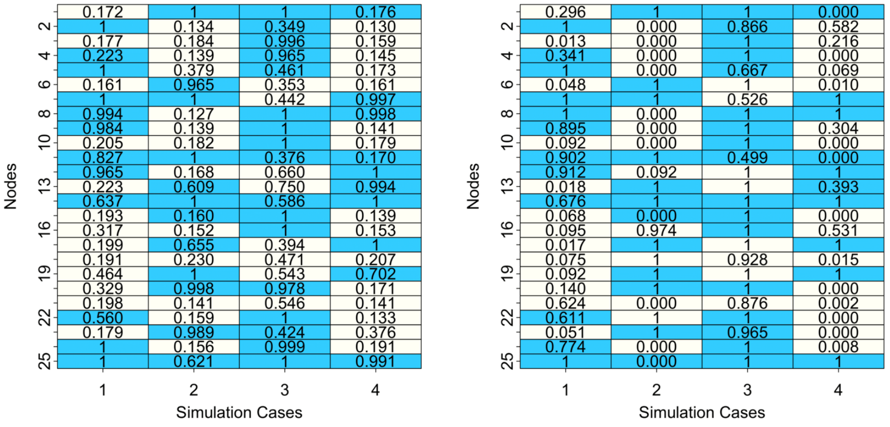
		BNLC			BNHC			Lasso			BLasso			BHS		
Sim	Cases	TPR	FPR	FDR	TPR	FPR	FDR	TPR	FPR	FDR	TPR	FPR	FDR	TPR	FPR	FDR
1	1	0.65	0.01	0.07	0.72	0.12	0.10	0.50	0.22	0.19	0.54	0.19	0.16	0.50	0.18	0.14
	2	0.64	0.00	0.00	0.63	0.02	0.08	0.40	0.14	0.16	0.48	0.18	0.16	0.46	0.11	0.12
	3	0.45	0.00	0.00	0.86	0.40	0.17	0.42	0.22	0.14	0.44	0.20	0.13	0.46	0.12	0.12
	4	0.72	0.09	0.10	0.70	0.12	0.11	0.54	0.16	0.16	0.64	0.20	0.15	0.62	0.15	0.13
2	1	0.63	0.00	0.00	0.84	0.08	0.07	0.44	0.20	0.12	0.44	0.20	0.16	0.52	0.12	0.14
	2	0.59	0.00	0.00	0.64	0.14	0.12	0.45	0.22	0.14	0.44	0.20	0.18	0.45	0.19	0.15
	3	0.46	0.02	0.07	0.59	0.08	0.06	0.31	0.16	0.17	0.36	0.18	0.20	0.32	0.15	0.18
	4	0.68	0.03	0.06	0.75	0.06	0.08	0.34	0.12	0.14	0.31	0.14	0.16	0.31	0.12	0.13

Table 5: True Positive Rates (TPR), False Positive Rates (FPR) and Realized False Discovery Rate (FDR) for edges for cases in *Simulation 1* and *Simulation 2*.

that ignore the network structure in the predictor perform the worst, as they yield lower FPRs and higher FPRs and FDRs.

5.6 Sensitivity to the Choice of Hyperparameters

This section assesses the sensitivity of the inferences from BNLC and BNHC to the choice of hyperparameters. In our sensitivity analysis for BNLC, we consider five scenarios that correspond to alternative values for various subsets of hyperparameters: (i) $a_{\triangle} = 1, b_{\triangle} = 9$; (ii) $\nu = 20, \frac{\zeta}{\iota} = 5$ (iii) $\nu = 50, \frac{\zeta}{\iota} = 5$ (iv) $\nu = 20, \frac{\zeta}{\iota} = 0.2$ (v) $\nu = 50, \frac{\zeta}{\iota} = 0.2$. Combination (i) ensures small prior mean for ξ_k's, while combinations (ii)-(v) allow a range of prior means for θ and $\boldsymbol{M}$. On the other hand, for BNHC we employ three different alternative combinations of hyperparameters: (i) $a_{\triangle} = 1$, $b_{\triangle} = 9$ (ii) $\nu = 10$ (iii) $\nu = 50$. To keep the discussion concise, we only present here results for



Nodes	1	2	3	4
	0.172	1	1	0.176
2	1	0.134	0.349	0.130
	0.177	0.184	0.996	0.159
4	0.223	0.139	0.965	0.145
	1	0.379	0.461	0.173
6	0.161	0.965	0.353	0.161
	1	1	0.442	0.997
8	0.994	0.127	1	0.998
	0.984	0.139	1	0.141
10	0.205	0.182	1	0.179
	0.827	1	0.376	0.170
	0.965	0.168	0.660	1
13	0.223	0.609	0.750	0.994
	0.637	1	0.586	1
	0.193	0.160	1	0.139
16	0.317	0.152	1	0.153
	0.199	0.655	0.394	1
	0.191	0.230	0.471	0.207
19	0.464	1	0.543	0.702
	0.329	0.998	0.978	0.171
	0.198	0.141	0.546	0.141
22	0.560	0.159	1	0.133
	0.179	0.989	0.424	0.376
	1	0.156	0.999	0.191
25	1	0.621	1	0.991

(a) BNLC

Nodes	1	2	3	4
	0.296	1	1	0.000
2	1	0.000	0.866	0.582
	0.013	0.000	1	0.216
4	0.341	0.000	1	0.000
	1	0.000	0.667	0.069
6	0.048	1	1	0.010
	1	1	0.526	1
8	1	0.000	1	1
	0.895	0.000	1	0.304
10	0.092	0.000	1	0.000
	0.902	1	0.499	0.000
	0.912	0.092	1	1
13	0.018	1	1	0.393
	0.676	1	1	1
	0.068	0.000	1	0.000
16	0.095	0.974	1	0.531
	0.017	1	1	1
	0.075	1	0.928	0.015
19	0.092	1	1	1
	0.140	1	1	0.000
	0.624	0.000	0.876	0.002
22	0.611	1	1	0.000
	0.051	1	0.965	0.000
	0.774	0.000	1	0.008
25	1	0.000	1	1

(b) BNHC

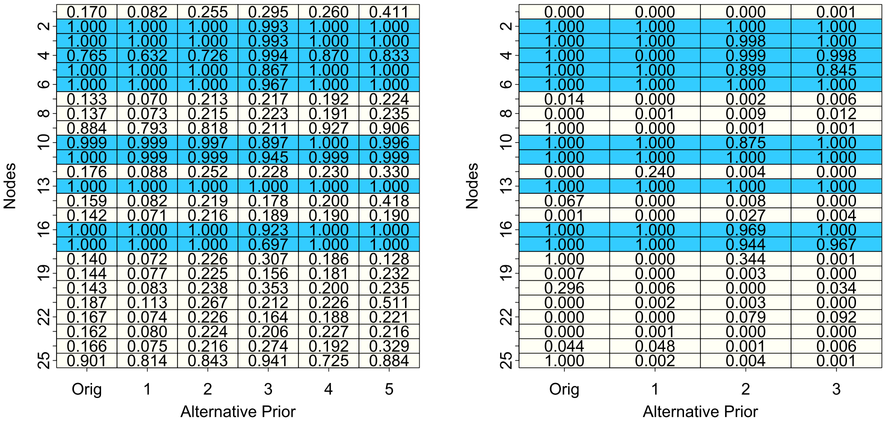
Figure 5: Simulation 2: Clear background denotes *uninfluential* and dark background denotes *influential* nodes in the truth for BNLC and BNHC models. Note that there are 25 rows (corresponding to 25 nodes) and 4 columns corresponding to 4 different cases in Simulation 2. The model-detected posterior probability of being influential has been super-imposed onto the corresponding node.

	BNLC						BNHC			
Prior	Orig	(i)	(ii)	(iii)	(iv)	(v)	Orig	(i)	(ii)	(iii)
MSE	0.16	0.14	0.30	0.22	0.10	0.22	0.20	0.19	0.28	0.28

Table 6: Mean Squared Error (MSE) of estimating the network coefficient in BNLC and BNHC for different combinations of hyper-parameters.

Case 4 in *Simulation 1*, but the outcomes are similar for all other experiments.

Table 6 presents the MSE associated with the network coefficients under the different priors. There seems to be a moderate effect of the priors on the MSE, particularly for BNLC. In particular, prior (ii) under BNLC seems to perform quite poorly when compared with the rest. Figure 6 shows the posterior probabilities of a node being identified as influential under each prior. Again, there seems to be a moderate impact of the prior on the estimated posterior probabilities. Indeed, for most nodes, $Pr(\xi_k = 1 \mid \text{Data})$ is quite comparable across all priors, and the nodes identified as influential by each model (at the standard threshold $t_n = 0.5$) are almost identical across all priors. However, there are some notable exceptions. For example, $Pr(\xi_4 = 1 \mid \text{Data})$ under BNHC is estimated to be 0 under prior (i), but it is estimated to be 1 under the original specification as well as the other two alternatives. Finally, Table 7 offers TPR and FPR values corresponding to the identification of influential edges for BNLC and BNHC under various combinations of hyper-parameters. For BNLC, all alternative prior specifications lead to higher TPRs, at the cost of higher FPRs. In the case BNHC, the



Nodes	Orig	1	2	3	4	5
	0.170	0.082	0.255	0.295	0.260	0.411
2	1.000	1.000	1.000	0.993	1.000	1.000
	1.000	1.000	1.000	0.993	1.000	1.000
4	0.765	0.632	0.726	0.994	0.870	0.833
	1.000	1.000	1.000	0.867	1.000	1.000
6	1.000	1.000	1.000	0.967	1.000	1.000
	0.133	0.070	0.213	0.217	0.192	0.224
8	0.137	0.073	0.215	0.223	0.191	0.235
	0.884	0.793	0.818	0.211	0.927	0.906
10	0.999	0.999	0.997	0.897	1.000	0.996
	1.000	0.999	0.999	0.945	0.999	0.999
	0.176	0.088	0.252	0.228	0.230	0.330
13	1.000	1.000	1.000	1.000	1.000	1.000
	0.159	0.082	0.219	0.178	0.200	0.418
	0.142	0.071	0.216	0.189	0.190	0.190
16	1.000	1.000	1.000	0.923	1.000	1.000
	1.000	1.000	1.000	0.697	1.000	1.000
	0.140	0.072	0.226	0.307	0.186	0.128
19	0.144	0.077	0.225	0.156	0.181	0.232
	0.143	0.083	0.238	0.353	0.200	0.235
	0.187	0.113	0.267	0.212	0.226	0.511
22	0.167	0.074	0.226	0.164	0.188	0.221
	0.162	0.080	0.224	0.206	0.227	0.216
	0.166	0.075	0.216	0.274	0.192	0.329
25	0.901	0.814	0.843	0.941	0.725	0.884

(a) BNLC Sensitivity

Nodes	Orig	1	2	3
	0.000	0.000	0.000	0.001
2	1.000	1.000	1.000	1.000
	1.000	1.000	0.998	1.000
4	1.000	0.000	0.999	0.998
	1.000	1.000	0.899	0.845
6	1.000	1.000	1.000	1.000
	0.014	0.000	0.002	0.006
8	0.000	0.001	0.009	0.012
	1.000	0.000	0.001	0.001
10	1.000	1.000	0.875	1.000
	1.000	1.000	1.000	1.000
	0.000	0.240	0.004	0.000
13	1.000	1.000	1.000	1.000
	0.067	0.000	0.008	0.000
	0.001	0.000	0.027	0.004
16	1.000	1.000	0.969	1.000
	1.000	1.000	0.944	0.967
	1.000	0.000	0.344	0.001
19	0.007	0.000	0.003	0.000
	0.296	0.006	0.000	0.034
	0.000	0.002	0.003	0.000
22	0.000	0.000	0.079	0.092
	0.000	0.001	0.000	0.000
	0.044	0.048	0.001	0.006
25	1.000	0.002	0.004	0.001

(b) BNHC Sensitivity

Figure 6: Figure shows $P(\xi_k = 1|Data)$ for BNLC and BNHC under different hyper-parameter combinations in the simulated data for Case 4 (Simulation 1). The first column in each matrix shows the values corresponding to the original prior specification.

	BNLC						BNHC			
Combinations	Orig.	(i)	(ii)	(iii)	(iv)	(v)	Orig.	(i)	(ii)	(iii)
TPR	0.72	0.80	0.76	0.82	0.83	0.78	0.70	0.64	0.88	0.82
FPR	0.09	0.16	0.21	0.17	0.21	0.18	0.12	0.19	0.24	0.18

Table 7: True Positive Rates (TPR) and False Positive Rates (FPR) of identifying influential edges in BNLC and BNHC for different combinations of hyper-parameters.

results are mixed, with prior (i) performing quite poorly (higher FPR and lower TPR than the original prior), and priors (ii) and (iii) exhibiting the same trade offs (higher TPR at the cost of also higher FDR) as the various alternative priors for BNLC.

Overall, we note that the prior can have a moderate impact on the inference. This should not be surprising. Since this is a high-dimensional regression paradigm with number of parameters far exceeding the sample size, one expects the prior hyper-parameters to have some effect on the inference.

6 Brain Connectome Application

In this section, we apply the BNLC and BNHC priors to study the relationship between a subject's brain connectome and its intelligence based on a sample of $n = 114$ subjects, using the MRN (Mind Research Network) dataset collected at the University of New Mexico, and available at https://neurodata.io. This dataset contains information on the *full scale intelligence quotient* (FSIQ) for multiple individuals. Full scale intelligence quotient (FSIQ) is a measure of an individual's complete cognitive capacity, and is designed to provide a measure of an individual's overall level of general cognitive and intellectual functioning. FSIQ is derived by administering selected sub-tests from the Wechsler Intelligence Scales (WIS) that measure acquired knowledge, verbal reasoning, attention to verbal material, fluid reasoning, spatial processing, attention to detail and visual-motor integration (Caplan et al., 2011). A substantial body of literature has suggested that there is an IQ threshold (usually described as an IQ of approximately 120 points) that may be characterized as superior reasoning ability (Brown et al., 2009; Carson et al., 2003). Following this literature, we have converted the FSIQ scores into a binary response variable y, which takes value 0 if FSIQ is less or equal to 120, and takes value 1 if FSIQ is greater than 120. Thus, we classify the subjects in our study as belonging to the *low IQ* group if $y = 0$, and the *high IQ* group if $y = 1$.

Along with FSIQ measurements, brain connectome information was gathered using weighted diffusion tensor imaging (DTI). DTI is a brain imaging technique that enables measurement of the restricted diffusion of water in tissue in order to produce neural tract images. The brain imaging data we use has been pre-processed using the NeuroData MRI to Graphs (NDMG) pipeline (Kiar et al., 2016, 2017a,b). For the purpose of our analysis, the human brain is divided according to the Desikan atlas (Desikan et al., 2006), which identifies 34 cortical regions of interest (ROIs) both in the left and right hemispheres of the human brain, implying 68 cortical ROIs in all. This results in a brain network of a 68×68 matrix for each individual. Our scientific goals in this setting include identification of brain regions or network nodes significantly related to FSIQ and classification of a subject into the low IQ or high IQ group based on his/her brain connectome information.

The analyses we present in this section use the same hyperparameters as our simulation studies. BNLC and BNHC are both fitted with $R = 4$, which is found to be sufficient for this study (see sensitivity analysis in Section 6.2). The MCMC chains are run for $50,000$ iterations, with the first $30,000$ iterations discarded as burn-in. As before, convergence is assessed by comparing different simulated sequences of representative parameters started at different initial values (Gelman et al., 2014a). The posterior mean for the effective dimensionality of the model is 2.17 for BNLC and 2 for BNHC, and the posterior probabilities associated with 4 dimensions for BNLC and BNHC are given by 0.0082 and 0.0000, respectively.

6.1 Findings from the Brain Connectome Application

We first focus on identifying ROIs that are influential on FSIQ. At a threshold of $t_n = 0.5$, BNLC identifies 38 influential ROIs, 20 in the left and 18 in the right hemisphere. On the other hand, using the same threshold, BNHC identifies 48 influential ROIs, 26 in the left hemisphere and the rest in the right hemisphere. Table 8 lists the 29 ROIs that are identified as influential by both methods. A large number of these are part of the *frontal* lobes in both the hemispheres. Numerous studies have linked the frontal region to an individual's intelligence and cognitive functions (e.g., see Yoon et al., 2017; Stuss et al., 1985; Razumnikova, 2007; Miller and Milner, 1985; Kolb and Milner, 1981). The methods also agree in finding a significant association between FSIQ and ROIs in

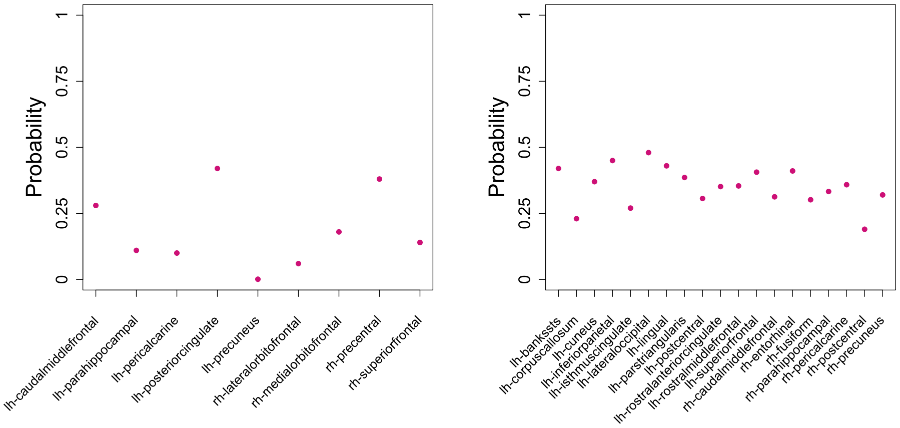
Hemisphere	Lobe	Node
Left	Temporal	fusiform, middle temporal gyrus, parahippocampal, temporal pole, transverse temporal
	Cingulate	isthmus cingulate cortex
	Frontal	pars opercularis, pars orbitalis, pars triangularis, frontal pole
	Occipital	lingual
	Parietal	inferior parietal lobule, precuneus, supramarginal gyrus
	Insula	insula
Right	Temporal	parahippocampal, superior temporal gyrus, temporal pole
	Cingulate	caudal anterior cingulate, isthmus cingulate cortex
	Frontal	lateral orbitofrontal, medial orbitofrontal, pars opercularis, pars orbitalis, rostral middle frontal gyrus, superior frontal gyrus
	Occipital	pericalcarine
	Parietal	supramarginal gyrus
	Insula	insula

Table 8: Nodes identified as influential by both BNLC and BNHC.

the left *inferior parietal lobule*, the left *precuneus* and the *supramarginal gyri* in both the hemispheres, and in the *parietal* lobe. These regions have also been found to be significantly related to FSIQ in Yoon et al. (2017).

Figure 7(a) shows the values of $Pr(\xi_k \mid Data)$ under BNHC for the nine nodes that were identified as influential by BNLC but not by BNHC. Note that most of these probabilities seem to be much lower than 0.5, indicating that BNHC is relatively confident about excluding these nodes. Similarly, Figure 7(b) shows the node inclusion probabilities under BNLC for those nodes that BNHC identified as influential but BNLC did not. In contrast to Figure 7(a), in this case the probabilities tend to be close to 0.5, suggesting that BNLC is not very confident about excluding these nodes. These observations, together with the fact that BNHC tends to include more nodes overall, matches what we saw in our simulation studies and suggests that BNLC is more conservative in identifying influential nodes.

Figure 8 shows the significant edges identified by the BNLC and BNHC models. In this part of the analysis, we only consider edges that connect nodes that were identified as being influential by each prior. BNLC and BNHC identify 142 and 291 edges as being influential, respectively. For the most part, these significant edges are spread across the various nodes. However, we can clearly see that some nodes have no significant edges involving them. Under BNLC (which, as we discussed before, seems to be more conservative) there are only three such nodes (the frontal and temporal poles in the left hemisphere, and the temporal pole in the right hemisphere). However, under BNHC, there are 14 nodes with no significant links (including the middle temporal and parsopercularis regions of the left hemisphere and the precental and precuneus regions of the right hemisphere, among others). While this might be somewhat surprising at first sight, it is not a mistake. Recall that the edges highlighted in the plots were identified by controlling FDR, so they should be interpreted as the set of edges that are *most likely* to be influential, rather than as a comprehensive list. Interestingly, the sets of nodes with at least one significant edge are very similar in the two models, and both



(a) Nodes Selected by BNLC, but not by BNHC

(b) Nodes Selected by BNHC, but not by BNLC

Figure 7: Posterior probabilities of nodes selected as *influential* by one method, but not by another, of being active.

contain the set of 29 common influential nodes that we listed in Table 8.

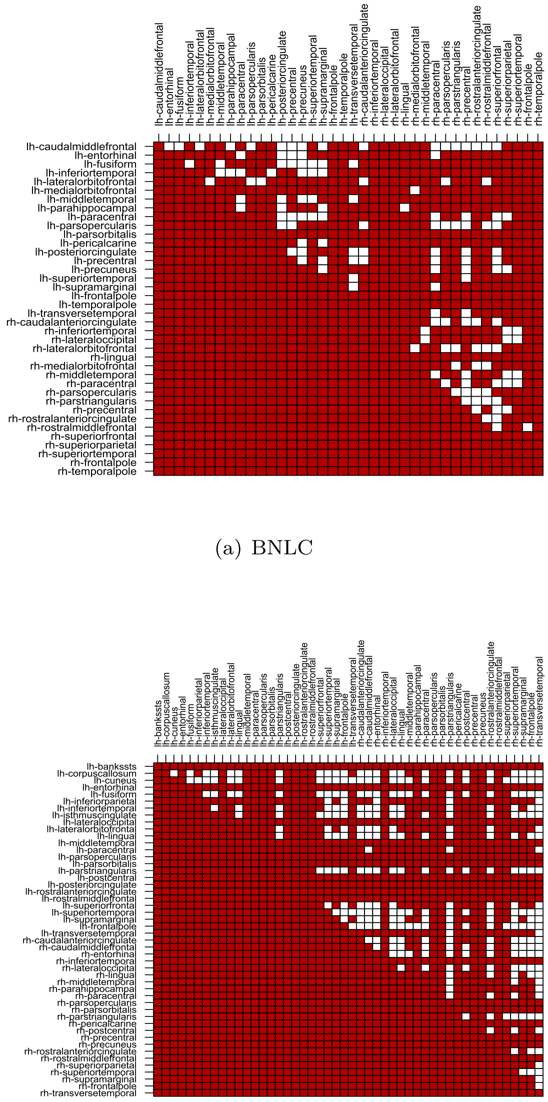
Finally, in order to examine the predictive ability of the Bayesian network classification model, we carry out a 10-fold cross validation exercise and report in Table 9 the (average) AUC for BNLC and BNHC, along with all competing methods. Overall, all AUC values are relatively low. BNLC seems to perform best, followed by BNHC and Lasso. The other two approaches (BLasso and BHS) perform quite poorly in this setting, yielding AUC values below 0.5.

BNLC	BNHC	Lasso	BLasso	BHS
0.617	0.598	0.532	0.461	0.484

Table 9: Average AUC values for a 10-fold cross validation exercise involving the competing approaches.

6.2 Sensitivity to the Choice of Hyperparameters

In this section, we carry out a sensitivity analysis using the same alternative set of hyperparameters introduced in Section 5.6. First, we report in Table 10 the number of nodes identified as influential under each alternative hyperparameter setting, as well as the number of these nodes that intersect with those identified in the original analysis. While there is some variation in the number and exact identity of influential nodes, the different hyperparameter settings largely seem to agree with each other (31 nodes are common to all hyperparameter settings under BNLC, and 40 are common to all under



(a) BNLC

(b) BNHC

Figure 8: Plot showing whether an edge connecting two influential nodes is influential or not. Note that the map is a $M \times M$ symmetric matrix, where M denotes the number of influential nodes, and each cell denotes an edge connecting the corresponding pair of nodes. The axis labels are the abbreviated names of the influential ROIs in the left (starting with 'lh -') and the right (starting with 'rh -') hemispheres of the brain. Full names of the ROIs can be obtained from the widely available Desikan brain atlas. A white cell represents an influential edge, while a red cell represents a non-influential edge.

BNHC). Next, we identify significant edges that connect nodes identified as influential by all sets of hyperparameters under each model (see Table 11). As before, the overlap is substantial. Finally, we investigate the sensitivity of the models to the choice of R. To accomplish this, we rerun each of our two models with $R = 8$ and $R = 10$, and report the posterior means of the effective dimensionality, along with AUC (see Table 12). While we see a small increase in R_{eff} as R increases (particularly for BNHC), the change has almost no effect on the AUC.

	BNLC					BNHC		
Combinations	(i)	(ii)	(iii)	(iv)	(v)	(i)’	(ii)’	(iii)’
# Nodes detected	35	39	34	40	37	45	49	44
# Intersections with original analysis	34	36	34	37	37	42	45	43

Table 10: Number of nodes identified as influential for all combinations are presented. The table also presents the number of intersections of influential nodes between different combinations and the original analysis.

	BNLC					BNHC		
Combinations	(i)	(ii)	(iii)	(iv)	(v)	(i)’	(ii)’	(iii)’
# Edges detected	122	113	125	118	107	272	265	262
# Intersections with original analysis	117	112	119	111	101	263	264	257

Table 11: Number of edges identified as influential for all combinations are presented. The table also presents the number of intersections of influential nodes between different combinations and the original analysis.

	BNLC			BNHC		
	$R = 4$	$R = 8$	$R = 10$	$R = 4$	$R = 8$	$R = 10$
Posterior mean Eff. Dim.	2.17	2.78	2.96	2.00	2.74	3.04
AUC	0.61	0.63	0.59	0.59	0.60	0.59

Table 12: AUC and posterior mean of effective dimensionality for BNLC and BNHC under different choices of R.

7 Conclusion

We have developed a binary Bayesian network regression model that enables the classification of multiple networks with labeled nodes, identifies influential network nodes and edges, and predicts the class in which a newly observed network belongs. Our contribution lies in carefully constructing a class of network global-local shrinkage priors on the network predictor coefficient while recognizing the latent network structure in the predictor variable. Our simulation studies show competitive performance in terms of inference and classification. On the other hand, the results we obtain in our application to brain connectome data both corroborate results that had already been described in the literature and suggest new relationships between brain ROIs and FSIQ scores.

There are several natural directions for future research. A major contribution of the proposed framework is a theoretical analysis of the asymptotic properties of the model under a condition in which the size of the predictor matrix increases with the sample size at a superlinear rate. Developing a similar theory for the Network Horseshoe prior proposed in this article faces more challenges due to the more complex prior structure in the parameters. We plan to tackle this problem as part of future work. It is also important to note that the theoretical results in this article hinge upon the assumption that the rank of the true network coefficient is known. As part of future work, we will investigate the model theoretically when the fitted network coefficient is low-rank but the true network coefficient is full rank. We also emphasize that Theorem 3.1 guarantees asymptotically consistent classification, but provides no consistency guarantee for the regression coefficients or their effective dimension. While Wei and Ghosal (2020) have established such a result for ordinary high-dimensional logistic regression, a similar result would be substantially more challenging to establish in our framework. We will consider this as a part of our future theoretical explorations. Finally, this article illustrates our approach on a dataset where a continuous outcome FSIQ is discretized to construct a discrete response variable. We plan to do future work emphasizing efficacy of our approach on a real data application with a network predictor and a binary outcome.

Supplementary Material

Supplementary Material: High-Dimensional Bayesian Network Classification with Network Global-Local Shrinkage Priors (DOI: 10.1214/23-BA1378SUPP). The supplementary material has four sections. Section 1 discusses the proof of Theorem 3.1. Section 2 discusses the MCMC algorithm for the Network Lasso Shrinkage prior. Section 3 discusses the MCMC algorithm for the Network Horsheshoe Shrinkage prior. Section 4 discusses the edge selection procedure.

References

Armagan, A., Dunson, D. B., and Lee, J. (2013a). "Generalized double Pareto shrinkage." *Statistica Sinica*, 23(1): 119–143. MR3076161. 1132, 1135, 1136, 1137

Armagan, A., Dunson, D. B., Lee, J., Bajwa, W. U., and Strawn, N. (2013b). "Posterior consistency in linear models under shrinkage priors." *Biometrika*, 100(4): 1011–1018. MR3142348. doi: `https://doi.org/10.1093/biomet/ast028`. 1133

Bai, R. and Ghosh, M. (2018). "High-dimensional multivariate posterior consistency under global–local shrinkage priors." *Journal of Multivariate Analysis*, 167: 157–170. MR3830639. doi: `https://doi.org/10.1016/j.jmva.2018.04.010`. 1133

Belitser, E. and Nurushev, N. (2015). "Needles and straw in a haystack: robust confidence for possibly sparse sequences." *arXiv preprint arXiv:1511.01803*. MR4036032. doi: `https://doi.org/10.3150/19-BEJ1122`. 1133

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2019). "Lasso meets horseshoe: A survey." *Statistical Science*, 34(3): 405–427. MR4017521. doi: https://doi.org/10.1214/19-STS700. 1133

Brown, T. E., Reichel, P. C., and Quinlan, D. M. (2009). "Executive function impairments in high IQ adults with ADHD." *Journal of Attention Disorders*, 13(2): 161–167. 1150

Bullmore, E. and Sporns, O. (2009). "Complex brain networks: graph theoretical analysis of structural and functional systems." *Nature Reviews. Neuroscience*, 10(3): 186–198. 1132, 1140

Caplan, B., Kreutzer, J. S., and DeLuca, J. (2011). *Encyclopedia of Clinical Neuropsychology; With 199 Figures and 139 Tables.*. Springer. 1150

Carson, S. H., Peterson, J. B., and Higgins, D. M. (2003). "Decreased latent inhibition is associated with increased creative achievement in high-functioning individuals." *Journal of personality and social psychology*, 85(3): 499. 1150

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The Horseshoe Estimator for sparse signals." *Biometrika*, 97(2): 465–480. MR2650751. doi: https://doi.org/10.1093/biomet/asq017. 1132, 1135, 1141

Castillo, I., Rousseau, J., et al. (2015). "A Bernstein–von Mises theorem for smooth functionals in semiparametric models." *The Annals of Statistics*, 43(6): 2353–2383. MR3405597. doi: https://doi.org/10.1214/15-AOS1336. 1133

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. MR3059077. doi: https://doi.org/10.1214/12-AOS1029. 1133

Craddock, R. C., Holtzheimer III, P. E., Hu, X. P., and Mayberg, H. S. (2009). "Disease state prediction from resting state functional connectivity." *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 62(6): 1619–1628. 1132

Daianu, M., Jahanshad, N., Nir, T. M., Toga, A. W., Jack Jr, C. R., Weiner, M. W., and Thompson, P. M., for the Alzheimer's Disease Neuroimaging Initiative (2013). "Breakdown of brain connectivity between normal aging and Alzheimer's disease: a structural k-core network analysis." *Brain connectivity*, 3(4): 407–422. 1132

Deshpande, M., Kuramochi, M., Wale, N., and Karypis, G. (2005). "Frequent substructure-based approaches for classifying chemical compounds." *IEEE Transactions on Knowledge and Data Engineering*, 17(8): 1036–1050. 1132

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. 1131, 1150

Du, X. and Ghosal, S. (2018). "Bayesian discriminant analysis using a high dimensional predictor." *Sankhya A*, 80(1): 112–145. MR3968360. doi: https://doi.org/10.1007/s13171-018-0140-z. 1132

Durante, D., Dunson, D. B., et al. (2018). "Bayesian inference and testing of group differences in brain networks." *Bayesian Analysis*, 13(1): 29–58. MR3737942. doi: https://doi.org/10.1214/16-BA1030. 1132

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. MR0125031. 1131

Fei, H. and Huan, J. (2010). "Boosting with structure information in the functional space: an application to graph classification." In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 643–652. ACM. 1132

Frank, O. and Strauss, D. (1986). "Markov graphs." *Journal of the American Statistical Association*, 81(395): 832–842. MR0860518. 1131

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. 1141

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014a). *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL. MR3235677. 1150

Gelman, A., Hwang, J., and Vehtari, A. (2014b). "Understanding predictive information criteria for Bayesian models." *Statistics and computing*, 24(6): 997–1016. MR3253850. doi: https://doi.org/10.1007/s11222-013-9416-2. 1141

Ghosal, S., Roy, A., et al. (2006). "Posterior consistency of Gaussian process prior for nonparametric binary regression." *The Annals of Statistics*, 34(5): 2413–2429. MR2291505. doi: https://doi.org/10.1214/009053606000000795. 1137

Guha, S. and Rodriguez, A. (2021). "Bayesian regression with undirected network predictors with an application to brain connectome data." *Journal of the American Statistical Association*, 116(534): 581–593. MR4270005. doi: https://doi.org/10.1080/01621459.2020.1772079. 1133, 1135

Guha, S. and Rodriguez, A. (2023). "Supplementary Material: High-Dimensional Bayesian Network Classification with Network Global-Local Shrinkage Priors." *Bayesian Analysis*. doi: https://doi.org/10.1214/23-BA1378SUPP. 1134, 1138, 1144

Guhaniyogi, R. and Rodriguez, A. (2020). "Joint modeling of longitudinal relational data and exogenous variables." *Bayesian Analysis*, 15(2): 477–503. MR4078722. doi: https://doi.org/10.1214/19-BA1160. 1131

Helma, C., King, R. D., Kramer, S., and Srinivasan, A. (2001). "The predictive toxicology challenge 2000–2001." *Bioinformatics*, 17(1): 107–108. 1132

Hoff, P. D. (2005). "Bilinear mixed-effects models for dyadic data." *Journal of the American Statistical Association*, 100(469): 286–295. MR2156838. doi: https://doi.org/10.1198/016214504000001015. 1131, 1135

Hoff, P. D. (2009). "Multiplicative latent factor models for description and prediction of social networks." *Computational and mathematical organization theory*, 15(4): 261. 1131

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. MR1951262. doi: https://doi.org/10.1198/016214502388618906. 1131

Jeong, S. and Ghosal, S. (2021). "Posterior contraction in sparse generalized linear models." *Biometrika*, 108(2): 367–379. MR4259137. doi: https://doi.org/10.1093/biomet/asaa074. 1133

Kiar, G., Gorgolewski, K., and Kleissas, D. (2017a). "Example Use Case of sic with the ndmg Pipeline (sic: ndmg)." *GigaScience Database*. 1150

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. 1150

Kiar, G., Gray Roncal, W., Mhembere, D., Bridgeford, E., Burns, R., and Vogelstein, J. (2016). "ndmg: NeuroData's MRI graphs pipeline." 1150

Kolb, B. and Milner, B. (1981). "Performance of complex arm and facial movements after focal brain lesions." *Neuropsychologia*, 19(4): 491–503. 1150

Martin, R., Mess, R., Walker, S. G., et al. (2017). "Empirical Bayes posterior concentration in sparse high-dimensional linear models." *Bernoulli*, 23(3): 1822–1847. MR3624879. doi: https://doi.org/10.3150/15-BEJ797. 1133

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. 1150

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. MR1947255. doi: https://doi.org/10.1198/016214501753208735. 1131

Olde Dubbelink, K. T., Hillebrand, A., Stoffers, D., Deijen, J. B., Twisk, J. W., Stam, C. J., and Berendse, H. W. (2013). "Disrupted brain network topology in Parkinson's disease: a longitudinal magnetoencephalography study." *Brain*, 137(1): 197–207. 1132

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, 103(482): 681–686. MR2524001. doi: https://doi.org/10.1198/016214508000000337. 1132, 1135, 1141, 1142

Polson, N. G., Scott, J. G., and Windle, J. (2013). "Bayesian inference for logistic models using Pólya–Gamma latent variables." *Journal of the American statistical Association*, 108(504): 1339–1349. MR3174712. doi: https://doi.org/10.1080/01621459.2013.829001. 1139

Razumnikova, O. M. (2007). "Creativity related cortex activity in the remote associates task." *Brain Research Bulletin*, 73(1): 96–102. 1150

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. MR4019153. doi: https://doi.org/10.1214/19-AOAS1252. 1133, 1141

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. 1132

Rodriguez, A. (2012). "Modeling the dynamics of social networks using Bayesian hierarchical blockmodels." *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(3): 218–234. MR2929964. doi: https://doi.org/10.1002/sam.10150. 1131

Scott, J. G. and Berger, J. O. (2010). "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." *The Annals of Statistics*, 2587–2619. MR2722450. doi: https://doi.org/10.1214/10-AOS792. 1136

Song, Q. and Liang, F. (2017). "Nearly optimal Bayesian shrinkage for high dimensional regression." *arXiv preprint arXiv:1712.08964*. MR4535982. doi: https://doi.org/10.1007/s11425-020-1912-6. 1133

Sosa, J. and Rodríguez, A. (2021). "A latent space model for cognitive social structures data." *Social Networks*, 65: 85–97. 1131

Srinivasan, A., Muggleton, S. H., Sternberg, M. J., and King, R. D. (1996). "Theories for mutagenicity: A study in first-order and feature-based induction." *Artificial Intelligence*, 85(1-2): 277–299. 1132

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. 1150

Teh, Y. W., Grür, D., and Ghahramani, Z. (2007). "Stick-breaking construction for the Indian buffet process." In *Artificial Intelligence and Statistics*, 556–563. 1136

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1): 267–288. MR1379242. 1132, 1141

Van Der Pas, S. L., Kleijn, B. J., Van Der Vaart, A. W., et al. (2014). "The horseshoe estimator: Posterior concentration around nearly black vectors." *Electronic Journal of Statistics*, 8(2): 2585–2618. MR3285877. doi: https://doi.org/10.1214/14-EJS962. 1133

Vishwanathan, S. V. N., Schraudolph, N. N., Kondor, R., and Borgwardt, K. M. (2010). “Graph kernels.” *Journal of Machine Learning Research*, 11(Apr): 1201–1242. MR2645450. doi: https://doi.org/10.1093/chemse/bjq147. 1132

Vogelstein, J. T., Roncal, W. G., Vogelstein, R. J., and Priebe, C. E. (2013). “Graph classification using signal-subgraphs: Applications in statistical connectomics.” *IEEE transactions on pattern analysis and machine intelligence*, 35(7): 1539–1551. MR3338694. doi: https://doi.org/10.1007/s00357-015-9170-6. 1132

Wei, R. and Ghosal, S. (2020). “Contraction properties of shrinkage priors in logistic regression.” *Journal of Statistical Planning and Inference*, 207: 215–229. MR4066132. doi: https://doi.org/10.1016/j.jspi.2019.12.004. 1133, 1134, 1137, 1138, 1155

Yoon, Y. B., Shin, W.-G., Lee, T. Y., Hur, J.-W., Cho, K. I. K., Sohn, W. S., Kim, S.-G., Lee, K.-H., and Kwon, J. S. (2017). “Brain structural networks associated with intelligence and visuomotor ability.” *Scientific reports*, 7(1): 2177. 1150, 1151

Zhang, J., Cheng, W., Wang, Z., Zhang, Z., Lu, W., Lu, G., and Feng, J. (2012). “Pattern classification of large-scale functional brain networks: identification of informative neuroimaging markers for epilepsy.” *PloS one*, 7(5): e36733. 1132

Zhang, R. and Ghosh, M. (2019). “Ultra High-dimensional Multivariate Posterior Contraction Rate Under Shrinkage Priors.” *arXiv preprint arXiv:1904.04417*. MR4322321. doi: https://doi.org/10.1016/j.jmva.2021.104835. 1133