High Dimensional Bayesian Network Classification with Low Rank and Sparse Shrinkage Priors

Sharmistha Guha and Abel Rodriguez

December 15, 2018

Abstract

This article proposes a novel Bayesian classification framework for networks with labeled nodes. While literature on statistical modeling of network data typically involves analysis of a single network, the recent emergence of complex datasets in several biological applications, including brain imaging studies, presents a pressing need to devise a network classifier for every individual. This article considers one such application from a brain connectome study, where the overarching goal is to classify subjects into two separate groups based on their brain network data, along with identifying influential regions of interest (ROIs) (referred to as *nodes*). Existing approaches either treat all edge weights as a long vector or summarize the network information with a few summary measures. Both these approaches ignore the full network structure, may lead to less desirable inference in small samples and are not designed to identify significant network nodes. We propose a novel binary logistic regression framework with the network as the predictor and a binary response, the network predictor coefficient being modeled using an additive structure having a low-rank and a sparse component. The framework is able to accurately detect both nodes and edges in the network influencing the classification. Our framework is implemented using an efficient Markov Chain Monte Carlo algorithm. Theoretically, we show asymptotically optimal classification for the proposed framework when the number of network edges grows faster than the sample size. The framework is empirically validated by extensive simulation studies and analysis of a brain connectome data.

Keywords: Brain connectome; Edge selection; High dimensional binary regression; Low-rank prior; Node selection; Network predictor; Posterior consistency.

1 Introduction

Of late, the statistical literature has paid heavy attention to the unsupervised analysis of a single network, thought to be generated from a variety of classic models, including random graph models (Erdos and Rényi, 1960), exponential random graph models (Frank and Strauss, 1986), social space models (Hoff *et al.*, 2002; Hoff, 2005; Hoff, 2009) and stochastic

block models (Nowicki and Snijders, 2001). These models have found prominence in social networking applications where the nodes of the network are exchangeable. However, there are pertinent biological and physiological applications in which network nodes are labeled and a network is available corresponding to every individual. Section 6 presents one such example from a brain connectome study, where brain networks are available for multiple individuals who are classified as subjects with high or low intelligence quotients (IQ). In this study, 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 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 are *labeled* corresponding to different ROIs (shared among networks corresponding to all individuals) and entries correspond to estimates of the number of *fibers* connecting pairs of brain regions. The scientific goal in this setting pertains to developing a predictive rule for classifying a newly observed brain network with labeled nodes. Additionally, it is of specific interest for neuroscientists to identify influential brain regions (nodes in the brain network).

Earlier literature on network or graph classification has been substantially motivated by the problem of classification of chemical compounds (Srinivasan *et al.*, 1996; Helma *et al.*, 2001), where a graph represents a compound's molecular structure. In such analyses, certain discriminative patterns in a graph are identified and used as features for training a standard classification method (Deshpande *et al.*, 2005; Fei and Huan, 2010). Another type of method is based on graph kernels (Vishwanathan *et al.*, 2010), which defines a similarity measure between two networks. Both these approaches are computationally feasible for small networks, do not account for uncertainty and do not facilitate influential network node identification. When the number of network nodes is moderately large, a common approach to network classification is to use a few summary measures (average degree, clustering coefficient, or average path length) from the network in the context of a flexible classification approach (see, for e.g., Bullmore and Sporns, 2009 and references therein). Clearly, the success of this approach is highly dependent upon selecting the right summaries to include. Furthermore, global summary statistics collapse all local network information, which can affect the accuracy of classification, not allowing identification of local differences. Furthermore, identification of the impact of specific nodes on the response, which is of clear interest in our setting, is not feasible. An alternative approach proceeds to vectorize the network predictor and treat edge weights together as a long vector followed by developing a high dimensional regression model with this long vector of edge weights as predictors (Richiardi *et al.*, 2011; Craddock *et al.*, 2009; Zhang *et al.*, 2012). This approach can take advantage of the recent developments in high dimensional binary 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 treats the links of the network as exchangeable, ignoring the fact that coefficients involving common nodes can be expected to be correlated a priori. Ignoring this correlation may lead to unsatisfactory predictive performance and can potentially impact model selection. In a related work, Vogelstein *et al.*, 2013 propose to look for a minimal set of nodes which best explains the difference between two

groups of networks, which requires solving a combinatorial problem. Durante and Dunson, 2017 propose a high dimensional Bayesian tensor factorization model for a population of networks that allows to test for local edge differences between groups. Their approach tends to focus mainly on classification and is not designed to detect important nodes impacting the response.

Our goal in this paper is to develop a high-dimensional Bayesian network classifier that not only uses all the individual edge weights, but also respects the network structure of the data and infers on influential nodes impacting classification. To achieve this goal, we formulate a high dimensional logistic network model with the network as the predictor on the binary response corresponding to each individual. The network predictor coefficient is proposed to assume an additive low-rank and sparse structure. While the low-rank component of the predictor coefficient is designed to mainly capture the impact of interactions between different nodes on the regression function, the sparse component accounts for the additional effects due to the edges connecting the nodes. Low-rank modeling of predictor coefficient matrices has been observed in the tensor regression modeling literature (Zhou *et al.*, 2013; Guhaniyogi *et al.*, 2017), though modeling predictor coefficient matrices using low-rank plus sparse decomposition is less common. A few articles have emphasized the advantage of modeling coefficient matrices using a low-rank plus sparse structure in the context of unsupervised analysis of matrix valued data in the frequentist literature. For example, low-rank plus sparse decomposition of Hankel matrices representing the input-output structure of a linear time invariant system (LTI) has appeared in the literature (Fazel *et al.*, 2003; Chandrasekaran *et al.*, 2011). Our proposal involves node specific latent variables in the low-rank component which are assigned a discrete mixture prior distribution to facilitate important node identification. On the other hand, Bayesian shrinkage priors are assigned on the elements of the second component to ensure a "near sparse" structure a posteriori. We coin the prior on both components together as the *network shrinkage prior*. The proposed framework respects the network structure of the predictor, leads to accurate classification and allows us to identify nodes that have influence on the response. Theoretical results guaranteeing asymptotically optimal prediction by the proposed approach have also been demonstrated.

Recently, Relión *et al.*, 2017 have proposed a penalized optimization scheme that enables classification of networks, in addition to identifying important nodes. Although this model seems to perform well for the classification 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. Recently Guha and Rodriguez, 2018 have proposed a Bayesian network regression framework for a network predictor and a continuous response. In contrast, our framework is designed for the network classification problem. Additionally, we offer novel theoretical developments ensuring optimal classification which is not provided in Guha and Rodriguez, 2018.

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 paper with a brief discussion of the proposed methodology and a discussion of possible future work.

2 Model Formulation

2.1 Notations

Let $\boldsymbol{M}_i$ represent the weighted undirected network predictor with an associated binary response $y_i \in \{0, 1\}$, for $i = 1, ..., n$. The number of samples in the study is denoted by n. The weights corresponding to the edges belong to $\mathbb{R}$ and all graphs share the same labels on their nodes. For example, in the brain connectome application discussed subsequently, $\boldsymbol{M}_i$ encodes the network connections between different regions of the brain for the ith individual and y_i is an indicator signifying if the I.Q. level of ith individual has been been found to be 'high' or 'low'. Let the network corresponding to any individual consist of V nodes. Thus $\boldsymbol{M}_i$ is a $V \times V$ symmetric matrix, with the (k, l)th entry of $\boldsymbol{M}_i$ denoted by $m_{i[kl]} \in \mathbb{R}$ and $m_{i[kl]} = m_{i[lk]}$, $k > l$. Our network specification allows no self relationships between nodes, i.e. $m_{i,[kk]} \equiv 0$ for all $k = 1, ..., V$. 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 $\boldsymbol{A}$ and $\boldsymbol{B}$ by $\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F = Trace(\boldsymbol{B}'\boldsymbol{A})$. 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. Frobenius norm of a matrix $\boldsymbol{A}$ is defined as $||\boldsymbol{A}||_F = \sqrt{\langle \boldsymbol{A}, \boldsymbol{A} \rangle_F}$. Additionally, for any vector $\boldsymbol{a} = (a_1, ..., a_p)'$, define the L_1, L_2 and L_∞ norms by $||\boldsymbol{a}||_1 = \sum_{l=1}^{p} |a_l|$, $||\boldsymbol{a}||_2 = \sqrt{\sum_{l=1}^{p} a_l^2}$ and $||\boldsymbol{a}||_\infty = \max_l |a_l|$ respectively. $|| \cdot ||_0$ denotes the L_0-norm, i.e. the number of non-zero entries for vectors. The $|| \cdot ||_1$, $|| \cdot ||_2$ and $|| \cdot ||_\infty$ norms of a matrix are defined analogously.

2.2 Bayesian Network Classification Model

In the context of network classification, we propose the high dimensional logistic regression model of the binary response y_i on the undirected network predictor $\boldsymbol{M}_i$ as

$$y_i \sim Bernoulli\left[\frac{\exp(\psi_i)}{1 + \exp(\psi_i)}\right], \quad \psi_i = \mu + \langle \boldsymbol{M}_i, \boldsymbol{\Gamma} \rangle_F, \tag{1}$$

where $\boldsymbol{\Gamma}$ is a $V \times V$ symmetric network coefficient matrix whose (k, l)th element is given by $2\gamma_{kl}$, with $\gamma_{kk} = 0$, for all $k = 1, ..., V$.

Model (1) can be expressed in the form of a generalized linear model. To be more

specific, $\langle \boldsymbol{M}_i, \boldsymbol{\Gamma} \rangle_F = \sum_{1 \le k < l \le V} m_{i[kl]} \gamma_{kl}$, so that $\psi_i = \mu + \sum_{1 \le k < l \le V} m_{i[kl]} \gamma_{kl}$ and the probability mass function of y_i can be written as

$$p(y_i) = \frac{\exp(\psi_i)^{y_i}}{1 + \exp(\psi_i)} \tag{2}$$

Equation (2) connects the binary network regression model with the high dimensional binary regression framework, with $m_{i[kl]}$'s as predictors and γ_{kl}'s as the corresponding coefficients. To be more precise, if $\boldsymbol{x}_i = (m_{i[12]}, ..., m_{i[(V-1)V]})' \in \mathbb{R}^{V(V-1)/2}$ is the collection of all upper triangular elements of $\boldsymbol{M}_i$, and $\boldsymbol{\gamma} = (2\gamma_{12}, ..., 2\gamma_{(V-1)V})' \in \mathbb{R}^{V(V-1)/2}$ is the vector of corresponding upper triangular elements of $\boldsymbol{\Gamma}$, then (1) can be written as

$$y_i \sim Bernoulli\left[f_{\boldsymbol{\gamma}}(\boldsymbol{x}_i)\right], \;\; f_{\boldsymbol{\gamma}}(\boldsymbol{x}_i) = \frac{\exp(\mu + \boldsymbol{x}_i'\boldsymbol{\gamma})}{(1 + \exp(\mu + \boldsymbol{x}_i'\boldsymbol{\gamma}))}. \tag{3}$$

A few remarks are in order. Although the binary network regression model is proposed for the logit link, it assumes natural extension for any other link function. While ordinary linear regression indexes predictor coefficients by the natural numbers $\mathcal{N}$, Model (2) indexes the predictor coefficients by their positions in the matrix $\boldsymbol{\Gamma}$ to encode the information of the edges as well as the nodes connecting the edges. As mentioned earlier, we are interested in identifying nodes and edges which contribute to the regression. Additionally, our goal remains estimating the coefficients γ_{kl} and subsequently making accurate classifications. The next section describes a low-rank and sparse shrinkage prior on network coefficients to achieve these goals.

2.3 Additive low rank and sparse shrinkage prior on network predictor coefficients

Ordinary regression with high dimensional vector predictors has recently been of interest in Bayesian statistics. An overwhelming literature in the last decade has focused on shrinkage priors which shrink coefficients corresponding to unimportant variables close to zero while minimizing the shrinkage of coefficients corresponding to influential variables (for e.g., see Park and Casella, 2008, Armagan *et al.*, 2013, Carvalho *et al.*, 2010). 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 (3), a global-local scale mixture prior would take the form

$$\gamma_{kl} | \phi_{kl}, \sigma^2 \sim N(0, \sigma^2 \phi_{kl}), \;\; \sigma^2 \sim H_1(\cdot), \;\; \phi_{kl} \sim H_2(\cdot), \tag{4}$$

where σ^2 is the global scale parameter and ϕ_{kl}'s are the predictor specific scale parameters controlling the shrinkage of γ_{kl}'s. The Bayesian Lasso shrinkage prior emerges by considering $H_1(\sigma^2) = \delta_1(\sigma^2)$ ($\delta_1(\cdot)$ is the Dirac-delta function at 1) and $H_2(\phi_{kl})$ as a double exponential

density. On the other hand, for the popular horseshoe shrinkage prior $H(\cdot), H_2(\cdot)$ is defined in a way such that $\phi_{kl} \sim C^+(0,1)$ and $\sigma \sim C^+(0,1)$. Unlike the discrete mixture prior distributions (George and McCulloch, 1993), the shrinkage prior on $\boldsymbol{\gamma}$ assigns zero probability at the point zero, thus the exact number of nonzero elements of $\boldsymbol{\gamma}$ is always $q = V(V-1)/2$.

A direct application of this global-local (GL) prior on coefficients $\boldsymbol{\gamma}$ in (3) misses out on important restrictions on $\boldsymbol{\gamma}$ imposed by the network structure in $\boldsymbol{M}_i$. To elucidate further, note that if node k contributes minimally to the response, one would expect to have smaller estimates for most 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 above does not necessarily conform to such an important restriction.

To capture the interaction between nodes in modeling a network, Hoff, 2005 employs the low-rank bilinear model with node specific latent variables. Low rank approximation of matrices is a popular technique of dimension reduction in many areas of science and engineering (Fazel *et al.*, 2003), including matrix completion problems, principal component analysis and factor analysis. As an extension to low-rank approximations, decomposing a given matrix into sparse and low-rank components has gained considerable interest, with applications in video surveillance (Candès *et al.*, 2011), neuroimaging and recommender systems. Finding the best low-rank plus sparse representation of an observed matrix via rank constrained optimization is computationally expensive due to the nonconvex nature of the problem. It has been noted that the low-rank approximation may be too restrictive and not robust, and a low-rank plus sparse framework (Fan *et al.*, 2013; Luo, 2011) is more stable and yields more accurate inference.

Motivated by the literature on low-rank and sparse decomposition of unsupervised analysis of matrices, we propose an additive "low-rank and sparse" framework $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_1 + \boldsymbol{\Gamma}_2$ for the network predictor coefficient matrix. Our proposal assigns a "low-rank" and a "sparse" structure for $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$, respectively. To elaborate further, 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. Define $\boldsymbol{U} = [\boldsymbol{u}_1 : \cdots : \boldsymbol{u}_V]$ as the matrix of node specific latent variables and $\boldsymbol{\Gamma}_1 = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'$, where $\boldsymbol{\Lambda}$ is an $R \times R$ diagonal matrix with the rth diagonal entry $\lambda_r \in \{0,1\}$ signifying the effect of the rth dimension of the latent variables on $\boldsymbol{\Gamma}_1$. 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.

Let $\boldsymbol{\Gamma}_2 = ((\gamma_{2,kl}))_{k,l=1}^V$. To impose a sparse structure on $\boldsymbol{\Gamma}_2$, one possible option is to assign a discrete mixture prior distribution on $\gamma_{2,kl}$. These priors have the advantage of inducing exact sparsity on a subset of parameters, but may face computational challenges when the number of predictors becomes moderately large. Instead, we propose to assign the Bayesian Lasso shrinkage prior of the form (4) to each $\gamma_{2,kl}$, with an additional symmetry restriction $\gamma_{2,kl} = \gamma_{2,lk}$. Hence, conditional on the global and local scale parameters, the γ_{kl}'s

are independent and are represented by the sum of two components:

$$\gamma_{kl} = \boldsymbol{u}_k' \boldsymbol{\Lambda} \boldsymbol{u}_l + \gamma_{2,kl}, \;\; k < l \tag{5}$$

Note that *low-rank only* modeling of $\boldsymbol{\Gamma}$ may be a bit restrictive, and an additional sparse structure provides extra flexibility in estimating more general $\boldsymbol{\Gamma}$ coefficient. In fact, in the context of brain connectome applications, the low-rank component $\boldsymbol{u}_k' \boldsymbol{\Lambda} \boldsymbol{u}_l$ mainly captures the impact of the (k,l)th edge (also referred to as the (k,l)th edge effect) on the regression function due to the interaction between the kth and the lth nodes. The impact of the (k,l)th edge after accounting for the node effects is captured by $\gamma_{2,kl}$. The node k is considered "inactive" in explaining the response if $\boldsymbol{u}_k = 0$, while an edge between the kth and the lth node is "inactive" if either the kth or the lth node is inactive as well as the additional edge effect $\gamma_{2,kl} = 0$. We assume that the "edge effects" are mainly caused by the interactions between nodes, and hence most of the $\gamma_{2,kl}$'s are zero. For computational ease, Bayesian Lasso shrinkage priors are assigned on $\gamma_{2,kl}$'s (see (4)) so that the posterior distribution of unimportant $\gamma_{2,kl}$'s are centered around zero with small variability but never exactly coincide with zero.

In order to directly make inference on which nodes are "active", we assign the *spike-and-slab* (Ishwaran and Rao, 2005) mixture distribution prior on the latent factor $\boldsymbol{u}_k$ as below

$$\boldsymbol{u}_k \sim \begin{cases} N(\boldsymbol{0}, \boldsymbol{Q}), & \text{if } \xi_k = 1 \\ \delta_{\boldsymbol{0}}, & \text{if } \xi_k = 0 \end{cases}, \;\; \xi_k \sim Ber(\Delta), \;\; \boldsymbol{Q} \sim IW(\boldsymbol{S}, \nu), \;\; \Delta \sim U(0,1) \tag{6}$$

where $\delta_{\boldsymbol{0}}$ is the Dirac-delta function at $\boldsymbol{0}$ and $\boldsymbol{Q}$ is a covariance matrix of order $R \times R$. $\boldsymbol{S}$ 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. 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 impacting the response. The location parameter μ is assigned a flat prior distribution.

In order to learn how many components of $\boldsymbol{u}_k$ are informative for (5), 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.

3 Posterior Contraction of the Binary Network Classification Model

This section establishes convergence results for (1) with $\gamma_{2,kl}$'s following the Bayesian Lasso shrinkage prior. From the hierarchical specification given in (4), the Bayesian Lasso shrinkage prior is given by $\gamma_{kl}|\phi_{kl} \sim N(0, \phi_{kl})$, $\phi_{kl} \sim Exp(\lambda_n/2)$. For the theoretical study, a common practice is to fix λ_n as a function of n (see Armagan *et al.*, 2013). Our theoretical investigations will also fix λ_n with the fixed values specified later.

To begin with, we consider an asymptotic setting in which the dimensions of the network predictor grow with n. This paradigm attempts to capture the fact that the number of elements in $\boldsymbol{M}_i$, given by V_n^2 can be substantially larger than sample size n. Since model (1) is equivalent to model (3), the size of the coefficient $\boldsymbol{\gamma}$ in (3) is also a function of n, given by $q_n = \frac{V_n(V_n-1)}{2}$. This creates theoretical challenges, related to (but distinct from) those faced in showing posterior consistency for high dimensional continuous (Armagan *et al.*, 2013) and binary regressions (Wei and Ghosal, 2017). Without loss of generality, we assume that the centering parameter $\mu = 0$ in both the true and the data generating models.

Let $\boldsymbol{y}_n = (y_1, ..., y_n)'$ and the log-likelihood function is given by

$$w_{\boldsymbol{\gamma},n}(\boldsymbol{y}_n) = \sum_{i=1}^{n}[(\boldsymbol{x}_i'\boldsymbol{\gamma})y_i - z(\boldsymbol{x}_i'\boldsymbol{\gamma})], \;\; z(\boldsymbol{x}_i'\boldsymbol{\gamma}) = \log(1 + \exp(\boldsymbol{x}_i'\boldsymbol{\gamma})).$$

We use superscript (0) in order to indicate true parameters. Thus, the true data generating model is given by

$$y_i \sim Bernoulli\left[\frac{\exp(\psi_i^{(0)})}{1+\exp(\psi_i^{(0)})}\right], \;\; \psi_i^{(0)} = \langle \boldsymbol{M}_i, \boldsymbol{\Gamma}^{(0)} \rangle_F. \tag{7}$$

$\boldsymbol{\Gamma}^{(0)}$ is assumed to have a "low-rank plus sparse structure." To be more precise, let $\boldsymbol{\Gamma}^{(0)} = \boldsymbol{\Gamma}_1^{(0)} + \boldsymbol{\Gamma}_2^{(0)}$, where $\boldsymbol{\Gamma}_1^{(0)} = \boldsymbol{U}^{(0)'}\boldsymbol{U}^{(0)}$ with $\boldsymbol{U}^{(0)}$ being an $R_0 \times V_n$ matrix having the rth column $\boldsymbol{u}_k^{(0)}$, $k = 1, ..., V$. $\boldsymbol{\Gamma}_2^{(0)}$ is a sparse matrix, and we denote the number of nonzero elements of $\boldsymbol{\gamma}_2^{(0)}$ by $s_{2,n}^0$, i.e. $||\boldsymbol{\gamma}_2^{(0)}||_0 = s_{2,n}^0$. Also, let $\mathcal{S}^0 = \{j \in \mathbb{N}^2 : \gamma_j^{(0)} \neq 0\}$ denote the indices of the true nonzero coefficients in (3). Similarly denote $\boldsymbol{\phi}_{\mathcal{S}^0} = (\phi_j : j \in \mathcal{S}^0)$ as the vector of ϕ_{kl}'s corresponding to the indices $\mathcal{S}^0$.

We introduce the function $C_{\boldsymbol{y}_n,n}(\cdot)$ to quantify the curvature of $w_{\boldsymbol{\gamma},n}(\boldsymbol{y}_n)$ around $\boldsymbol{\gamma}^{(0)}$,

$$C_{\boldsymbol{y}_n,n}(\boldsymbol{\gamma}) = w_{\boldsymbol{\gamma},n}(\boldsymbol{y}_n) - w_{\boldsymbol{\gamma}^{(0)},n}(\boldsymbol{y}_n) - \nabla w_{\boldsymbol{\gamma}^{(0)},n}(\boldsymbol{y}_n)'(\boldsymbol{\gamma} - \boldsymbol{\gamma}^{(0)}), \tag{8}$$

where $\nabla w_{\boldsymbol{\gamma}^{(0)},n}(\boldsymbol{y}_n)$ is the derivative of $w_{\boldsymbol{\gamma}^{(0)},n}(\boldsymbol{y}_n)$ w.r.t. $\boldsymbol{\gamma}$, evaluated at $\boldsymbol{\gamma}^{(0)}$.

Define $\mathcal{A}_n = \left\{\boldsymbol{\gamma}_n : \frac{1}{n}\sum_{i=1}^{n}|f_{\boldsymbol{\gamma}}(\boldsymbol{x}_i) - f_{\boldsymbol{\gamma}^{(0)}}(\boldsymbol{x}_i)| > \epsilon\right\}$ as a neighborhood around the true density. Further suppose $\pi_n(\cdot)$ and $\Pi_n(\cdot)$ are the prior and posterior densities of $\boldsymbol{\gamma}$ with n

observations, so that

$$\Pi_n(\mathcal{A}_n) = \frac{\int_{\mathcal{A}_n} p_{\boldsymbol{\gamma}}(\boldsymbol{y}_n)\pi_n(\boldsymbol{\gamma})}{\int p_{\boldsymbol{\gamma}}(\boldsymbol{y}_n)\pi_n(\boldsymbol{\gamma})},$$

where $p_{\boldsymbol{\gamma}_n}(\boldsymbol{y}_n) = \prod_{i=1}^{n} \exp(w_{\boldsymbol{\gamma},n}(y_i))$.

3.1 Main Results

To show the posterior contraction result, we make a couple of simplifications. It is assumed that the dimension R of $\boldsymbol{u}_k$ is fixed and is the same as R_0, the dimension of $\boldsymbol{u}_k^{(0)}$. Consequently, *effective dimensionality* is not required to be estimated, and hence $\boldsymbol{\Lambda} = \boldsymbol{I}$ is a non-random matrix. Additionally, we assume $\boldsymbol{Q}$ to be non-random and $\boldsymbol{Q} = \boldsymbol{I}$. We emphasize that both these assumptions are *not* essential for the posterior contraction rate result to be true, but are only introduced for simplifying calculations.

To begin with, we state the following assumptions under which posterior contraction will be shown.

(A) $\sup_{r=1,..,R;k=1,..,V_n} |u_{kr}^{(0)}| < \infty$;

(B) $V_n = o(\frac{n}{\log(n)})$;

(C) $||\boldsymbol{M}_i||_\infty$ is bounded for all $i = 1, .., ,$ w.l.o.g assume $||\boldsymbol{M}_i||_\infty \leq 1$.

(D) $s_{2,n}^0 = o\left\{\frac{n^{1-\rho/2}}{\sqrt{q_n}\log(n)}\right\}$, for some $\rho \in (0, 2)$;

(E) $||\boldsymbol{\Gamma}_2^{(0)}||_\infty < \infty$;

(F) $\lambda_n = \frac{C}{q_n n^{\rho/2}\log(n)}$ for some $C > 0$.

The following theorem shows contraction of the posterior asymptotically under mild sufficient conditions on $V_n, s_{2,n}^0$. The proof of the theorem is provided in Appendix C.

Theorem 3.1 *Under assumptions (A)-(F) for the Bayesian Lasso prior on* $\boldsymbol{\gamma}_2$, $P_{\boldsymbol{\gamma}^{(0)}}(\Pi(\mathcal{A}_n) \to 0) \geq 1 - \frac{2}{q_n}$.

4 Posterior Computation

Using the result in Polson *et al.*, 2013, the data augmented representation of the distribution of y_i given in (2) follows as below

$$p(y_i|\omega_i) = 2^{-b}\exp(k_i\psi_i)\exp(-\omega_i\psi_i^2/2), \;\; \omega_i \sim PG(1, 0), \tag{9}$$

where $k_i = y_i - 1/2$. 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 $\boldsymbol{X} = (\boldsymbol{x}_1 : \cdots : \boldsymbol{x}_n)'$ is an $n \times q$ matrix. Then the conditional likelihood of $\boldsymbol{y} = (y_1, ..., y_n)'$ given $\boldsymbol{\omega} = (\omega_1, ..., \omega_n)'$ and $\boldsymbol{\gamma}$ is given by

$$\begin{aligned} p(\boldsymbol{y} \,|\, \boldsymbol{X}, \boldsymbol{\gamma}, \boldsymbol{\omega}) &\propto \prod_{i=1}^{n} p(y_i \,|\, \boldsymbol{x}_i, \boldsymbol{\gamma}, \omega_i, ...) \\ &\propto \prod_{i=1}^{n} \exp\left\{(y_i - 0.5)(\mu + \boldsymbol{x}_i'\boldsymbol{\gamma}) - \omega_i(\mu + \boldsymbol{x}_i'\boldsymbol{\gamma})^2/2\right\} \\ &\propto \prod_{i=1}^{n} \exp\left\{-\frac{\omega_i}{2}\left[\frac{(y_i - 0.5)}{\omega_i} - (\mu + \boldsymbol{x}_i'\boldsymbol{\gamma})^2\right]\right\} \end{aligned}$$

In matrix notation, the likelihood may be written as

$$p(\boldsymbol{y} \,|\, \boldsymbol{X}, \boldsymbol{\gamma}, \boldsymbol{\omega}...) \propto N(\boldsymbol{z} \,|\, \mu\mathbf{1} + \boldsymbol{X}\boldsymbol{\gamma}, \boldsymbol{\Omega}^{-1})$$

where $\boldsymbol{z} = ((y_1 - 0.5)/\omega_1, ..., (y_n - 0.5)/\omega_n)'$ and $\boldsymbol{\Omega} = diag(\omega_1, ..., \omega_n)$. The full conditional distributions of the parameters are in closed form given in Appendix A. The posterior computation proceeds by running the Gibbs sampler with the full conditional distributions.

Let $\boldsymbol{\Omega}^{(1)}, ..., \boldsymbol{\Omega}^{(L)}$, $\boldsymbol{\Gamma}^{(1)}, ..., \boldsymbol{\Gamma}^{(L)}$ and $\mu^{(1)}, ..., \mu^{(L)}$ be the L post burn-in MCMC samples for $\boldsymbol{\Omega}$, $\boldsymbol{\Gamma}$ and μ respectively after suitable thinning. To classify a network $\boldsymbol{M}_*$ as a member of one of the two groups, we compute $S^{(l)} = \frac{\exp(\mu^{(1)} + \langle \boldsymbol{M}_*, \boldsymbol{\Gamma}^{(l)} \rangle)}{1 + \exp(\mu^{(1)} + \langle \boldsymbol{M}_*, \boldsymbol{\Gamma}^{(l)} \rangle)}$ for $l = 1, ..., L$. $\boldsymbol{M}_*$ 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 0.5, respectively. To judge sensitivity to the choice of the cut-off, the simulation section presents Area under Curve (AUC) of ROC curves with True Positive Rates (TPR) and False Positive Rates (FPR) of classification corresponding to a range of cut-off values.

In order to judge the importance of the kth node in terms of predicting the response, we rely on the post burn-in L samples $\xi_k^{(1)}, ..., \xi_k^{(L)}$ of ξ_k. Also, 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.

5 Simulation Studies

This section evaluates the inferential and predictive ability of our method, along with a number of competitors, using synthetic networks generated under various simulation settings. In each simulation, we assess the ability of the proposed approach to correctly identify influential nodes, to accurately estimate predictive edge coefficients and to classify a network with precise characterization of uncertainties. In this section, we evaluate the performance of our method using synthetic networks. Specific goals are recorded in the description of each simulation. Our proposed approach is referred to as the *Bayesian Network Classifier* (BNC). As competitors, we consider both penalized likelihood methods as well as Bayesian shrink-

age priors for binary high-dimensional regression. Classification performance of different competitors are assessed using the area under the curve (AUC) of the ROC curve.

To study all competitors under various data generation schemes, we simulate the response from (1) given by

$$y_i \sim Ber\left(\frac{\exp(\mu_0 + \langle \boldsymbol{M}_i, \boldsymbol{\Gamma}_0\rangle)}{1+\exp(\mu_0 + \langle \boldsymbol{M}_i, \boldsymbol{\Gamma}_0\rangle)}\right), \ \boldsymbol{\Gamma}_0 = \boldsymbol{\Gamma}_{0,1} + \boldsymbol{\Gamma}_{0,2} \tag{10}$$

where $\boldsymbol{\Gamma}_{0,1}$ and $\boldsymbol{\Gamma}_{0,2}$ are symmetric matrices with zero diagonal entries. μ_0 is fixed at 2 in all simulation scenarios. We consider two different schemes of generating the network $\boldsymbol{M}_i$, referred to as *Simulation 1* and *Simulation 2*.

Simulation 1. In *Simulation 1*, the network edges (i.e. the elements of the matrix $\boldsymbol{M}_i$) are simulated from $\mathrm{N}(0,1)$. Thus, *Simulation 1* generates dense networks with all nodes having inter-connections.

Simulation 2. In *Simulation 2*, nodes in a simulated network are organized into communities so that nodes in the same community tend to have stronger connections than nodes belonging to different communities. This simulation scenario simulates networks which closely mimic brain connectome networks (Bullmore and Sporns, 2009). To simulate networks with such community structures, we assign each node a community label, $A_k \in \{1, 2, ..., K\}$, $k = 1, ..., V$. The node assignments are the same for all networks in the population. Given the community labels, the edge weights for 'active' edges are simulated from a Gaussian distribution. More specifically, the (k, k')th element of $\boldsymbol{M}$ is simulated from $N(m_{A_k A_{k'}}, \sigma_0^2)$, where $m_{kl} = 0.5$ when $k = l$. When $k \neq l$, i.e. the concerned edges connect nodes belonging to different clusters, 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. We set $\sigma_0^2 = 1$ and $K = 3$ with 8, 9 and 8 nodes in the three communities.

Simulating the network predictor coefficient $\boldsymbol{\Gamma}_{0,1}$ *and* $\boldsymbol{\Gamma}_{0,2}$. In both Simulations 1 and 2, $\boldsymbol{\Gamma}_{0,1}$ is $\boldsymbol{\Gamma}_{0,2}$ assume a low-rank and a sparse structure respectively. Specifically, to simulate $\boldsymbol{\Gamma}_{0,1}$, we draw V latent variables $\boldsymbol{u}_{k,0}$, each of dimension R_g, from a mixture distribution given by

$$\boldsymbol{u}_{k,0} \sim \pi N_{R_g}(\boldsymbol{u}_{m,g}, u_{s,g}^2) + (1-\pi)\boldsymbol{\delta_0}; \ k \in \{1, ..., V\}, \tag{11}$$

where $\boldsymbol{\delta_0}$ is the Dirac-delta function and π is the probability of any $\boldsymbol{u}_{k,0}$ being nonzero. The (k,l)th element of the low-rank coefficient $\boldsymbol{\Gamma}_{0,1}$ is given by $\frac{\boldsymbol{u}'_{k,0}\boldsymbol{u}_{l,0}}{2}$, $k \neq l$. Note that if $\boldsymbol{u}_{k,0}$ is zero, then the kth node has no contribution to the mean function in (10), i.e., the kth node becomes inactive in predicting the response. Since $(1-\pi)$ is the probability of a node being inactive, it is referred to as the *node sparsity* parameter in the context of the data generation mechanism under *Simulations 1* and *2*. All elements of $\boldsymbol{u}_{m,g}$ are taken to be 0.5 and $u_{s,g}$ is taken to be 1.

To simulate $\boldsymbol{\Gamma}_{0,2}$, we set π_2, the proportion of nonzero elements of $\boldsymbol{\Gamma}_{0,2}$, randomly at either 0.05 or 0.1. π_2 is referred to as the *residual edge sparsity*. Once the locations of nonzero entries are simulated, the nonzero entries are drawn using one of the three following

strategies:
Strategy 1: Nonzero entries are simulated from $\mathrm{N}(1, 0.1)$.
Strategy 2: Nonzero entries are simulated from $\mathrm{N}(0.5, 0.1)$.
Strategy 3: All nonzero entries are fixed at 0.5.

For a comprehensive picture of *Simulation 1* and *Simulation 2*, we consider different cases as summarized in Table 1 and Table 2, respectively. In each of these cases, the network predictor coefficient and the response are generated by changing the node sparsity π, the residual edge sparsity π_2 and the true dimension R_g of the latent variables $\boldsymbol{u}_{k,0}$'s. The table also presents the maximum fitted dimension R of the latent variables $\boldsymbol{u}_k$ for the logistic regression model (2). Note that the various cases also allow model mis-specification with unequal choices of R and R_g.

Cases	R_g	R	Node Sparsity (π)	Residual Edge Sparsity (π_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	0.5	0.95	Strategy 1
Case - 4	3	4	0.7	0.95	Strategy 1
Case - 5	3	5	0.4	0.95	Strategy 1
Case - 6	2	5	0.5	0.90	Strategy 2
Case - 7	2	5	0.6	0.90	Strategy 3

Table 1: Table presents different cases for ***Simulation 1***. The true dimension R_g 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.

Cases	R_g	R	Node Sparsity (π)	Residual Edge Sparsity (π_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	0.3	0.95	Strategy 1
Case - 4	2	5	0.6	0.90	Strategy 3

Table 2: Table presents different cases for ***Simulation 2***. The true dimension R_g 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.

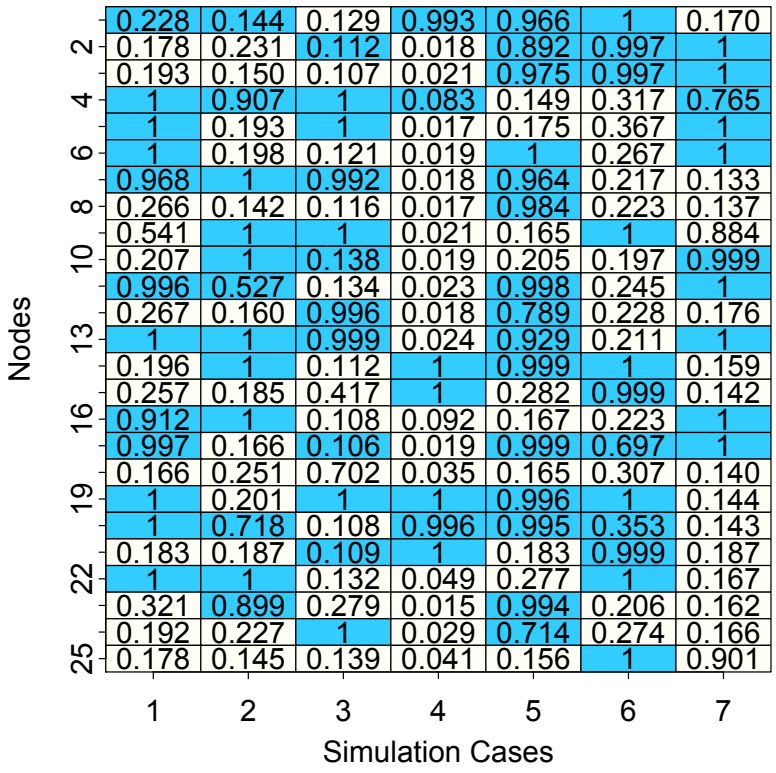


Figure 1: ***Simulation 1***: True activity status of a network node (clear blackground denotes *inactive* and dark background denotes *active*). Note that there are 25 rows (corresponding to 25 nodes) and 7 columns corresponding to 7 different cases in *Simulation 1*. The model-detected posterior probability of the node being active has been super-imposed onto the corresponding cell.

Competitors: As competitors, we use generic variable selection and shrinkage methods that treat edges between nodes together as a long predictor vector to run high dimensional regression, thereby ignoring the relational nature of the predictor. More specifically, we use Lasso (Tibshirani, 1996), which is a popular penalized optimization scheme, and the Bayesian Lasso (BLasso for short)(Park and Casella, 2008) and Horseshoe (BHS for short) priors (Carvalho *et al.*, 2010), which are popular Bayesian shrinkage regression methods, all three under the logistic regression framework. In particular, 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. We use the `glmnet` package in `R` (Friedman *et al.*, 2010) to implement binary Lasso regression. A thorough comparison with these methods indicate the relative advantage of exploiting the structure of the network predictor. We also implement an additional competitor that replaces the Bayesian Lasso shrinkage prior specification on the elements of $\boldsymbol{\Gamma}_2$ by a Horseshoe shrinkage prior. Similar to ours, this

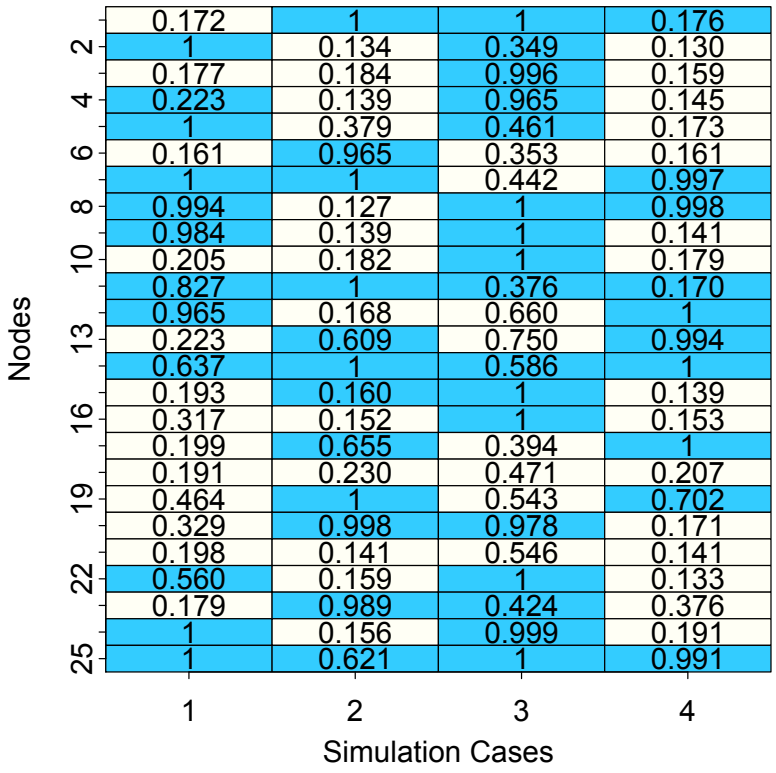
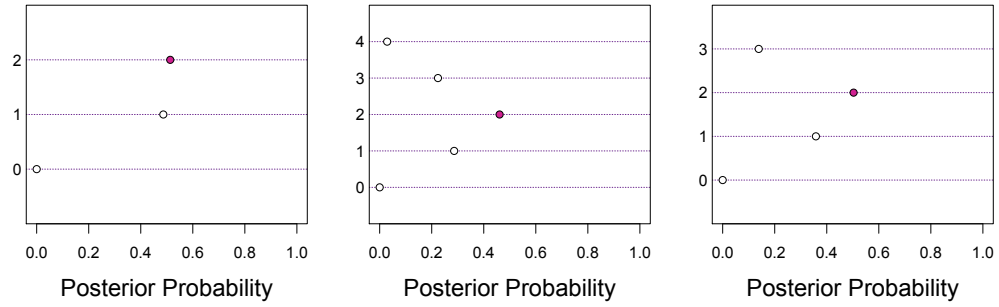


Figure 2: ***Simulation 2***: True activity status of a network node (clear blackground denotes *inactive* and dark background denotes *active*). Note that there are 25 rows (corresponding to 25 nodes) and 4 columns corresponding to 4 different cases in *Simulation 2*. The model-detected posterior probability of the node being active has been super-imposed onto the corresponding cell.

competitor also proposes a low-rank plus sparse shrinkage prior on the network coefficient. We refer to this method as the Bayesian Network Horseshoe (BNH for short). This will help us ascertain if there is any added advantage in replacing the Bayesian Lasso prior on the elements of $\mathbf{\Gamma}_2$ by a more complex structured shrinkage prior.

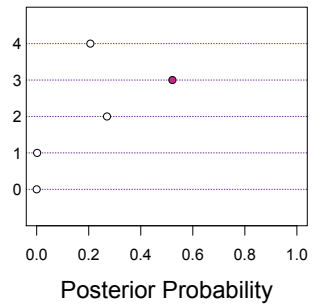
Additionally, we compare our method to a frequentist approach that develops network classification in the presence of a network predictor and binary response (Relión *et al.*, 2017). Relión *et al.*, 2017 develop a penalty on the network predictor coefficient that enables important node selection along with classification of the networks. They argue that their proposed penalty on the coefficient matrix incorporates the network information of the predictor, thereby yielding superior inference to any ordinary penalized optimization scheme. Hence comparison with Relión *et al.*, 2017 will potentially highlight the advantages of a carefully structured Bayesian network shrinkage prior over the penalized optimization scheme incorporating network information. In the absence of any open source code, we implement



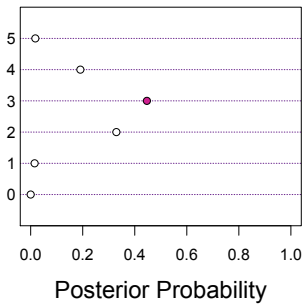
(a) Case 1

(b) Case 2

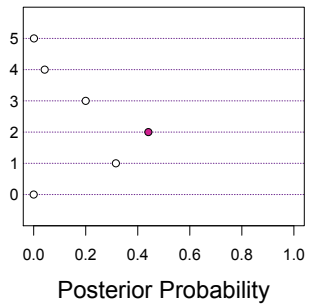
(c) Case 3



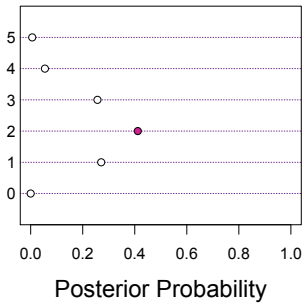
(d) Case 4



(e) Case 5

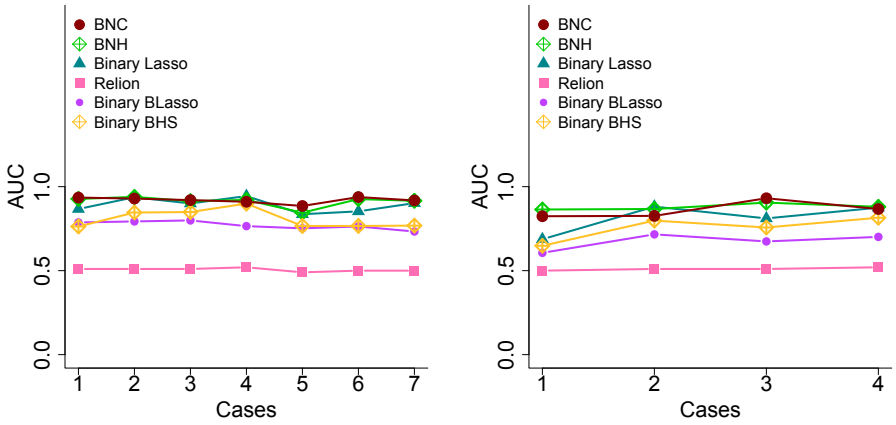


(f) Case 6



(g) Case 7

Figure 3: Plots showing posterior probability distribution of effective dimensionality in all 7 cases in ***Simulation 1***. Filled bullets indicate the true value of effective dimensionality.



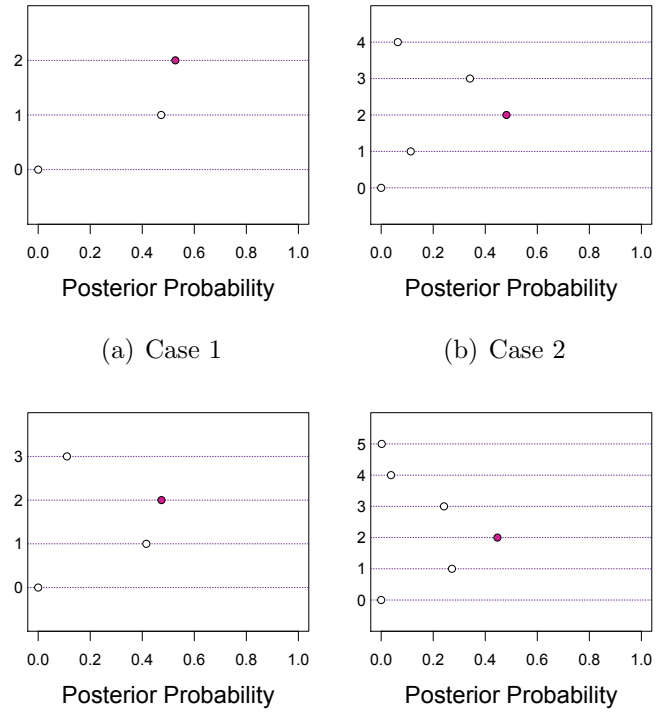
(a) ***Simulation 1*** (b) ***Simulation 2***

Figure 4: Figure shows predictive performance in the form of Area under Curve (AUC) of ROC for all cases in ***Simulations 1*** and ***2***.

the algorithm in Relión *et al.*, 2017 to the best of our ability. All Bayesian competitors are allowed to draw $50,000$ 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 *et al.*, 2014b). All posterior inference is carried out based on the rest $20,000$ MCMC samples after suitably thinning the post burn-in chain. We monitor the auto-correlation plots and effective sample sizes of the iterates, and they are found to be satisfactorily uncorrelated. In all of our simulations, we set $V = 25$ nodes and $n = 250$ samples.

6 Brain Connectome Application

In this section, we present the inferential and predictive ability of Bayesian network classification in the context of a weighted diffusion tension imaging (DTI) dataset. Our 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. It is derived from administration of selected subtests from the Wechsler Intelligence Scales (WIS), designed to provide a measure of an individual's overall level of general cognitive and intellectual functioning, and is a summary score derived from an individual's performance on a variety of tasks that measure acquired knowledge, verbal reasoning, attention to verbal materials, fluid reasoning, spatial processing, attentiveness to details, and visual-motor integration (Caplan *et al.*, 2011). We have converted the FSIQ scores into a binary response variable $\mathbf{y}$, which takes value 0 if FSIQ is less or equal to 120, and value 1 if FSIQ is greater than 120. Thus, we classify the subjects in our study as belonging to the



(a) Case 1 (b) Case 2

(c) Case 3 (d) Case 4

Figure 5: Plots showing posterior probability distribution of effective dimensionality in all 4 cases in ***Simulation 2***. Filled bullets indicate the true value of effective dimensionality.

	MSE					
Cases	**BNC**	BNH	Binary Lasso	Relión(2017)	Binary BL	Binary Horseshoe
Case - 1	**0.164**	0.983	1.197	1.387	0.980	1.160
Case - 2	**0.132**	0.606	0.796	1.004	0.611	0.728
Case - 3	**0.109**	0.303	0.505	0.757	0.407	0.542
Case - 4	**0.064**	0.118	0.187	0.386	0.129	0.153
Case - 5	**2.349**	3.568	3.943	4.368	3.502	3.993
Case - 6	**0.106**	0.467	0.906	1.056	0.695	0.856
Case - 7	**0.166**	0.200	0.485	0.617	0.329	0.415

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.

	MSE					
Cases	**BNC**	BNH	Binary Lasso	Relión(2017)	Binary BL	Binary Horseshoe
Case - 1	**0.279**	0.418	0.807	0.939	0.712	0.739
Case - 2	**0.180**	0.388	0.514	0.665	0.423	0.548
Case - 3	**0.134**	0.549	0.906	1.097	0.748	0.883
Case - 4	**0.066**	0.106	0.167	0.221	0.097	0.141

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.

low IQ group if $\mathbf{y} = 0$, and the *high IQ* group if $\mathbf{y} = 1$.

Along with FSIQ measurements, brain connectome information for $n = 114$ subjects is 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 NDMG pre-processing pipeline (Kiar *et al.*, 2016; Kiar *et al.*, 2017a; Kiar *et al.*, 2017b). 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, there 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, representing the brain network 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.

We fit our proposed model with $\mathbf{y}$ as the binary response and the adjacency matrix as the network predictor. Identical prior distributions for all the parameters as in the simulation studies have been used. BNC is fitted with $R = 4$, which is found to be sufficient for this study. The MCMC chain is run for $50,000$ iterations, with the first $30,000$ iterations discarded as burn-in. Convergence is assessed by comparing different simulated sequences of representative parameters started at different initial values (Gelman *et al.*, 2014a). All inference is based on the remaining $20,000$ post burn-in iterates appropriately thinned. Additionally, we monitor the auto-correlation plots and effective sample sizes of the iterates.

6.1 Findings from the Brain Connectome Application

As in the simulation studies, we emphasize on identifying influential brain regions of interest (ROIs) associated with IQ. Figure 6 plots the estimated posterior probability of each ROI being active. The model estimates posterior probabilities close to 1 for 4 ROIs, namely the *parsopercularis, pericalcarine and supramarginal* regions in the left hemisphere, and *parsorbitalis* region in the right hemisphere. For 6 more ROIs, the model is somewhat uncertain, with posterior probabilities of these ROIs being active varying from around 15% to 20%. The model shows strong conviction about the rest being not influential in the variation in IQ.

Figure 7 presents a heatmap of the estimated posterior means of $\boldsymbol{\gamma}$. As expected, the figure shows high sparsity with only 22 edge coefficients having estimated absolute value of the posterior mean greater than 1. Importantly, all "significantly" nonzero edge coefficients are found to be connected to the ROIs estimated to be influential.

In order to examine the predictive ability of the Bayesian network classification model, we report the area under curve (AUC) of the ROC curve for BNC as well with all competing methods. AUC for all competitors is computed using the popular 10-fold cross validation approach. AUC estimates of all competitors presented in Table 5 indicates better performance of BNC. Frequentist binary Lasso turns out to be the second best performer while BNH, BLasso or BHS all perform very similar to a random classifier. Finally, the distribution of the effective dimensionality for the model is investigated and it turns out that the distribution has a mode at 2 with probability ≈ 0.44. Hence, a choice of $R = 4$ seems to be sufficient for the analysis.

7 Summary and Future Work

We develop a binary Bayesian network regression that enables classifying multiple networks with "labeled nodes" into two groups, identifies influential network nodes and predicts the

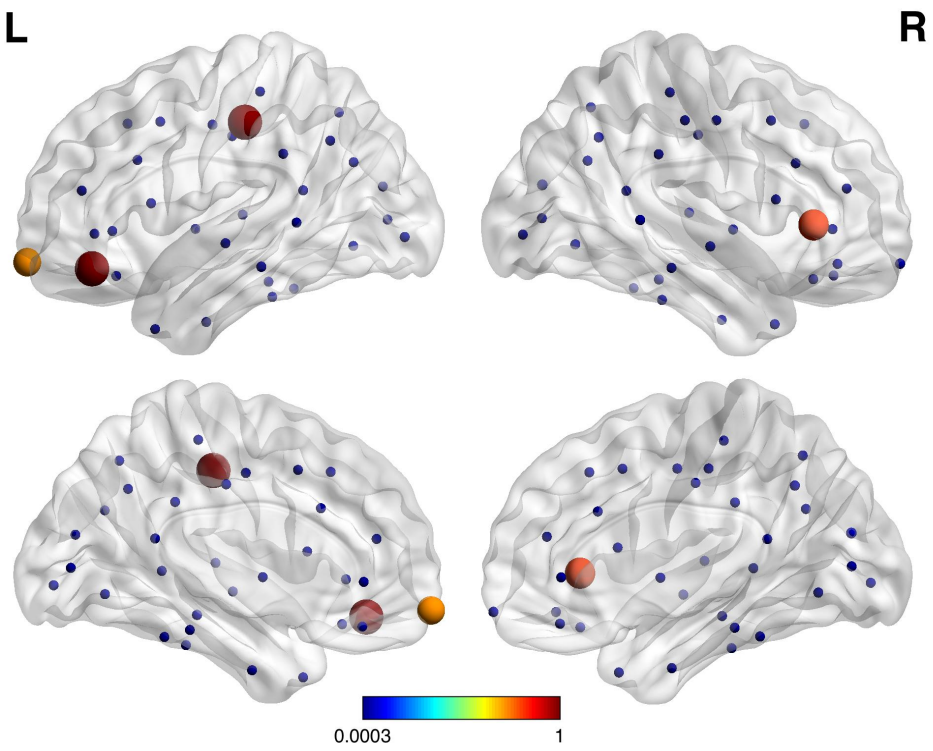


Figure 6: Lateral and medial views of the brain (left and right hemispheres) showing all 68 regions of interest (ROIs). The size and color of the ROIs vary according to the value of the posterior probabilities of them being actively related to the binary response.

Method	BNC	BNH	Binary Lasso	Relión(2017)	Binary BL	Binary BHS
AUC	0.597	0.484	0.532	0.466	0.461	0.484

Table 5: Predictive performance of Bayesian Network Classification (BNC) vis-a-vis competitors in terms of Area Under Curve (AUC) of the ROC. AUC has been calculated in each case using a 10-fold cross validation technique.

class to which a newly observed network belongs. Our contribution lies in carefully constructing an additive low-rank and sparse shrinkage prior on the network predictor coefficient, recognizing the latent network structure in the predictor variable. Another major contribution of the proposed framework remains theoretically understanding the Bayesian network classifier model. Specifically, we develop theory guaranteeing accurate classification as the sample size tends to infinity. The theoretical developments allow the number of possible interconnections in the network predictor to grow at a faster rate than the sample size. Empirical studies reveal advantages of the proposed approach in terms of accurate classification and

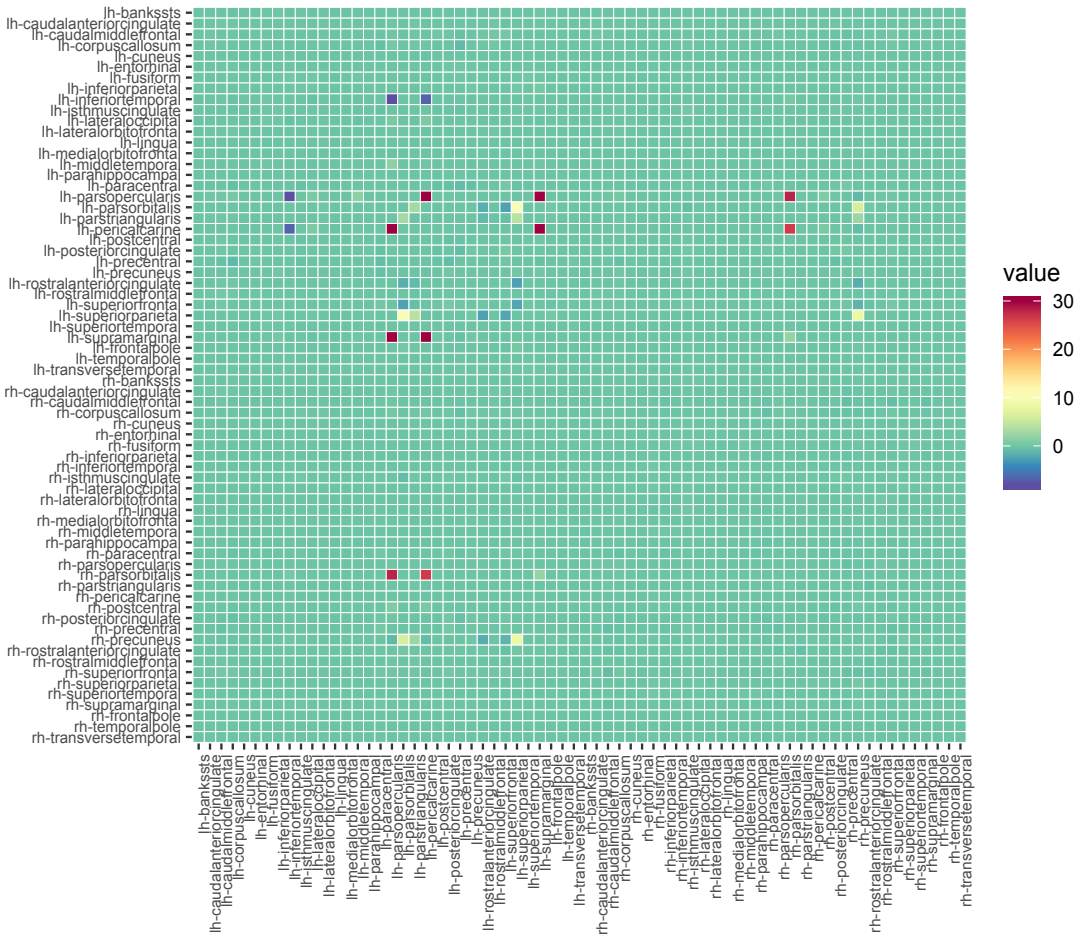


Figure 7: Heatmap showing posterior mean estimates of edge coefficients. Note that the heatmap is a $V \times V$ symmetric matrix, where $V = 68$ denotes the 68 ROIs or nodes, and each cell denotes an edge connecting the corresponding pair of nodes. The axis labels are the abbreviated names of the 68 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.

influential node identification over traditional high dimensional regression techniques which vectorize the network predictor to a high dimensional vector predictor. The framework is employed to analyze a brain connectome dataset that records connectivity between different regions of interest in the brain for multiple individuals and includes information on whether an individual has been found to possess low or high IQ. BNC is able to show satisfactory out of sample classification and identifies important brain regions actively influencing the IQ of an individual.

In future, we hope to develop a network regression model where the response and network

predictors share a nonlinear relationship. 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. We are currently developing models to address these important methodological issues.

Appendix A

This section provides full conditionals for all the parameters in the Bayesian binary network regression presented in Section 2. 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** in the binary setting can be written as

$$\begin{gathered}
\boldsymbol{z} \sim \mathrm{N}(\mu + \boldsymbol{X}\boldsymbol{\gamma}, \boldsymbol{\Omega}^{-1}) \\
\boldsymbol{\gamma} \sim \mathrm{N}(\boldsymbol{W}, \boldsymbol{D}),\ \ \boldsymbol{u}_k|\xi_k = 1 \sim N(\boldsymbol{u}_k \,|\, \boldsymbol{0}, \boldsymbol{M}),\ \ \boldsymbol{u}_k|\xi_k = 0 \sim \delta_{\boldsymbol{0}},\ \ \xi_k \sim Ber(\Delta),\ \ \mu \sim flat() \\
s_{k,l} \sim Exp(\theta^2/2), \quad \theta^2 \sim Gamma(\zeta, \iota), \quad \boldsymbol{M} \sim IW(\boldsymbol{S}, \nu), \quad \Delta \sim Beta(a_\Delta, b_\Delta) \\
p(\omega_i) \sim PG(1, 0),\ \ \lambda_r \sim Ber(\pi_r),\ \ \pi_r \sim Beta(1, r^\eta),\ \eta > 1.
\end{gathered}$$

The full conditional distributions of the model parameters are given below.

- $\mu \,|-\sim N\left(\frac{\boldsymbol{1}'\boldsymbol{\Omega}(\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\gamma})}{\boldsymbol{1}'\boldsymbol{\Omega}\boldsymbol{1}}, \frac{1}{\boldsymbol{1}'\boldsymbol{\Omega}\boldsymbol{1}}\right)$
- $\boldsymbol{\gamma} \,|-\sim N(\boldsymbol{\mu}_{\boldsymbol{\gamma}|\cdot}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}|\cdot})$, where $\boldsymbol{\mu}_{\boldsymbol{\gamma}|\cdot} = (\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X} + \boldsymbol{D}^{-1})^{-1}(\boldsymbol{X}'\boldsymbol{\Omega}(\boldsymbol{z} - \mu\boldsymbol{1}) + \boldsymbol{D}^{-1}\boldsymbol{W})$ and $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}|\cdot} = (\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X} + \boldsymbol{D}^{-1})^{-1}$
- $s_{k,l} \,|- \sim GIG\left[\frac{1}{2}, (\gamma_{k,l} - \boldsymbol{u}_k'\boldsymbol{\Lambda}\boldsymbol{u}_l)^2, \theta^2\right]$, where GIG denotes the generalized inverse Gaussian distribution.
- $\theta^2 \,|- \sim Gamma\left[\left(\zeta + \frac{V(V-1)}{2}\right), \left(\iota + \sum_{k<l} \frac{s_{k,l}}{2}\right)\right]$
- $\boldsymbol{u}_k \,|- \sim w_{\boldsymbol{u}_k}\, \delta_0(\boldsymbol{u}_k) + (1 - w_{\boldsymbol{u}_k})\, N(\boldsymbol{u}_k \,|\, \boldsymbol{m}_{\boldsymbol{u}_k}, \boldsymbol{\Sigma}_{\boldsymbol{u}_k})$, 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

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

- $\xi_k|- \sim Ber(1 - w_{\boldsymbol{u}_k})$
- $\Delta \,|- \sim Beta\left[(a_\Delta + \sum_{k=1}^V \xi_k), (b_\Delta + \sum_{k=1}^V (1 - \xi_k))\right]$.
- $\boldsymbol{M} \,|- \sim IW[(\boldsymbol{S} + \sum_{k:\boldsymbol{u}_k \neq \boldsymbol{0}} \boldsymbol{u}_k\boldsymbol{\Lambda}\boldsymbol{u}_k'), (\nu + \{\#k : \boldsymbol{u}_k \neq \boldsymbol{0}\})]$.

- $\lambda_r \,|\, - \sim Ber(p_{\lambda_r})$, where $p_{\lambda_r} = \frac{\pi_r N(\boldsymbol{\gamma} \,|\, \boldsymbol{W}_1, \tau^2 \boldsymbol{D})}{\pi_r N(\boldsymbol{\gamma} \,|\, \boldsymbol{W}_1, \tau^2 \boldsymbol{D}) + (1-\pi_r) N(\boldsymbol{\gamma} \,|\, \boldsymbol{W}_0, \tau^2 \boldsymbol{D})}$. 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 $r = 1, .., R$.
- $\pi_r \,|\, - \sim Beta(\lambda_r + 1, 1 - \lambda_r + r^\eta)$, for $r = 1, .., R$.

  Using the relationship, $PG(x \,|\, b, c) \propto \exp(-\frac{c^2 x}{2}) PG(x \,|\, 1, 0)$ (Polson *et al.* (2013)), we obtain
- $\omega_i \,|\, - \sim PG(1, \mu + \boldsymbol{x}_i' \boldsymbol{\gamma})$, for $i = 1, .., n$.

Lemma 7.1 *Let $\boldsymbol{\gamma}_{\boldsymbol{W}}$ be a random variable such that*

$$\boldsymbol{\gamma}_{\boldsymbol{W}} \,|\, - \sim \mathrm{N}\left[(\boldsymbol{D}^{-1} + \boldsymbol{X}^T \Omega \boldsymbol{X})^{-1} \boldsymbol{X}^T \Omega (\boldsymbol{z} - \mu \underline{1} - \boldsymbol{X}\boldsymbol{W}), (\boldsymbol{D}^{-1} + \boldsymbol{X}^T \Omega \boldsymbol{X})^{-1}\right]. \tag{12}$$

Then the following results hold.

(a) $\boldsymbol{\gamma} \overset{D}{=} \boldsymbol{\gamma}_{\boldsymbol{W}} + \boldsymbol{W}$

(b) Let, $\boldsymbol{\Delta}_{\gamma_1} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{D})$, $\boldsymbol{\Delta}_{\gamma_2} \sim \mathrm{N}(\boldsymbol{0}, I)$, $\boldsymbol{\Delta}_{\gamma_3} = \Omega^{\frac{1}{2}} \boldsymbol{X} \boldsymbol{\Delta}_{\gamma_1} + \boldsymbol{\Delta}_{\gamma_2}$,
$\boldsymbol{\gamma}_{\boldsymbol{W}} = \boldsymbol{\Delta}_{\gamma_1} + \boldsymbol{D}\boldsymbol{X}^T \Omega^{\frac{1}{2}} (\Omega^{\frac{1}{2}} \boldsymbol{X} \boldsymbol{D} \boldsymbol{X}^T \Omega^{\frac{1}{2}} + \boldsymbol{I})^{-1} \left[\Omega^{\frac{1}{2}} (\boldsymbol{z} - \mu \underline{1} - \boldsymbol{X}\boldsymbol{W}) - \boldsymbol{\Delta}_{\gamma_3}\right]$.

Remark: This algorithm ensures that samples from the posterior full conditionals of $\boldsymbol{\gamma}$ can be obtained by sampling from the posterior full conditionals of $\boldsymbol{\gamma}_{\boldsymbol{W}}$. Lemma 7.1 shows that obtaining samples from the full conditional of $\boldsymbol{\gamma}_{\boldsymbol{W}}$ only requires inverting an $n \times n$ matrix. Assuming $n << q$, which is typically encountered in the real data applications, the computational complexity of the proposed approach is substantially mitigated.

As noted in Section 2.3 of the main text, straightforward posterior draw from the full conditional of $\boldsymbol{\gamma}$ as above faces substantial computational difficulties. Section 2.3 states Lemma 7.1 that provides a computational strategy to draw posterior samples of $\boldsymbol{\gamma}$ efficiently. Proof of Lemma 7.1 is given below.

Proof of Lemma 7.1

(a) Note that

$$\begin{aligned}
E(\boldsymbol{\gamma}_{\boldsymbol{W}} + \boldsymbol{W}) &= \boldsymbol{W} + (\boldsymbol{D}^{-1} + \boldsymbol{X}^T \Omega \boldsymbol{X})^{-1} \boldsymbol{X}^T \Omega (\boldsymbol{z} - \mu \underline{1} - \boldsymbol{X}\boldsymbol{W}) \\
&= \boldsymbol{W} - (\boldsymbol{D}^{-1} + \boldsymbol{X}^T \Omega \boldsymbol{X})^{-1} \boldsymbol{X}^T \Omega \boldsymbol{X} \boldsymbol{W} + (\boldsymbol{D}^{-1} + \boldsymbol{X}^T \Omega \boldsymbol{X})^{-1} \boldsymbol{X}^T \Omega (\boldsymbol{z} - \mu \underline{1}) \\
&= \boldsymbol{W} - (\boldsymbol{D}^{-1} + \boldsymbol{X}^T \Omega \boldsymbol{X})^{-1} (\boldsymbol{D}^{-1} + \boldsymbol{X}^T \Omega \boldsymbol{X} - \boldsymbol{D}^{-1}) \boldsymbol{W} \\
&\quad + (\boldsymbol{D}^{-1} + \boldsymbol{X}^T \Omega \boldsymbol{X})^{-1} \boldsymbol{X}^T \Omega (\boldsymbol{z} - \mu \underline{1}) \\
&= \boldsymbol{W} - (\boldsymbol{I} - (\boldsymbol{D}^{-1} + \boldsymbol{X}^T \Omega \boldsymbol{X})^{-1} \boldsymbol{D}^{-1}) \boldsymbol{W} + (\boldsymbol{D}^{-1} + \boldsymbol{X}^T \Omega \boldsymbol{X})^{-1} \boldsymbol{X}^T \Omega (\boldsymbol{z} - \mu \underline{1}) \\
&= (\boldsymbol{D}^{-1} + \boldsymbol{X}^T \Omega \boldsymbol{X})^{-1} \boldsymbol{D}^{-1} \boldsymbol{W} + (\boldsymbol{D}^{-1} + \boldsymbol{X}^T \Omega \boldsymbol{X})^{-1} \boldsymbol{X}^T \Omega (\boldsymbol{z} - \mu \underline{1}) \\
&= (\boldsymbol{D}^{-1} + \boldsymbol{X}^T \Omega \boldsymbol{X})^{-1} (\boldsymbol{D}^{-1} \boldsymbol{W} + \boldsymbol{X}^T \Omega (\boldsymbol{z} - \mu \underline{1})) = E(\boldsymbol{\gamma}).
\end{aligned}$$

Also note that $Var(\boldsymbol{\gamma_W} + \boldsymbol{W}) = Var(\boldsymbol{\gamma})$ trivially since $\boldsymbol{W}$ is a given in the Gibbs step.

(b) Note that

$$\begin{aligned} E(\boldsymbol{\gamma_W}) &= E\left(\boldsymbol{\Delta}_{\gamma_1} + \boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}}(\Omega^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}} + \boldsymbol{I})^{-1}\left[\Omega^{\frac{1}{2}}(\boldsymbol{z} - \mu\underline{1} - \boldsymbol{X}\boldsymbol{W}) - \boldsymbol{\Delta}_{\gamma_3}\right]\right) \\ &= \boldsymbol{0} + \boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}}(\Omega^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}} + \boldsymbol{I})^{-1}\left[\Omega^{\frac{1}{2}}(\boldsymbol{z} - \mu\underline{1} - \boldsymbol{X}\boldsymbol{W}) - 0\right] \\ &= \boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}}(\Omega^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}} + \boldsymbol{I})^{-1}\Omega^{\frac{1}{2}}(\boldsymbol{z} - \mu\underline{1} - \boldsymbol{X}\boldsymbol{W}). \end{aligned}$$

We need to prove that

$$\boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}}(\Omega^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}} + \boldsymbol{I})^{-1}\Omega^{\frac{1}{2}}(\boldsymbol{z} - \mu\underline{1} - \boldsymbol{X}\boldsymbol{W}) = (\boldsymbol{D}^{-1} + \boldsymbol{X}^T\Omega\boldsymbol{X})^{-1}\boldsymbol{X}^T\Omega(\boldsymbol{z} - \mu\underline{1} - \boldsymbol{X}\boldsymbol{W})$$

$$\text{i.e.} \qquad \boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}}(\Omega^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}} + \boldsymbol{I})^{-1} = (\boldsymbol{D}^{-1} + \boldsymbol{X}^T\Omega\boldsymbol{X})^{-1}\boldsymbol{X}^T\Omega^{\frac{1}{2}}$$

Using the Sherman-Morrison-Woodbury matrix identity, we have that $(\boldsymbol{D}^{-1} + \boldsymbol{X}^T\Omega\boldsymbol{X})^{-1} = (\boldsymbol{D}^{-1} + \boldsymbol{X}^T\Omega^{\frac{1}{2}}\boldsymbol{I}\Omega^{\frac{1}{2}}\boldsymbol{X})^{-1} = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}}(\Omega^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}} + \boldsymbol{I})^{-1}\Omega^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{D}$.
Hence

$$\begin{aligned} &R.H.S. = (\boldsymbol{D}^{-1} + \boldsymbol{X}^T\Omega\boldsymbol{X})^{-1}\boldsymbol{X}^T\Omega^{\frac{1}{2}} \\ &= (\boldsymbol{D} - \boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}}(\Omega^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}} + \boldsymbol{I})^{-1}\Omega^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{D})\boldsymbol{X}^T\Omega^{\frac{1}{2}} \\ &= \boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}} - \boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}}(\Omega^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}} + \boldsymbol{I})^{-1}\Omega^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}} \\ &= \boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}} - \boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}}(\Omega^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}} + \boldsymbol{I})^{-1}\left[\Omega^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}} + \boldsymbol{I} - \boldsymbol{I}\right] \\ &= \boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}} - \boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}}\left[\boldsymbol{I} - (\boldsymbol{I} + \Omega^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}})^{-1}\right] \\ &= \boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}}(\boldsymbol{I} + \Omega^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}})^{-1} = L.H.S. \end{aligned}$$

Using the fact that $Var(\boldsymbol{\Delta}_{\gamma_1}) = \boldsymbol{D}$, $Var(\boldsymbol{\Delta}_{\gamma_2}) = \boldsymbol{I}$, $Var(\boldsymbol{\Delta}_{\gamma_3}) = (\Omega^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^T\Omega^{\frac{1}{2}} + \boldsymbol{I})$

and $Cov(\boldsymbol{\Delta}_{\boldsymbol{\gamma}_1}, \boldsymbol{\Delta}_{\boldsymbol{\gamma}_3}) = \Omega^{\frac{1}{2}}\boldsymbol{XD}$, we have

$$
\begin{aligned}
&Var\left(\boldsymbol{\Delta}_{\boldsymbol{\gamma}_1} + \boldsymbol{DX}^T\Omega^{\frac{1}{2}}(\Omega^{\frac{1}{2}}\boldsymbol{XDX}^T\Omega^{\frac{1}{2}} + \boldsymbol{I})^{-1}\left[\Omega^{\frac{1}{2}}(\boldsymbol{z} - \mu\underline{1} - \boldsymbol{XW}) - \boldsymbol{\Delta}_{\boldsymbol{\gamma}_3}\right]\right) \\
&= Var(\boldsymbol{\Delta}_{\boldsymbol{\gamma}_1}) + \boldsymbol{DX}^T\Omega^{\frac{1}{2}}(\Omega^{\frac{1}{2}}\boldsymbol{XDX}^T\Omega^{\frac{1}{2}} + \boldsymbol{I})^{-1}Var(\boldsymbol{\Delta}_{\boldsymbol{\gamma}_3})(\boldsymbol{DX}^T\Omega^{\frac{1}{2}}(\Omega^{\frac{1}{2}}\boldsymbol{XDX}^T\Omega^{\frac{1}{2}} + \boldsymbol{I})^{-1})^T \\
&\quad - Cov(\boldsymbol{\Delta}_{\boldsymbol{\gamma}_1}, \boldsymbol{DX}^T\Omega^{\frac{1}{2}}(\Omega^{\frac{1}{2}}\boldsymbol{XDX}^T\Omega^{\frac{1}{2}} + \boldsymbol{I})^{-1}\boldsymbol{\Delta}_{\boldsymbol{\gamma}_3}) \\
&\quad - \left[Cov(\boldsymbol{\Delta}_{\boldsymbol{\gamma}_1}, \boldsymbol{DX}^T\Omega^{\frac{1}{2}}(\Omega^{\frac{1}{2}}\boldsymbol{XDX}^T\Omega^{\frac{1}{2}} + \boldsymbol{I})^{-1}\boldsymbol{\Delta}_{\boldsymbol{\gamma}_3})\right]^T \\
&= \boldsymbol{D} + \boldsymbol{DX}^T\Omega^{\frac{1}{2}}(\Omega^{\frac{1}{2}}\boldsymbol{XDX}^T\Omega^{\frac{1}{2}} + \boldsymbol{I})^{-1}(\Omega^{\frac{1}{2}}\boldsymbol{XDX}^T\Omega^{\frac{1}{2}} + \boldsymbol{I})(\Omega^{\frac{1}{2}}\boldsymbol{XDX}^T\Omega^{\frac{1}{2}} + \boldsymbol{I})^{-1}\Omega^{\frac{1}{2}}\boldsymbol{XD} \\
&\quad - 2\boldsymbol{DX}^T\Omega^{\frac{1}{2}}(\Omega^{\frac{1}{2}}\boldsymbol{XDX}^T\Omega^{\frac{1}{2}} + \boldsymbol{I})^{-1}\Omega^{\frac{1}{2}}\boldsymbol{XD} \\
&= \boldsymbol{D} + \boldsymbol{DX}^T\Omega^{\frac{1}{2}}(\Omega^{\frac{1}{2}}\boldsymbol{XDX}^T\Omega^{\frac{1}{2}} + \boldsymbol{I})^{-1}\Omega^{\frac{1}{2}}\boldsymbol{XD} - 2\boldsymbol{DX}^T\Omega^{\frac{1}{2}}(\Omega^{\frac{1}{2}}\boldsymbol{XDX}^T\Omega^{\frac{1}{2}} + \boldsymbol{I})^{-1}\Omega^{\frac{1}{2}}\boldsymbol{XD} \\
&= \boldsymbol{D} - \boldsymbol{DX}^T\Omega^{\frac{1}{2}}(\Omega^{\frac{1}{2}}\boldsymbol{XDX}^T\Omega^{\frac{1}{2}} + \boldsymbol{I})^{-1}\Omega^{\frac{1}{2}}\boldsymbol{XD} \\
&= (\boldsymbol{D}^{-1} + \boldsymbol{X}^T\Omega\boldsymbol{X})^{-1} \qquad \text{(Using the Sherman-Morrison-Woodbury matrix identity)} \\
&= Var(\boldsymbol{\gamma_W})
\end{aligned}
$$

Appendix B

Appendix A shows full conditionals for BNC model. This section provides full conditionals for all the parameters in the Bayesian binary network horseshoe regression used as a competitor for BNC. Assume $\boldsymbol{W} = (\boldsymbol{u}_1'\boldsymbol{\Lambda}\boldsymbol{u}_2, ..., \boldsymbol{u}_1'\boldsymbol{\Lambda}\boldsymbol{u}_V, ..., \boldsymbol{u}_{V-1}'\boldsymbol{\Lambda}\boldsymbol{u}_V)'$, $\boldsymbol{D} = diag(\tau^2 s_{1,2}^2, ..., \tau^2 s_{V-1,V}^2)$ and $\boldsymbol{\gamma} = (\gamma_{1,2}, ..., \gamma_{V-1,V})'$. Thus, with n data points, the hierarchical model with the **Network Horseshoe prior** in the binary setting can be written as

$$
\begin{aligned}
\boldsymbol{z} &\sim \mathrm{N}(\mu + \boldsymbol{X\gamma}, \boldsymbol{\Omega}^{-1}) \\
\boldsymbol{\gamma} \sim \mathrm{N}(\boldsymbol{W}, \boldsymbol{D}),\ \boldsymbol{u}_k|\xi_k = 1 &\sim N(\boldsymbol{u}_k \,|\, \boldsymbol{0}, \boldsymbol{M}),\ \boldsymbol{u}_k|\xi_k = 0 \sim \delta_{\boldsymbol{0}},\ \xi_k \sim Ber(\Delta),\ \mu \sim flat() \\
s_{k,l} \sim C^+(0,1),\quad \tau \sim C^+(0,1),\quad \boldsymbol{M} &\sim IW(\boldsymbol{S}, \nu),\quad \Delta \sim Beta(a_\Delta, b_\Delta) \\
p(\omega_i) \sim PG(1,0),\ \lambda_r &\sim Ber(\pi_r),\ \pi_r \sim Beta(1, r^\eta),\ \eta > 1.
\end{aligned}
$$

Note that, following Makalic and Schmidt (2015),

$$s_{k,l} \sim C^+(0,1),\quad \tau \sim C^+(0,1)$$

can be written in an augmented form as

$$s_{k,l}^2 \,|\, \nu_{k,l} \sim IG\left(\frac{1}{2}, \frac{1}{\nu_{k,l}}\right),\quad \nu_{k,l} \sim IG\left(\frac{1}{2}, 1\right),\quad \tau^2 \,|\, \sigma \sim IG\left(\frac{1}{2}, \frac{1}{\sigma}\right),\quad \sigma \sim IG\left(\frac{1}{2}, 1\right).$$

The full conditional distributions of the model parameters are given below:

- $\mu \,|\, - \sim N\left(\frac{\boldsymbol{1'\Omega(z - X\gamma)}}{\boldsymbol{1'\Omega 1}}, \frac{1}{\boldsymbol{1'\Omega 1}}\right)$

- $\boldsymbol{\gamma} \,|-\sim N(\boldsymbol{\mu}_{\boldsymbol{\gamma}|.}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}|.})$, where $\boldsymbol{\mu}_{\boldsymbol{\gamma}|.} = (\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X} + \boldsymbol{D}^{-1})^{-1}(\boldsymbol{X}'\boldsymbol{\Omega}(\boldsymbol{z} - \mu\mathbf{1}) + \boldsymbol{D}^{-1}\boldsymbol{W})$ and $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}|.} = (\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X} + \boldsymbol{D}^{-1})^{-1}$
- $s^2_{k,l} \,|-\sim IG\left[1, (\frac{1}{\nu_{k,l}} + \frac{(\gamma_{k,l} - \boldsymbol{u}_k'\boldsymbol{\Lambda}\boldsymbol{u}_l)^2}{2\tau^2})\right]$
- $\tau^2 \,|-\sim IG\left[\left(\frac{1}{2} + \frac{V(V-1)}{4}\right), \left(\frac{1}{\sigma} + \sum_{k<l} \frac{(\gamma_{k,l} - \boldsymbol{u}_k'\boldsymbol{\Lambda}\boldsymbol{u}_l)^2}{2s^2_{k,l}}\right)\right]$
- $\nu_{k,l} \,|-\sim IG\left[1, (1 + \frac{1}{s^2_{k,l}})\right]$
- $\sigma \,|-\sim IG\left[1, (1 + \frac{1}{\tau^2})\right]$
- $\boldsymbol{u}_k \,|-\sim w_{\boldsymbol{u}_k}\, \delta_0(\boldsymbol{u}_k) + (1 - w_{\boldsymbol{u}_k})\, N(\boldsymbol{u}_k \,|\, \boldsymbol{m}_{\boldsymbol{u}_k}, \boldsymbol{\Sigma}_{\boldsymbol{u}_k})$, 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

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

$$w_{\boldsymbol{u}_k} = \frac{(1-\pi) N(\boldsymbol{\gamma}_k \,|\, \mathbf{0}, \tau^2 \boldsymbol{H}_k)}{(1-\pi) N(\boldsymbol{\gamma}_k \,|\, \mathbf{0}, \tau^2 \boldsymbol{H}_k) + \pi N(\boldsymbol{\gamma}_k \,|\, \mathbf{0}, \tau^2 \boldsymbol{H}_k + \boldsymbol{U}^*_k \boldsymbol{M} \boldsymbol{U}^{*'}_k)}$$

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

  Using the relationship, $PG(x \,|\, b, c) \propto \exp(-\frac{c^2 x}{2}) PG(x \,|\, b, 0)$ (Polson *et al.* (2013)), we obtain
- $\omega_i \,|-\sim PG(1, \mu + \boldsymbol{x}_i'\boldsymbol{\gamma})$, for $i = 1, .., n$.

Appendix C

The proof of Theorem 3.1 relies in part on the existence of exponentially consistent sequence of tests.

Theorem 7.2 *There exists a sequence of test functions for testing* $H_0 : \boldsymbol{\gamma} = \boldsymbol{\gamma}^0$ *vs.* $H_1 : \boldsymbol{\gamma} \in \mathcal{A}_n$, *which satisfy*

$$E_{\boldsymbol{\gamma}^0}(\Phi_n) \leq \exp(-hn), \qquad \sup_{\boldsymbol{\gamma} \in \mathcal{A}_n} E_{\boldsymbol{\gamma}}(1 - \Phi_n) \leq \exp(-hn). \tag{13}$$

This sequence of test functions are referred to as the exponential consistent sequence of tests Φ_n *for testing* $H_0 : \boldsymbol{\gamma} = \boldsymbol{\gamma}^0$ *vs.* $H_1 : \boldsymbol{\gamma} \in \mathcal{A}_n$.

Proof The construction of the test is provided in the proof of Theorem 2 in Ghosal and Roy (2006).

We also state another result which will be subsequently used in the proof.

Lemma 7.3 *Let* $\boldsymbol{u}_k^{(0)} = (u_{k1}^{(0)}, ..., u_{kR}^{(0)})'$ *for* $k = 1, .., V_n$, *and* υ_{kl} *be the only positive root of the equation*

$$x^2 + x(||\boldsymbol{u}_k^{(0)}||_2|| + ||\boldsymbol{u}_l^{(0)}||_2) - \eta = 0, \;\; k < l. \tag{14}$$

Assume $\upsilon = \min_{k,l} \upsilon_{kl}$. *Then, for* $\boldsymbol{W} = (\boldsymbol{u}_1'\boldsymbol{u}_2, ..., \boldsymbol{u}_{V_n-1}'\boldsymbol{u}_{V_n})'$ *and* $\boldsymbol{W}^{(0)} = (\boldsymbol{u}_1^{(0)'}\boldsymbol{u}_2^{(0)}, ..., \boldsymbol{u}_{V_n-1}^{(0)'}\boldsymbol{u}_{V_n}^{(0)})'$

$$\Pi(||\boldsymbol{W} - \boldsymbol{W}^{(0)}||_\infty < \eta) \geq \Pi(||\boldsymbol{u}_k - \boldsymbol{u}_k^{(0)}||_2 \leq \upsilon, \; \forall \, k = 1, .., V_n). \tag{15}$$

Proof for $k < l$,

$$\begin{aligned}
|\boldsymbol{u}_k'\boldsymbol{u}_l - \boldsymbol{u}_k^{(0)'}\boldsymbol{u}_l^{(0)}| &= |\sum_{r=1}^{R} u_{kr}u_{lr} - \sum_{r=1}^{R} u_{kr}^{(0)}u_{lr}^{(0)}| \\
&= |\sum_{r=1}^{R}(u_{kr} - u_{kr}^{(0)})u_{lr}| + |\sum_{r=1}^{R}(u_{lr} - u_{lr}^{(0)})u_{kr}^{(0)}| \\
&\leq ||\boldsymbol{u}_k - \boldsymbol{u}_k^{(0)}||_2||\boldsymbol{u}_l||_2 + ||\boldsymbol{u}_l - \boldsymbol{u}_l^{(0)}||_2||\boldsymbol{u}_k^{(0)}||_2 \\
&\leq ||\boldsymbol{u}_k - \boldsymbol{u}_k^{(0)}||_2 \left[||\boldsymbol{u}_l - \boldsymbol{u}_l^{(0)}||_2 + ||\boldsymbol{u}_l^{(0)}||_2\right] + ||\boldsymbol{u}_l - \boldsymbol{u}_l^{(0)}||_2||\boldsymbol{u}_k^{(0)}||_2.
\end{aligned}$$

If $||\boldsymbol{u}_k - \boldsymbol{u}_k^{(0)}||_2 \leq \upsilon, \; \forall \, k = 1, .., V_n$, the above inequality implies

$$|\boldsymbol{u}_k'\boldsymbol{u}_l - \boldsymbol{u}_k^{(0)'}\boldsymbol{u}_l^{(0)}| \leq \upsilon(\upsilon + ||\boldsymbol{u}_l^{(0)}||_2) + \upsilon||\boldsymbol{u}_k^{(0)}||_2 \leq \eta, \; \forall \, k < l.$$

Hence $\Pi(||\boldsymbol{W} - \boldsymbol{W}^{(0)}||_\infty < \eta) \geq \Pi(||\boldsymbol{u}_k - \boldsymbol{u}_k^{(0)}||_2 \leq \upsilon, \; \forall \, k = 1, .., V_n)$.

Proof of Theorem 3.1
Suppose $\boldsymbol{y}_n \in \mathcal{E}_n = \left\{\boldsymbol{y} : ||\nabla w_{\boldsymbol{\gamma}^{(0)},n}(\boldsymbol{y})||_\infty \leq 2\sqrt{nq_n}\right\}$. Then

$$P_{\boldsymbol{\gamma}^{(0)}}(\boldsymbol{y}_n \in \mathcal{E}_n) \geq 1 - P_{\boldsymbol{\gamma}^{(0)}}(\max_{1 \leq j \leq q_n} |\sum_{i=1}^{n}(y_i - \nabla z(\boldsymbol{x}_i'(\boldsymbol{\gamma} - \boldsymbol{\gamma}^{(0)})))x_{ij}| > 2\sqrt{nq_n}) \geq 1 - \frac{2}{q_n},$$

where the last step follows from the Hoeffding inequality. In what follows, we will assume that $\boldsymbol{y}_n \in \mathcal{E}_n$. It can be observed that

$$\Pi_n(\mathcal{A}_n) = \frac{\int_{\mathcal{A}_n} p_{\boldsymbol{\gamma}}(\boldsymbol{y}_n)\pi_n(\boldsymbol{\gamma})}{\int p_{\boldsymbol{\gamma}}(\boldsymbol{y}_n)\pi_n(\boldsymbol{\gamma})} = \frac{\int_{\mathcal{A}_n} \frac{p_{\boldsymbol{\gamma}}(\boldsymbol{y}_n)}{p_{\boldsymbol{\gamma}^{(0)}}(\boldsymbol{y}_n)}\pi_n(\boldsymbol{\gamma})}{\int \frac{p_{\boldsymbol{\gamma}}(\boldsymbol{y}_n)}{p_{\boldsymbol{\gamma}^{(0)}}(\boldsymbol{y}_n)}\pi_n(\boldsymbol{\gamma})} = \frac{\mathcal{N}_n}{\mathcal{D}_n} \leq \Phi_n + (1-\Phi_n)\frac{\mathcal{N}_n}{\mathcal{D}_n}, \tag{16}$$

where Φ_n is the exponentially consistent sequence of tests given in Theorem 7.2. In proving Theorem 3.1, we will proceed in three steps as following.

(a) Step 1 shows that $\Phi_n \to 0$, as $n \to \infty$, almost surely.

(b) Step 2 shows that $\exp(hn/2)(1-\Phi_n)\mathcal{N}_n \to 0$, as $n \to \infty$, almost surely.

(c) Finally, step 3 shows that $\exp(hn/2)\mathcal{D}_n \to \infty$, as $n \to \infty$, when $\boldsymbol{y}_n \in \mathcal{E}_n$.

(a) Step 1
(13) in Theorem 7.2 yields,

$$P_{\boldsymbol{\gamma}^{(0)}}\left(\Phi_n > \exp(-nh/2)\right) \leq E_{\boldsymbol{\gamma}^{(0)}}\left(\Phi_n\right)\exp(nh/2) \leq \exp(-nh/2).$$

Therefore $\sum_{n=1}^{\infty} P_{\boldsymbol{\gamma}^{(0)}}\left(\Phi_n > \exp(-nh/2)\right) < \infty$.
Applying Borel-Cantelli lemma $P_{\boldsymbol{\gamma}^{(0)}}\left(\Phi_n > \exp(-nh/2)\text{ i.o.}\right) = 0$. Thus,

$$\Phi_n \to 0 \quad a.s. \tag{17}$$

(b) Step 2
We have

$$\begin{aligned} E_{\boldsymbol{\gamma}^{(0)}}((1-\Phi_n)\mathcal{N}_n) &= \int (1-\Phi_n)\int_{\mathcal{A}_n} \frac{p_{\boldsymbol{\gamma}}(\boldsymbol{y}_n)}{p_{\boldsymbol{\gamma}^{(0)}}(\boldsymbol{y}_n)}\pi_n(\boldsymbol{\gamma})p_{\boldsymbol{\gamma}^{(0)}}(\boldsymbol{y}_n) \\ &= \int_{\mathcal{A}_n}\int (1-\Phi_n)p_{\boldsymbol{\gamma}}(\boldsymbol{y}_n)\pi_n(\boldsymbol{\gamma}) \\ &\leq \sup_{\boldsymbol{\gamma}\in\mathcal{A}_n} E_{\boldsymbol{\gamma}}(1-\Phi_n) \leq \exp(-nh). \end{aligned}$$

Applying Borel-Cantelli lemma, $P_{\boldsymbol{\gamma}^{(0)}}\left((1-\Phi_n)\mathcal{N}_n\exp(nh/2) > \exp(-nh/4)\text{ i.o.}\right) = 0$ so

$$\exp(nh/2)(1-\Phi_n)\mathcal{N}_n \to 0 \quad a.s.. \tag{18}$$

(c) Step 3

$$\begin{aligned}\int \frac{p_{\boldsymbol{\gamma}}(\boldsymbol{y}_n)}{p_{\boldsymbol{\gamma}^{(0)}}(\boldsymbol{y}_n)}\pi(\boldsymbol{\gamma}) &= \int \exp\left(\nabla w_{\boldsymbol{\gamma}^{(0)},n}(\boldsymbol{y}_n)'(\boldsymbol{\gamma}-\boldsymbol{\gamma}^{(0)}) + C_{\boldsymbol{y}_n,n}(\boldsymbol{\gamma})\right)\pi(\boldsymbol{\gamma}) \\ &\geq \int \exp\left(-||\nabla w_{\boldsymbol{\gamma}^{(0)},n}(\boldsymbol{y}_n)||_\infty ||\boldsymbol{\gamma}-\boldsymbol{\gamma}^{(0)}||_2 - \frac{n}{8}||\boldsymbol{\gamma}-\boldsymbol{\gamma}^{(0)}||_2^2\right)\pi(\boldsymbol{\gamma}) \\ &\geq \int \exp\left(-2\sqrt{nq_n}||\boldsymbol{\gamma}-\boldsymbol{\gamma}^{(0)}||_2 - \frac{n}{8}||\boldsymbol{\gamma}-\boldsymbol{\gamma}^{(0)}||_2^2\right)\pi(\boldsymbol{\gamma}) \\ &\geq \exp\left(-2\sqrt{nq_n}\frac{\eta}{n^{\rho/2}} - \frac{n\eta^2}{8n^\rho}\right)\Pi(||\boldsymbol{\gamma}-\boldsymbol{\gamma}^{(0)}||_2 < \frac{\eta}{n^{\rho/2}}),\end{aligned}$$

where the inequality in the second line follows from the Taylor series expansion after taking into account that $\nabla^2 z(\cdot) \leq 1/4$. The inequality in the third line follows from the fact that $\boldsymbol{y}_n \in \mathcal{E}_n$.

Observe that

$$\Pi(||\boldsymbol{\gamma}-\boldsymbol{\gamma}^{(0)}||_2 < \frac{\eta}{n^{\rho/2}}) \geq \Pi(||\boldsymbol{\gamma}_2-\boldsymbol{\gamma}_2^{(0)}||_2 < \frac{\eta}{2n^{\rho/2}})\Pi(||\boldsymbol{W}-\boldsymbol{W}^{(0)}||_2 < \frac{\eta}{2n^{\rho/2}}),$$

where $\boldsymbol{W}$ and $\boldsymbol{W}^{(0)}$ are as defined in Lemma 7.3. We will show sequentially (i) $-\log\Pi(||\boldsymbol{W}-\boldsymbol{W}^{(0)}||_2 < \frac{\eta}{2n^{\rho/2}}) = o(n)$ and (ii) $-\log\left\{\Pi(||\boldsymbol{\gamma}_2-\boldsymbol{\gamma}_2^{(0)}||_2 < \frac{\eta}{2n^{\rho/2}})\right\} = o(n)$.

(i) Note that,

$$\begin{aligned}\Pi(||\boldsymbol{W}-\boldsymbol{W}^{(0)}||_2 < \frac{\eta}{2n^{\rho/2}}) &\geq \Pi(||\boldsymbol{u}_k-\boldsymbol{u}_k^{(0)}||_2 \leq \upsilon_n,\ \forall\, k=1,..,V_n) \\ &\geq \mathrm{E}\left[\Pi(||\boldsymbol{u}_k-\boldsymbol{u}_k^{(0)}||_2 \leq \upsilon_n,\ \forall\, k=1,..,V_n|\Delta)\right] \\ &\geq \mathrm{E}\left[\prod_{k=1}^{V_n}\left\{\exp\left(-\frac{1}{2}\boldsymbol{u}_k^{(0)'}\boldsymbol{u}_k^{(0)}\right)\Pi(||\boldsymbol{u}_k||_2 \leq \upsilon_n|\Delta)\right\}\right], \qquad (19)\end{aligned}$$

where the first inequality follows from Lemma 7.3 by considering η as $\frac{\eta}{2n^{\rho/2}}$ with a slight abuse of notation, and υ_n is defined accordingly. The last inequality follows from Anderson Lemma. We will now make use of the fact that $\int_{-a}^{a}\exp(-x^2/2)dx \geq \exp(-a^2)2a$ to conclude

$$\begin{aligned}\Pi(||\boldsymbol{u}_k||_2 \leq \upsilon_n|\Delta) &\geq \prod_{r=1}^{R}\Pi(|u_{kr}| \leq \frac{\upsilon_n}{R}|\Delta) = \prod_{r=1}^{R}\left((1-\Delta) + \frac{\Delta}{\sqrt{2\pi}}\int_{-\upsilon_n/R}^{\upsilon_n/R}\exp(-x^2/2)\right) \\ &\geq \prod_{r=1}^{R}\left((1-\Delta) + \frac{\Delta}{\sqrt{2\pi}}\exp(-\upsilon_n^2/R^2)\frac{2\upsilon_n}{R}\right) \geq \left[(1-\Delta) + \frac{\Delta}{\sqrt{2\pi}}\exp(-\upsilon_n^2/R^2)\frac{2\upsilon_n}{R}\right]^R.\end{aligned}$$

$$\prod_{k=1}^{V_n}\Pi(||\boldsymbol{u}_k||_2 \leq \upsilon_n) \geq \mathrm{E}\left[(1-\Delta)+\frac{\Delta}{\sqrt{2\pi}}\exp(-\upsilon_n^2/R^2)\frac{2\upsilon_n}{R}\right]^{RV_n}$$
$$= \mathrm{E}\left[\sum_{h_1=1}^{RV_n}\binom{RV_n}{h_1}(1-\Delta)^{h_1}\Delta^{RV_n-h_1}\left(\frac{2\upsilon_n}{R}\right)^{RV_n-h_1}\exp(-(RV_n-h_1)\upsilon_n^2/R^2)\right]$$
$$\geq \sum_{h_1=1}^{RV_n}\binom{RV_n}{h_1}Beta(RV_n-h_1+1,h_1+1)\left(\frac{2\upsilon_n}{R}\right)^{RV_n-h_1}\exp(-(RV_n-h_1)\upsilon_n^2/R^2)$$
$$\geq \sum_{h_1=1}^{RV_n}\frac{(RV_n)!}{h_1!(RV_n-h_1)!}\frac{h_1!(RV_n-h_1)!}{(RV_n+1)!}\left(\frac{2\upsilon_n}{R}\right)^{RV_n-h_1}\exp(-(RV_n-h_1)\upsilon_n^2/R^2)$$
$$\geq \frac{RV_n}{RV_n+1}\left(\frac{2\upsilon_n}{R}\right)^{RV_n}\exp(-V_n\upsilon_n^2/R).$$

Where the last inequality follows from Lemma 7.3 by considering the fact that, $\upsilon_n = \min_{k,l}\frac{-[||\boldsymbol{u}_k^{(0)}||+||\boldsymbol{u}_l^{(0)}||]+\sqrt{[||\boldsymbol{u}_k^{(0)}||+||\boldsymbol{u}_l^{(0)}||]^2+2\eta/n^{\rho/2}}}{2} \leq \frac{\sqrt{\eta}}{\sqrt{2}n^{\rho/4}}$. Hence, $0 < \frac{2\upsilon_n}{R} < 1$ for large n. It now follows from (19) that

$$-\log\Pi(||\boldsymbol{W}-\boldsymbol{W}^{(0)}||_2 < \frac{\eta}{2n^{\rho/2}}) \leq \sum_{k=1}^{V_n}\frac{\boldsymbol{u}_k^{(0)\prime}\boldsymbol{u}_k^{(0)}}{2}+\frac{V_n\eta}{2Rn^{\rho/2}}-(RV_n)\log\left(\frac{2\sqrt{\eta}}{\sqrt{2}Rn^{\rho/4}}\right)+\log(RV_n+1)$$
$$-\log(RV_n) = o(n),$$

by the assumptions (A) and (B). This proves (i).

We will now prove (ii). It follows that

$$\Pi(||\boldsymbol{\gamma}_2-\boldsymbol{\gamma}_2^{(0)}||_2 < \frac{\eta}{2n^{\rho/2}}) \geq \Pi(|\gamma_{2j}-\gamma_{2j}^{(0)}| < \frac{\eta}{2\sqrt{q_n}n^{\rho/2}}, j\in\mathcal{S}^0)\Pi(\sum_{j\notin\mathcal{S}^0}|\gamma_{2j}|^2 < \frac{(q_n-s_{2,n}^0)\eta^2}{4q_nn^\rho}). \tag{20}$$

We will lower bound two components of the product in (20) individually. By Chebyshev's inequality

$$\Pi(\sum_{j\notin\mathcal{S}^0}|\gamma_{2j}|^2 < \frac{(q_n-s_{2,n}^0)\eta^2}{4q_nn^\rho}) \geq \left(1-\frac{E[\sum_{j\notin\mathcal{S}^0}|\gamma_{2j}|^2]4q_nn^\rho}{(q_n-s_{2,n}^0)\eta^2}\right)$$
$$= \left(1-\frac{2\lambda_nq_nn^\rho}{\eta^2}\right). \tag{21}$$

$$\Pi(|\gamma_{2j}-\gamma_{2j}^{(0)}|<\frac{\eta}{2\sqrt{q_n}n^{\rho/2}},j\in\mathcal{S}^0)=E\left[\Pi(|\gamma_{2j}-\gamma_{2j}^{(0)}|<\frac{\eta}{2\sqrt{q_n}n^{\rho/2}},j\in\mathcal{S}^0|\boldsymbol{\phi}_{\mathcal{S}^0})\right]$$
$$=E\left[\prod_{j\in\mathcal{S}^0}\Pi(|\gamma_{2j}-\gamma_{2j}^{(0)}|<\frac{\eta}{2\sqrt{q_n}n^{\rho/2}}|\boldsymbol{\phi}_{\mathcal{S}^0})\right].$$

Using the fact that $\int_a^b e^{-x^2/2}dx\geq e^{-(a^2+b^2)/2}(b-a)$, one obtains

$$\prod_{j\in\mathcal{S}^0}\Pi(|\gamma_{2j}-\gamma_{2j}^{(0)}|<\frac{\eta}{2\sqrt{q_n}n^{\rho/2}}|\boldsymbol{\phi}_{\mathcal{S}^0})\geq\prod_{j\in\mathcal{S}^0}\left\{\left(\frac{\eta}{\sqrt{2q_nn^\rho\pi\phi_j}}\right)\exp\left(-\frac{|\gamma_{2j}^0|^2+\eta^2/(4q_nn^\rho)}{\phi_j}\right)\right\}.$$

Thus

$$\Pi(|\gamma_{2j}-\gamma_{2j}^{(0)}|<\frac{\eta}{2\sqrt{q_n}n^{\rho/2}},j\in\mathcal{S}^0)\geq E\left[\prod_{j\in\mathcal{S}^0}\left\{\left(\frac{\eta}{\sqrt{2q_nn^\rho\pi\phi_j}}\right)\exp\left(-\frac{|\gamma_{2j}^0|^2+\eta^2/(4q_nn^\rho)}{\phi_j}\right)\right\}\right]$$
$$\geq\left(\frac{\eta\lambda_n}{\sqrt{2q_nn^\rho\pi}}\right)^{s_{2,n}^0}\prod_{j\in\mathcal{S}^0}\int_{\phi_j}\left\{\frac{1}{\sqrt{\phi_j}}\exp\left(-\frac{|\gamma_{2j}^0|^2+\eta^2/(4q_nn^\rho)}{\phi_j}-\frac{\lambda_n\phi_j}{2}\right)d\phi_j\right\}.$$

Use the change of variable $\frac{1}{\phi_j}=z_j$ and the normalizing constant from the inverse Gaussian density to deduce

$$\int_{\phi_j}\left\{\frac{1}{\sqrt{\phi_j}}\exp\left(-\frac{|\gamma_{2j}^0|^2+\eta^2/(4q_nn^\rho)}{\phi_j}-\frac{\lambda_n\phi_j}{2}\right)d\phi_j\right\}$$
$$=\int_{z_j}\left\{\frac{1}{\sqrt{z_j^3}}\exp\left(-(|\gamma_{2j}^0|^2+\eta^2/(4q_nn^\rho)z_j-\frac{\lambda_n}{2z_j}\right)dz_j\right\}=\sqrt{\left(\frac{2\pi}{\lambda_n}\right)}\exp\left(-\lambda_n\sqrt{2\left(|\gamma_{2j}^0|^2+\eta^2/(4q_nn^\rho)\right)}\right).$$

Therefore,

$$\Pi(|\gamma_{2j}-\gamma_{2j}^{(0)}|<\frac{\eta}{2\sqrt{q_n}n^{\rho/2}},j\in\mathcal{S}^0)\geq\left(\frac{\eta\sqrt{\lambda_n}}{\sqrt{q_nn^\rho}}\right)^{s_{2,n}^0}\exp\left(-\lambda_n\sum_{j\in\mathcal{S}^0}\sqrt{2\left(|\gamma_{2j}^0|^2+\eta^2/(4q_nn^\rho)\right)}\right).\tag{22}$$

Combining results from (21) and (22)

$$\Pi(||\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_2^{(0)}||_2 < \frac{\eta}{2n^{\rho/2}}) \geq \left(\frac{\eta\sqrt{\lambda_n}}{\sqrt{q_n n^{\rho}}}\right)^{s^0_{2,n}} \exp\left(-\lambda_n \sum_{j\in\mathcal{S}^0} \sqrt{2\left(|\gamma^0_{2j}|^2 + \eta^2/(4q_n n^{\rho})\right)}\right) \left(1 - \frac{2\lambda_n q_n n^{\rho/2}}{\eta^2}\right).$$

Using the fact that $\lambda_n = \frac{1}{q_n n^{\rho/2}\log(n)}$,

$$-\log\Pi(||\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_2^{(0)}||_2 < \frac{\eta}{2n^{\rho/2}}) \leq s^0_{2,n}[\eta + \log(q_n) + (3\rho/4)\log(n) + \log(\log(n))/2] + \frac{\sqrt{2\left(|\gamma^0_{2j}|^2 + \eta^2/(4q_n n^{\rho})\right)}}{q_n n^{\rho/2}\log(n)} - \log\left(1 - \frac{2}{\eta^2\log(n)}\right) = o(n), \tag{23}$$

under assumptions (B)-(F).

Finally,

$$\begin{aligned} -\log(\mathcal{D}_n) &\leq 2\sqrt{nq_n}\frac{\eta}{n^{\rho/2}} + \frac{n\eta^2}{8n^{\rho}} - \log\Pi(||\boldsymbol{\gamma} - \boldsymbol{\gamma}^{(0)}||_2 < \frac{\eta}{n^{\rho/2}}) \\ &= 2\eta\sqrt{q_n}n^{(1-\rho)/2} + \frac{\eta^2}{8}n^{1-\rho} - \log\Pi(||\boldsymbol{\gamma} - \boldsymbol{\gamma}^{(0)}||_2 < \frac{\eta}{n^{\rho/2}}). \end{aligned}$$

Using (23), the fact that $(1-\rho)/2 \in (-1/2, 1/2)$ and assumption (B), we obtain $-\log(\mathcal{D}_n) = o(n)$. Thus (c) follows.

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.

Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, **58**(3), 11.

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

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

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, **21**(2), 572–596.

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.

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.

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.

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.

Durante, D. and Dunson, D. B. (2017). Bayesian inference and testing of group differences in brain networks. *Bayesian Analysis*, **doi:10.1214/16-BA1030**. Advance publication.

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.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75**(4), 603–680.

Fazel, M., Hindi, H., and Boyd, S. P. (2003). Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices. In *American Control Conference, 2003. Proceedings of the 2003*, volume 3, pages 2156–2162. IEEE.

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*, pages 643–652. ACM.

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.

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.

Gelman, A., Hwang, J., and Vehtari, A. (2014b). Understanding predictive information criteria for bayesian models. *Statistics and computing*, **24**(6), 997–1016.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, **88**(423), 881–889.

Guha, S. and Rodriguez, A. (2018). Bayesian regression with undirected network predictors with an application to brain connectome data. *arXiv preprint arXiv:1803.10655*.

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

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

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). Multiplicative latent factor models for description and prediction of social networks. *Computational and mathematical organization theory*, **15**(4), 261.

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.

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.

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

Luo, X. (2011). High dimensional low rank and sparse covariance matrix estimation via convex minimization. *Arxiv preprint*.

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.

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.

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.

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.

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

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.

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.

Wei, R. and Ghosal, S. (2017). Contraction properties of shrinkage priors in logistic regression. *Preprint at http://www4. stat. ncsu. edu/˜ ghoshal/papers*.

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.

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.