Published in final edited form as:

Biometrics. 2017 September ; 73(3): 769–779. doi:10.1111/biom.12650.

On Joint Estimation of Gaussian Graphical Models for Spatial and Temporal Data

Zhixiang Lin[1,2], **Tao Wang**[3], **Can Yang**[4], and **Hongyu Zhao**[5,*]

[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, U.S.A

[2]Department of Statistics, Stanford University, Stanford, California, U.S.A

[3]Department of Bioinformatics and Biostatistics, Shanghai Jiao Tong University, Shanghai, China

[4]Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong

[5]Department of Biostatistics, School of Public Health, Yale University, New Haven, Connecticut, U.S.A

Summary

In this paper, we first propose a Bayesian neighborhood selection method to estimate Gaussian Graphical Models (GGMs). We show the graph selection consistency of this method in the sense that the posterior probability of the true model converges to one. When there are multiple groups of data available, instead of estimating the networks independently for each group, joint estimation of the networks may utilize the shared information among groups and lead to improved estimation for each individual network. Our method is extended to jointly estimate GGMs in multiple groups of data with complex structures, including spatial data, temporal data and data with both spatial and temporal structures. Markov random field (MRF) models are used to efficiently incorporate the complex data structures. We develop and implement an efficient algorithm for statistical inference that enables parallel computing. Simulation studies suggest that our approach achieves better accuracy in network estimation compared with methods not incorporating spatial and temporal dependencies when there are shared structures among the networks, and that it performs comparably well otherwise. Finally, we illustrate our method using the human brain gene expression microarray dataset, where the expression levels of genes are measured in different brain regions across multiple time periods.

Keywords

Bayesian Variable Selection; Gaussian Graphical Model; Neighborhood Selection; Markov Random Field; Spatial and Temporal Data

[*] hongyu.zhao@yale.edu.

Supplementary Materials

Web Appendices and Figures, referenced in Section 2–5, are available with this paper at the Biometrics website on Wiley Online Library. The Matlab code is available at https://github.com/linzx06/Spatial-and-Temporal-GGM.

1. Introduction

The analysis of biological networks, including protein-protein interaction networks (PPI), biological pathways, transcriptional regulatory networks and gene co-expression networks, has led to numerous advances in the understanding of the organization and functionality of biological systems (e.g., Kanehisa and Goto 2000; Shen-Orr et al. 2002; Rual et al. 2005; Zhang and Horvath 2005). The work presented in this paper was motivated from the analysis of the human brain gene expression microarray data, where the expression levels of genes were measured in numerous spatial loci, which represent different brain regions, during different time periods of brain development (Kang et al., 2011). Although these data offer rich information on the network information among genes, only naive methods have been used for network inference. For example, Kang et al. (2011) pooled all the data from different spatial regions and time periods to construct a single gene network. However, only a limited number of data points are available for a specific region and time period, making region- and time- specific inference challenging.

Our aim here is to develop sound statistical methods to characterize the changes in the networks across time periods and regions, as well as the common network edges that are shared. This is achieved through a joint modeling framework to infer individual graphs for each region in each time period, where the degrees of spatial and temporal similarity are learnt adaptively from the data. Our proposed joint modeling framework may better capture the edges that are shared among graphs, and also allow the graphs to differ across regions and time periods.

We represent the biological network with a graph $G = (V, E)$ consisting of vertices $V = \{1, \ldots, p\}$ and edges $E \subset V \times V$. In this paper, we focus on conditionally independent graphs, where $(i, j) \in E$ if and only if node i and node j are not conditionally independent given all the other nodes. Gaussian graphical models (GGMs) have been proven among the best to infer conditionally independent graphs. In GGM, the p-dimensional $X = (X_1, \ldots, X_p)$ is assumed to follow a multivariate Gaussian distribution $\mathscr{N}\left(\mu, \sum\right)$. Denote $\Theta = \Sigma^{-1}$ the precision matrix. It can be shown that the conditional independence of X_i and X_j is equivalent to $\Theta_{ij} \neq 0$: $X_i \perp\!\!\!\perp X_j / X_{V \setminus \{i,j\}} \Leftrightarrow \Theta_{ij} = 0$. In GGM, estimating the conditional independence graph is equivalent to estimating the non-zero entries in Θ. Various approaches have been proposed to estimate the graph (Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Friedman et al., 2008; Cai et al., 2011; Dobra et al., 2011; Wang et al., 2012; Orchard et al., 2013). Among these methods, Friedman et al. (2008) developed a fast and simple algorithm, named the graphical lasso (glasso), using a coordinate descent procedure for the lasso. They considered optimizing the penalized likelihood, with ℓ_1 penalty on the precision matrix. As extensions of glasso, several approaches have been proposed to jointly estimate GGMs in multiple groups of data. Guo et al. (2011) expressed the elements of the precision matrix for each group as a product of binary common factors and group-specific values. They incorporated an ℓ_1 penalty on the common factors, to encourage shared sparse structure, and another ℓ_1 penalty on the group-specific values, to allow edges included in the shared structure to be set to zero for specific groups. Danaher et al. (2014) extended glasso more directly by extending the ℓ_1 penalty for each precision matrix with additional penalty functions that encourage shared structure. They proposed two possible choices of penalty functions: 1. Fused lasso penalty that penalizes the difference of the precision matrices, which encourages common values among the precision matrices; 2. Group lasso penalty. Chun et al. (2014) proposed a class of non-convex penalties for more flexible joint sparsity constraints. As an alternative to the penalized methods, Peterson et al. (2014) proposed a Bayesian approach. They formulated the model in the *G*-Wishart prior framework and modeled the similarity of multiple graphs through a Markov Random Field (MRF) prior. However, their approach is only applicable when the graph size is small (~ 20) and the number of groups is also small (~ 5).

In this paper, we formulate the model in a Bayesian variable selection framework to estimate the graph structure (George and McCulloch, 1993, 1997). The precision matrix is estimated in a second step with the graph structure fixed (Hastie et al., 2009). Meinshausen and Bühlmann (2006) proposed a neighborhood selection procedure for estimating GGMs, where the neighborhood of node i was selected by regressing on all the other nodes. Intuitively, our approach is the Bayesian analog of the neighborhood selection procedure. Our framework is applicable to the estimation of both single graph and multiple graphs. For the joint estimation of multiple graphs, we incorporate the MRF model. Compared with Peterson et al. (2014), we use a different MRF model and a different inferential procedure. In small scale simulations (20 nodes), our method performed slightly worse than Peterson et al. (2014), but better than the other competing methods (Friedman et al., 2008; Guo et al., 2011; Danaher et al., 2014). One advantage of our approach is that it can naturally model complex data structures, such as spatial data, temporal data and data with both spatial and temporal structures. Another advantage is the computational efficiency. For the estimation of a single graph with 100 nodes (the typical size of biological pathways is around that range), the computational time on a laptop is ~ 30 seconds for 1,000 iterations of Gibbs sampling, which is ~ 3-folds faster than Bayesian Graphical Lasso, which implements a highly efficient block Gibbs sampler and is among the fastest algorithms for estimating GGMs in the Bayesian framework (Wang et al., 2012). For multiple graphs, the computational time increases roughly linear with the number of graphs. Our procedure also enables parallel computing and the computational time can be substantially reduced if multicore processors are available. For single graph estimation, we show the graph selection consistency of the proposed method in the sense that the posterior probability of the true model converges to one.

The rest of the paper is organized as follows. We introduce the Bayesian neighborhood selection procedure for single graph and the extension to multiple graphs in Section 2. The theoretical properties are presented in Section 3. The simulation results are demonstrated in Section 4 and the application to the human brain gene expression microarray dataset is presented in Section 5. We conclude the paper with a brief summary in Section 6.

2. Statistical Model and Methods

2.1 The Bayesian Neighborhood Selection Procedure

We first consider estimating the graph structure when there is only one group of data. Consider the *p*-dimensional multivariate normal random variable $X \sim \mathscr{N}\left(\mu, \sum\right)$. We further assume that X is centered and $X \sim \mathscr{N}\left(0, \sum\right)$. Let $\Theta = \Sigma^{-1}$ denote the precision matrix. Let the $n \times p$ matrix $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ contain n independent observations of X. For $A \subseteq \{1, \ldots, p\}$, define $\mathbf{X}_A = (\mathbf{X}_j, j \in A)$. Let Γ_i denote the subset of $\{1, \ldots, p\}$, excluding the *i*th entry only. For any square matrix C, let $C_{i\Gamma_i}$ denote the *i*th row, excluding the *i*th element in that row. Consider estimating the neighborhood of node *i*. It is well known that the following conditional distribution holds:

$$\mathbf{X}_i|\mathbf{X}_{\Gamma_i} \sim \mathscr{N}\left(-\mathbf{X}_{\Gamma_i}\Theta_{i\Gamma_i}^T\Theta_{ii}^{-1}, \Theta_{ii}^{-1}\boldsymbol{I}\right), \quad (1)$$

where $\boldsymbol{I}$ is the $n \times n$ identity matrix, Θ_{ii} is a scalar and finding the neighborhood of X_i is equivalent to estimating the non-zero coefficients in the regression of X_i on X_{Γ_i}. Let $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ be matrices of dimension $p \times p$, where $\beta_{i\Gamma_i} = -\Theta_{ii}^{-1}\Theta_{i\Gamma_i}$ and $\boldsymbol{\gamma}$ is the binary latent state matrix. The diagonal elements in $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are not assigned values. Conditioning on γ_{ij}, β_{ij} is assumed to follow a normal mixture distribution (George and McCulloch, 1993, 1997):

$$\beta_{ij}|\gamma_{ij} \sim (1-\gamma_{ij})\mathscr{N}(0, \tau_{i0}^2) + \gamma_{ij}\mathscr{N}(0, \tau_{i1}^2), \text{ for } j \in \Gamma_i,$$

where $\tau_{i0}/\tau_{i1} = \delta$ and $0 < \delta < 1$. The prior on γ_{ij} is Bernoulli: $p(\gamma_{ij} = 1) = 1 - p(\gamma_{ij} = 0) = q$. δ, τ_{i1} and q are prefixed hyperparameters and are discussed in the Supplementary Materials. The off-diagonal entries in $\boldsymbol{\gamma}$ represent the presence or absence of the edges, which is the goal of our inference.

Let $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_p)$, where $\sigma_i^2 = \Theta_{ii}^{-1}$. The inverse gamma (IG) conjugate prior is assumed for σ_i^2:

$$\sigma_i^2|\gamma \sim IG\left(\nu_i/2, \lambda\nu_i/2\right).$$

In this paper, we assume that $\nu_i = 0$ and the IG prior reduces to a flat prior (Li and Zhang, 2010).

For each node, we implement the Bayesian procedure to select the neighbors of that node. The graph selection consistency of our approach is shown in Section 3. Performing neighborhood selection is not equivalent to making inference on Θ (Dobra et al., 2004). Though $f(\mathbf{X}_i|\mathbf{X}_{\Gamma_i})$ is a conditional likelihood for Gaussian distribution, the function $\prod_i f(\mathbf{X}_i|\mathbf{X}_{\Gamma_i})$ is not a Gaussian likelihood and it has the likelihood properties of a misspecified model with p regressions (Varin et al., 2011). We implement MCMC to estimate the posterior distributions and details for the algorithm are provided in the Supplementary Materials. In our approach, the symmetric constraint of the graph structure can be incorporated when sampling γ by assuming $\gamma_{ij} = \gamma_{ji}$ for $j \neq i$. Under the Gaussian assumption, the regression coefficients β are directly related to the corresponding entries in Θ and are also related to the partial correlation. The symmetry of Θ can be satisfied if we impose the constraint when sampling β. Without the constraint on β, the partial correlation may not be estimated coherently: the magnitudes and signs may be different between nodes i on j and nodes j on i. However, imposing the constraint will lead to substantial loss in computational efficiency since β have to be updated one at a time, instead of one row at a time. Moreover, our simulation results suggest that the two approaches are comparable for graph structure estimation, whether or not the constraint is imposed (data not shown). We do not impose the constraint on β in practice. Although most applications tend to focus more on the graph structure estimation, it may be desirable to obtain a symmetric positive definite estimate for Θ. To achieve this, we propose a two-step approach. First, we estimate the graph structure with $\hat{G}$ following the neighborhood selection approach. Second, we estimate Θ with the mode of the conditional likelihood $p(X / \hat{G}, \Theta)$, where Θ is subjected to the constraint of $\hat{G}$. The mode can be obtained with a fast iterative algorithm presented in Hastie et al. (2009). There are two limitations in the two-step approach. First, uncertainty of the graph structure in step 1 is not taken into account in step 2. Second, priors on the regression coefficients are part of the modeling framework in step 1, but not in step 2. Examples implementing the two-step procedure are provided in the Supplementary Materials.

2.2 Extension to mutiple graphs

When there is similarity shared among multiple graphs, jointly estimating multiple graphs can improve inference. We propose to jointly estimate multiple graphs by specifying a Markov Random Field (MRF) prior on the latent states. Our model can naturally incorporate complex data structures, such as spatial data, temporal data and data with both spatial and temporal structures. Consider jointly estimating multiple graphs for data with both spatial and temporal structures. Denote B the set of spatial loci and T the set of time points. Our proposed model can be naturally implemented when there is missing data, i.e. no data points taken in certain locus at certain time point. For now, we assume that there is no missing data. The latent states for the whole dataset are represented by a $/B/\times/T/\times p \times p$ array γ, where $|\,|$ denotes the cardinality of a set. Let $\gamma_{bt\cdot\cdot}$ denote the latent state matrix for locus b at time t. In the real data example, b is a categorical variable representing the brain region and t is a discrete variable that represents the time period during brain development. Same as that in Section 2.1, the diagonal entries in $\gamma_{bt\cdot\cdot}$ are not assigned values.

Consider estimating the neighborhood of node i. Let γ_{btij} denote the latent state for node $j \in \Gamma_i$ in locus b at time t. Denote $\gamma_{\cdot\cdot ij} = \{\gamma_{btij} : \forall b \in B, \forall t \in T\}$, $E^s_{ij}=\{(\gamma_{ijbt}, \gamma_{ijb't'}):b \neq b', t=t'\}$ and $E^t_{ij}=\{(\gamma_{btij}, \gamma_{b't'ij}):b=b' \text{ and } |t-t'|=1\}$. Here E^s_{ij} contain all the pairs capturing spatial similarity and E^t_{ij} contain all the pairs capturing temporal dependency between adjacent time periods. We do not consider the direction of the pairs: $(\gamma_{ijbt}, \gamma_{ijb't'})$ and $(\gamma_{ijb't'}, \gamma_{ijbt})$ are the same. Let $I_1(\cdot)$ and $I_0(\cdot)$ represent the indicator functions of 1 and 0, respectively. The prior for $\boldsymbol{\gamma}_{\cdot\cdot ij}$ is specified by a pairwise interaction MRF model (Besag, 1986; Lin et al., 2015):

$$p(\gamma_{\cdot\cdot ij}|\mathbf{\Phi}) \propto \exp\left\{\eta_1 \sum_{b\in B,t\in T} I_1(\gamma_{ijbt}) + \right. \\ \eta_s \sum_{E_{ij}^s} \left[I_0(\gamma_{btij}) I_0(\gamma_{b't'ij}) + I_1(\gamma_{btij}) I_1(\gamma_{b't'ij})\right] + \\ \eta_t \sum_{E_{ij}^t} \left[I_0(\gamma_{btij}) I_0(\gamma_{b't'ij}) + I_1(\gamma_{btij}) I_1(\gamma_{b't'ij})\right], \tag{2}$$

and conditional independence is assumed:

$$p\left(\gamma|\mathbf{\Phi}\right) = \prod_i \prod_{j\in\Gamma_i} p(\gamma_{\cdot\cdot ij}|\mathbf{\Phi}), \tag{3}$$

where $\mathbf{\Phi} = \{\eta_1, \eta_s, \eta_t\}$ are set to be the same for all i and j. $\eta_1 \in \mathbb{R}$ and when there is no interaction terms, $1/(1 + \exp(-\eta_1))$ corresponds to q in the Bernoulli prior. $\eta_s \in \mathbb{R}$ represents the magnitude of spatial similarity and $\eta_t \in \mathbb{R}$ represents the magnitude of temporal similarity. In the simulation and real data example, η_1 is prefixed, whereas η_s and η_t are estimated from the dataset. The priors on η_s and η_t are assumed to follow uniform distribution in [0, 2]. Sensitivity analyses on the choice of η_1 and the priors on η_s and η_t are provided in the Supplementary Materials.

Let $\boldsymbol{\gamma}_{\cdot\cdot ij}/\boldsymbol{\gamma}_{btij}$ denote the subset of $\boldsymbol{\gamma}_{\cdot\cdot ij}$ excluding $\boldsymbol{\gamma}_{btij}$. Then we have:

$$p(\gamma_{btij}|\gamma_{\cdot\cdot ij}/\gamma_{btij}, \mathbf{\Phi}) = \frac{\exp\{\gamma_{btij} F(\gamma_{btij}, \mathbf{\Phi})\}}{1+\exp\{F(\gamma_{btij}, \mathbf{\Phi})\}}, \tag{4}$$

where

$$F(\gamma_{btij}, \mathbf{\Phi}) = \eta_1 + \eta_s \sum_{b'\in B, b'\neq b} (2\gamma_{b'tij} - 1) + \eta t \left\{I_{t\neq 1}\left[2\gamma_{b(t-1)ij} - 1\right] + I_{t\neq T}\left[2\gamma_{b(t+1)ij} - 1\right]\right\}.$$

In the MCMC for multiple graphs, the Metropolis-Hastings (MH) algorithm is implemented to update η_s and η_t. The normalizing constant in $p(\boldsymbol{\gamma}_{\cdot\cdot ij} | \mathbf{\Phi})$ is generally not tractable as one has to sum over all $2^{|B|+|T|}$ possible configurations of $\boldsymbol{\gamma}_{\cdot\cdot ij}$. The likelihood ratio need in the MH step is approximated with the ratio of pseudolikelihoods (Besag, 1986) and the pseudolikelihood is calculated as: $\prod_{b\in B} \prod_{t\in T} p(\boldsymbol{\gamma}_{btij} / \boldsymbol{\gamma}_{\cdot\cdot ij}/\boldsymbol{\gamma}_{btij}, \mathbf{\Phi})$. Comparison between using the ratio of pseudolikelihoods with the bridge sampler (Meng and Wong, 1996) is shown in the Supplementary Materials.

In the prior specification (2), we made the following assumptions: a) the spatial similarity does not change over time and b) the time periods are evenly spaced and can be represented by integer labels. The first assumption can be relaxed by allowing η_s to change over time. For the second assumption, η_t can be adjusted to a parametric function of the time interval. When there is only spatial or only temporal structure in the dataset, prior (2) can be adjusted by removing the summation over the corresponding pairs. The posterior probability-based false discovery rate (FDR) control (Newton et al., 2004) can be implemented to evaluate the marginal posterior probabilities (Supplementary Materials).

3. Theoretical Properties

We rewrite p as p_n to represent a sequence p_n that changes with n. Let $1 \le p^* \le p_n$. Throughout, we assume that **X** satisfies the sparse Riesz condition (Zhang and Huang, 2008) with rank p^*; that is, there exist some constants $0 < c_1 < c_2 < \infty$ such that

$$c_1 \le \frac{\|\mathbf{X}_A u\|^2}{n\|u\|^2} \le c_2,$$

for any $A \subseteq \{1, \ldots, p_n\}$ with size $/A/ = p^*$ and any nonzero vector $u \in {}^{p^*}$.

Consider estimating the neighborhood for the ith node. We borrow some notations from Narisetty et al. (2014). For the simplicity of notation, let $\beta^i \equiv \beta_{i\Gamma_i} = -\Theta_{i\Gamma_i}^T \Theta_{ii}^{-1}$ and $\gamma^i \equiv \gamma_{i\Gamma_i}$. Write τ_{i0}, τ_{i1} and q as τ_{0n}, τ_{1n} and q_n, respectively, to represent sequences that change with n. We use a $(p_n - 1) \times 1$ binary vector k^i to index an arbitrary model. The corresponding design matrix and parameter vector are denoted by $\mathbf{X}_{k^i} \equiv (\mathbf{X}_{\Gamma_i})_{k^i}$ and $\boldsymbol{\beta}^i_{k^i}$, respectively. Let t^i represent the true neighborhood of node i.

Denote by $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ the largest and smallest eigenvalues of a matrix, respectively. For $\upsilon > 0$, define

$$m(\upsilon) \equiv m_n(\upsilon) = (p_n - 1) \wedge \frac{n}{(2+\upsilon)\log(p_n - 1)}$$

and

$$\lambda_{m,i}(\upsilon) = \min_{k^i : |k^i| \le m(\upsilon)} \lambda_{\min}\left(\frac{\mathbf{X}'_{k^i}\mathbf{X}_{k^i}}{n}\right).$$

For $K > 0$, let

$$\Delta_i(K) = \min_{\left\{k^i : |k^i| \le K|t^i|, k^i)\overline{\supset} t^i\right\}} \|(I - P_{k^i})\mathbf{X}_{t^i}\boldsymbol{\beta}^i_{t^i}\|_2^2,$$

where $/k^i/$ denotes the size of the model k^i and P_{k^i} is the projection matrix onto the column space of $\mathbf{X}_{k^i}$.

For sequences a_n and b_n, $a_n \sim b_n$ means $a_n/b_n \to c$ for some constant $c > 0$, $a_n \prec b_n$ (or $b_n \succ a_n$) means $a_n = o(b_n)$, and $a_n \preceq b_n$ (or $b_n \succeq a_n$) means $a_n = O(b_n)$. We need the following conditions.

- **(A)** $p_n \to \infty$ and $p_n = O(n^{\theta})$ for some $\theta > 0$;
- **(B)** $q_n = p_n^{\alpha-1}$ for some $0 \le \alpha < 1 \wedge (1/\theta)$;
- **(C)** $n\tau_{0n}^2 = o(1)$ and $n\tau_{1n}^2 \sim n \vee p_n^{2+2\delta_1}$ for some $\delta_1 > 1 + \alpha$;
- **(D)** $|t^i| \prec n/\log p_n$ and $\|\beta_{t^i}^i\|_2^2 \prec \tau_{1n}^2 \log p_n$;
- **(E)** there exist $1 + \alpha < \delta_2 < \delta_1$ and $K > 1 + 8/(\delta_2 - 1 - \alpha)$ such that, for some large $C > 0$, $\Delta_i(K)/\sigma_i^2 > C|t^i| \log\left(n \vee p_n^{2+2\delta_1}\right)$;
- **(F)** $p^* \geqslant (K + 1)|t^i|$;
- **(G)** $\lambda_{\max}(\mathbf{X}'\mathbf{X}/n) \prec \left(n\tau_{0n}^2\right)^{-1} \wedge \left(n\tau_{1n}^2\right)$ and there exist some $0 < \upsilon < \delta_2$ and $0 < \kappa < 2(K - 1)$ such that

$$\lambda_{m,i}(\upsilon) \succeq \frac{\left(n \vee p_n^{2+2\delta_2}\right)}{n\tau_{1n}^2} \vee p_n^{-\kappa}.$$

Theorem 1

Assume conditions (A)–(G). For some $c > 0$ and $s > 1$ we have, with probability at least $1 - cp_n^{-s}$, $P(\gamma^i = t^i | \mathbf{X}, \sigma_i^2) > 1 - r_n$, where r_n goes to 0 as the sample size increases to ∞.

To establish graph-selection consistency, we need slightly stronger conditions than (D)–(G). Let

$$t^* = \max_{1 \le i \le p_n} |t^i|, \Delta^*(K) = \min_{1 \le i \le p_n} \left(\Delta_i(K)/\sigma_i^2\right) \text{ and } \lambda_m^*(\upsilon) = \min_{1 \le i \le p_n} \lambda_{m,i}(\upsilon).$$

- **(D′)** $t^* \prec n/\log p_n$ and $\max_{1 \le i \le p_n} \|\beta_{t^i}^i\|_2^2 \prec \tau_{1n}^2 \log p_n$;
- **(E′)** there exist $1 + \alpha < \delta_2 < \delta_1$ and $K > 1 + 8/(\delta_2 - 1 - \alpha)$ such that, for some large $C > 0$, $\Delta^*(K) > C\log(n \vee p_n^{2+2\delta_1})$;
- **(F′)** $p^* \geqslant (K + 1)t^*$;
- **(G′)** $\lambda_{\max}(\mathbf{X}'\mathbf{X}/n) \prec (n\tau_{0n}^2)^{-1} \wedge (n\tau_{1n}^2)$ and there exist some $0 < \upsilon < \delta_2$ and $0 < \kappa < 2(K - 1)$ such that

$$\lambda_m^*(\upsilon) \succeq \frac{\left(n \vee p_n^{2+2\delta_2}\right)}{n\tau_{1n}^2} \vee p_n^{-\kappa}.$$

Let $\mathscr{G}$ denote the true graph structure and $\boldsymbol{\gamma}$ is the latent state matrix for all the nodes.

Theorem 2

Assume conditions (A)–(C) and (D′)–(G′). We have, as $n \to \infty$, $P(\gamma = \mathscr{G} | \mathbf{X}, \boldsymbol{\sigma}^2) \to 1$.

The proofs of Theorem 1 and 2 are provided in the Supplementary Materials. In the proof of Theorem 1, we borrowed the general framework and some ideas from Narisetty et al. (2014). The key difference in our proof is that (1) we need to simultaneously control the posterior probability for p regressions, while allowing p to diverge with the sample size; (2) we allow the true model size to diverge while Narisetty et al. (2014) assumed that it is fixed. Some prior specifications are also different. The maximum a posteriori (MAP) estimate is hard to achieve in practice as the searching space is too large: $2^{\text{number of possible edges}}$. Instead, we use the marginal posterior probability to select the edges. The consistency of joint posterior probability implies the consistency of marginal posterior probability.

4. Simulation examples

4.1 Joint estimation of multiple graphs

We first considered the simulation of three graphs. For all three graphs, $p = 100$ and $n = 150$. We first simulated the graph structure. We randomly selected 5% or 10% among all the possible edges and set them to be edges in graph 1. For graphs 2 and 3, we removed a portion (20% or 100%) of edges that were present in graph 1 and added back the same number of edges that were not present in graph 1. 20% represents the case that there is moderate shared structure. 100% represents the extreme case that there is little shared structure other than those shared by chance. For the entries in the precision matrices, we considered two settings: a) the upper-diagonal entries were sampled from uniform [−0.4, −0.1] ∪ [0.1, 0.4] independently and then set the matrix to be symmetric b) Same as that in a), except that for the shared edges, the corresponding entries were set to be the same. To make the precision matrix positive definite, we set the diagonal entry in a row to be the sum of absolute values of all the other entries in that row, plus 0.5.

The simulation results are presented in Figure 1. Our method (MRF) was compared with Guo's method (Guo et al., 2011), JGL (Danaher et al., 2014) and graphical lasso (*glasso*) (Friedman et al., 2008). In *glasso*, the graphs are estimated independently. In JGL, there are two options, fused lasso (JGL-Fused) and group lasso (JGL-Group). For Guo's method, *glasso* and JGL, we varied the sparsity parameter to generate the curves. For our method, we varied the threshold for the marginal posterior probabilities of γ to generate the curves. There are two tuning parameters in JGL, λ_1 and λ_2, where λ_1 controls sparsity and λ_2 controls the strength of sharing. We performed a grid search for λ_2 in {0, 0.05, …, 0.5} and selected the best curve. In Figure 1, our method performed slightly better than Guo's method. When there is little shared structure among graphs, our method performed slightly better than *glasso*, which is possibly due to the fact that we used a different modeling framework. When the entries were different for the shared edges, JGL-Fused did not perform well. However, when the entries were the same, JGL-Fused performed much better. The fused lasso penalty encourages entries in the precision matrix to be the same and JGL-Fused gains efficiency when the assumption is satisfied. We also performed simulations under high dimension settings ($p < n$) and simulations with a larger scale ($p = 500$). The results are similar and are shown in the Supplementary Materials. Moreover, we compared the methods for the detection of differential edges and shared edges in the graphs. We did not include JGL in the comparison as the similarity of the graphs is controlled with a tuning parameter. For the detection of shared edges, our method is better than Guo's, and Guo's method is better than *glasso*; for the detection of differential edges, our method is comparable to Guo's, and *glasso* is slightly better than both methods. In addition, we compared our method with Peterson et al. (2014), a Bayesian approach using G-Wishart priors. Peterson's method may not be applicable when p is moderately large or the number of graphs is more than a few (Supplementary Materials). We performed simulations with smaller scale and more replicates ($p = 20$, $n = 100$), where the setting is similar to that in Peterson et al. (2014). The results are shown in the Supplementary Materials. Our method performed slightly worse than Peterson's method, but better than Guo's method and JGL-Fused. Neighborhood selection methods may favor random graphs over graphs with hub structures. We also performed simulation for single graph, where the degree of nodes follows a power law distribution. Our method is comparable with *glasso*.

4.2 Joint estimation of multiple graphs with temporal dependency

In this setting, we assumed that the graph structure evolved over time by Hidden Markov Model (HMM). We set $p = 50$. At time $t = 1$, we randomly selected 10% among all the possible edges and set them to be edges. At time $t + 1$, we removed 20% of the edges at time t and added back the same number of edges that were not present at time t. The entries in the precision matrix were set the same as that in a) in Section 4.1. We present the simulation results in Figure 2, varying n and $|T|$. We compared our method with Guo's and JGL-Group, where the graphs were treated as parallel. Our method performed better than Guo's method and JGL-Group in all three settings, and the difference was greater when either n or $|T|$ increases. We did not include JGL-Fused in the comparison as the computational time for JGL-Fused increases substantially when the number of graphs is more than a few.

4.3 Joint estimation of multiple graphs with both spatial and temporal dependency

We simulated graphs in $|B| = 3$ spatial loci and $|T| = 10$ time periods. We set $p = 50$, $n = 100$, and sparsity~ 0.1. We first set the graphs in different loci at the same time point to be the same. The graph structure evolved over time by HMM similarly as that in Section 4.2, and 40% of the edges changed between adjacent time points. For all graphs, we then added some perturbations by removing a portion (10%, 20%, 50%) of edges and adding back the same number of edges. In each time period, the 3 graphs have similar degree of similarity with each other. The entries in the precision matrix were set the same as that in a) in Section 4.1. The simulation results are presented in Figure 3. Our method achieved better performance than all the other methods. We also compared the posterior distribution of η_s and η_t between real data and simulated data (Supplementary Materials). Based on η_s and η_t, the temporal dependency in the simulations is weaker than that in the real data; the simulations with 10% and 20% perturbations have stronger spatial dependency than that in the real data, and for the 50% one, it has similar degree of spatial similarity.

4.4 Computational time

We evaluated the computational speed of our approach in the estimation of single GGM and multiple GGMs. For single GGM, we compared our method (B-NS) with Bayesian Graphical Lasso (B-GLASSO) (Wang et al., 2012) in Figure 4a. Our algorithm took 0.5 and 4.5 minutes to generate 1,000 iterations for $p = 100$ and $p = 200$, and B-GLASSO took 1.6 and 17.9 minutes. The performance of graph structure estimation is comparable (Supplementary Materials). We also evaluated the speed of our algorithm for the joint estimation of multiple graphs, where n and p were both fixed to 100. The CPU time was roughly linear as the number of graphs increased (Figure 4b). When multiple processors are available, parallel computing will result in substantial gain in computational speed (Figure 4). Our model enables parallel computing in two levels: 1. for single graph estimation, the rows in β can be updated in parallel ("rows parallel"); 2. for multiple graphs estimation, the matrix β for each graph can be updated in parallel ("graphs parallel"). "graphs parallel" requires more memory and tends to outperform "rows parallel" on data with a smaller scale. The computations presented in Figures 4a and 4b were implemented on a dual-core CPU 2.4 GHz laptop running OS X 10.9.5 using MATLAB 2014a. The other computations were performed on the Yale University Biomedical High Performance Computing Center. The computational cost of our algorithm is $O(p^3)$.

5. Application to the human brain gene expression dataset

Next we apply our method to the human brain gene expression microarray dataset (Kang et al., 2011). In the dataset, the expression levels of 17,568 genes were measured in 16 brain regions across 15 time periods. The time periods are not evenly spaced over time and each time period represents a distinct stage of brain development. The median number of biological replicates per time per region is 5. Because of the small sample size, we collapsed the 16 regions into two regions: neocortical regions (11 regions) and non-neocortical regions (5 regions). The neocortical regions are more similar with each other (Kang et al., 2011). We excluded the data from time periods 1 and 2 in our analysis because they represent very early stage of brain development, when most of the brain regions sampled in future time periods have not differentiated. We first studied the network of 7 high confidence genes associated Autism Spectrum Disorders (ASD): GRIN2B, DYRK1A, ANK2, TBR1, POGZ, CUL3, and SCN2A(Willsey et al., 2013). ASD is a neurodevelopment disorder that affects the brain and have an early onset in childhood. With a good understanding on the networks of the 7 ASD genes, we hope to gain insight into how these genes interact to yield clues on their roles in autism etiology. The posterior mean and standard deviation for η_s were 0.61 and 0.47, respectively. The posterior mean and standard deviation for η_t were 1.27 and 0.48, respectively. The estimated model parameters suggest strong temporal dependency of the network structure. The estimated graphs are shown in Figure 5, and in the Supplementary Materials.

Time period 10 corresponds to early childhood (1 years ≤ age ≤ 6 years), which is the typical period that patients show symptoms of autism. Of particular interest are the genes that are connected with TBR1, which is a transcription factor that may directly regulate the expression of numerous other genes. GRIN2B is a potential target of TBR1 (Bedogni et al.,

2010) and Tbr1 has been shown to mediate the expression of Grin2b in adult mouse brain (Chuang et al., 2014). Interestingly, the edge between TBR1 and GRIN2B tends to be shared over time in the non-neocortical regions, but not in the neocortical regions: the edge inclusion probability tend to decrease after period 10 in the neocortical regions. Further biological experiments are required to validate the temporal dynamics of TBR1 and GRIN2B interaction. The marginal posterior probabilities of edge inclusion are compared between our approach (MRF) and the simpler approach not considering the data structure (B-NS) (Figure 6a). For a fair comparison, we set $q = 1/(1 + \exp(-\eta_1))$ in B-NS. Considering the data structure leads to a better separation of the marginal posterior probability.

To demonstrate the temporal dependency, we shuffled the time periods and re-implemented our approach. η_t tends to be smaller in the shuffled data while η_s has similar posterior distribution (Figures 6b, and 6c). Collapsing the neocortical and non-neocortical regions, we implemented our approach on three manually curated biological pathways: long-term potentiation (65 genes), long-term depression (57 genes), and GABAergic synapse (86 genes). Long-term potentiation and long-term depression are associated with memory and learning; Gamma aminobutyric acid (GABA) is the most abundant inhibitory neurotransmitter in the mammalian central nervous system. When we shuffled the time periods, η_t tends to be smaller in long-term potentiation and long-term depression and it is slightly smaller for GABAergic synapse (Figures 6d, 6e, and 6f). Only in long-term potentiation, one gene is overlapped with the ASD gene set. The estimated graphs for the three pathways are shown in the Supplementary Materials. A large fraction of the top edges are shared in the adjacent periods (Supplementary Materials). Before birth, from period 6 to 7, the networks tend to rewire and the trend is similar in the three pathways (Supplementary Materials).

6. Conclusion

In this paper, we proposed a Bayesian neighborhood selection procedure to estimate Gaussian Graphical Models. Incorporating the Markov Random Field prior, our method was extended to jointly estimating multiple GGMs in data with complex structures. Compared with the non-Bayesian methods, there is no tuning parameter controlling the degree of structure sharing in our model. Instead, the parameters that represent similarity between graphs are learnt adaptively from the data. Simulation studies suggest that incorporating the complex data structure in the jointly modeling framework would benefit the estimation. For the human brain gene expression data, we applied our method on the autism genes and three biological pathways related to the nervous system. We identified some interesting connections in the networks of autism genes. We also demonstrated the graph selection consistency of our procedure for the estimation of single graph.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Professor Forrest W. Crawford for the helpful discussions. This study was supported in part by the National Science Foundation grant DMS-1106738 and the National Institutes of Health grants R01 GM59507 and P01 CA154295. Tao Wang is supported by National Natural Science Foundation of China grant 11601326. Can Yang was supported in part by grant 61501389 from National Science Funding of China, grants 22302815 and 12316116 from the Hong Kong Research Grant Council, and grant FRG2/15-16/011 from Hong Kong Baptist University.

References

Bedogni F, Hodge RD, Elsen GE, Nelson BR, Daza RA, Beyer RP, Bammler TK, Rubenstein JL, Hevner RF. Tbr1 regulates regional and laminar identity of postmitotic neurons in developing neocortex. Proceedings of the National Academy of Sciences. 2010; 107:13129–13134.

Besag J. On the statistical analysis of dirty pictures. Journal of the Royal Statistical Society. Series B (Methodological). 1986:259–302.

Cai T, Liu W, Luo X. A constrained l1 minimization approach to sparse precision matrix estimation. Journal of the American Statistical Association. 2011; 106:594–607.

Chuang HC, Huang TN, Hsueh YP. Neuronal excitation upregulates tbr1, a high-confidence risk gene of autism, mediating grin2b expression in the adult brain. Frontiers in cellular neuroscience. 2014; 8:280. [PubMed: 25309323]

Chun H, Zhang X, Zhao H. Gene regulation network inference with joint sparse gaussian graphical models. Journal of Computational and Graphical Statistics. 2014:00–00.

Danaher P, Wang P, Witten DM. The joint graphical lasso for inverse covariance estimation across multiple classes. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2014; 76:373–397. [PubMed: 24817823]

Dobra A, Hans C, Jones B, Nevins JR, Yao G, West M. Sparse graphical models for exploring gene expression data. Journal of Multivariate Analysis. 2004; 90:196–212.

Dobra A, Lenkoski A, Rodriguez A. Bayesian inference for general gaussian graphical models with application to multivariate lattice data. Journal of the American Statistical Association. 2011; 106

Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008; 9:432–441. [PubMed: 18079126]

George EI, McCulloch RE. Variable selection via gibbs sampling. Journal of the American Statistical Association. 1993; 88:881–889.

George EI, McCulloch RE. Approaches for bayesian variable selection. Statistica sinica. 1997; 7:339–373.

Guo J, Levina E, Michailidis G, Zhu J. Joint estimation of multiple graphical models. Biometrika. 2011:asq060.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., Tibshirani, R. The elements of statistical learning. Vol. 2. Springer; 2009.

Kanehisa M, Goto S. Kegg: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000; 28:27–30. [PubMed: 10592173]

Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AM, Pletikos M, Meyer KA, Sedmak G, et al. Spatio-temporal transcriptome of the human brain. Nature. 2011; 478:483–489. [PubMed: 22031440]

Li F, Zhang NR. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. Journal of the American Statistical Association. 2010; 105

Lin Z, Sanders SJ, Li M, Sestan N, Zhao H, et al. A markov random field-based approach to characterizing human brain development using spatial–temporal transcriptome data. The Annals of Applied Statistics. 2015; 9:429–451. [PubMed: 26877824]

Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. The Annals of Statistics. 2006:1436–1462.

Meng XL, Wong WH. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. Statistica Sinica. 1996:831–860.

Narisetty NN, He X, et al. Bayesian variable selection with shrinking and diffusing priors. The Annals of Statistics. 2014; 42:789–817.

Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics. 2004; 5:155–176. [PubMed: 15054023]

Orchard P, Agakov F, Storkey A. Bayesian inference in sparse gaussian graphical models. arXiv preprint. 2013 arXiv:1309.7311.

Peterson, C., Stingo, F., Vannucci, M. Journal of the American Statistical Association. 2014. Bayesian inference of multiple gaussian graphical models; p. 00-00.

Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, et al. Towards a proteome-scale map of the human protein–protein interaction network. Nature. 2005; 437:1173–1178. [PubMed: 16189514]

Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of escherichia coli. Nature genetics. 2002; 31:64–68. [PubMed: 11967538]

Varin C, Reid N, Firth D. An overview of composite likelihood methods. Statistica Sinica. 2011:5–42.

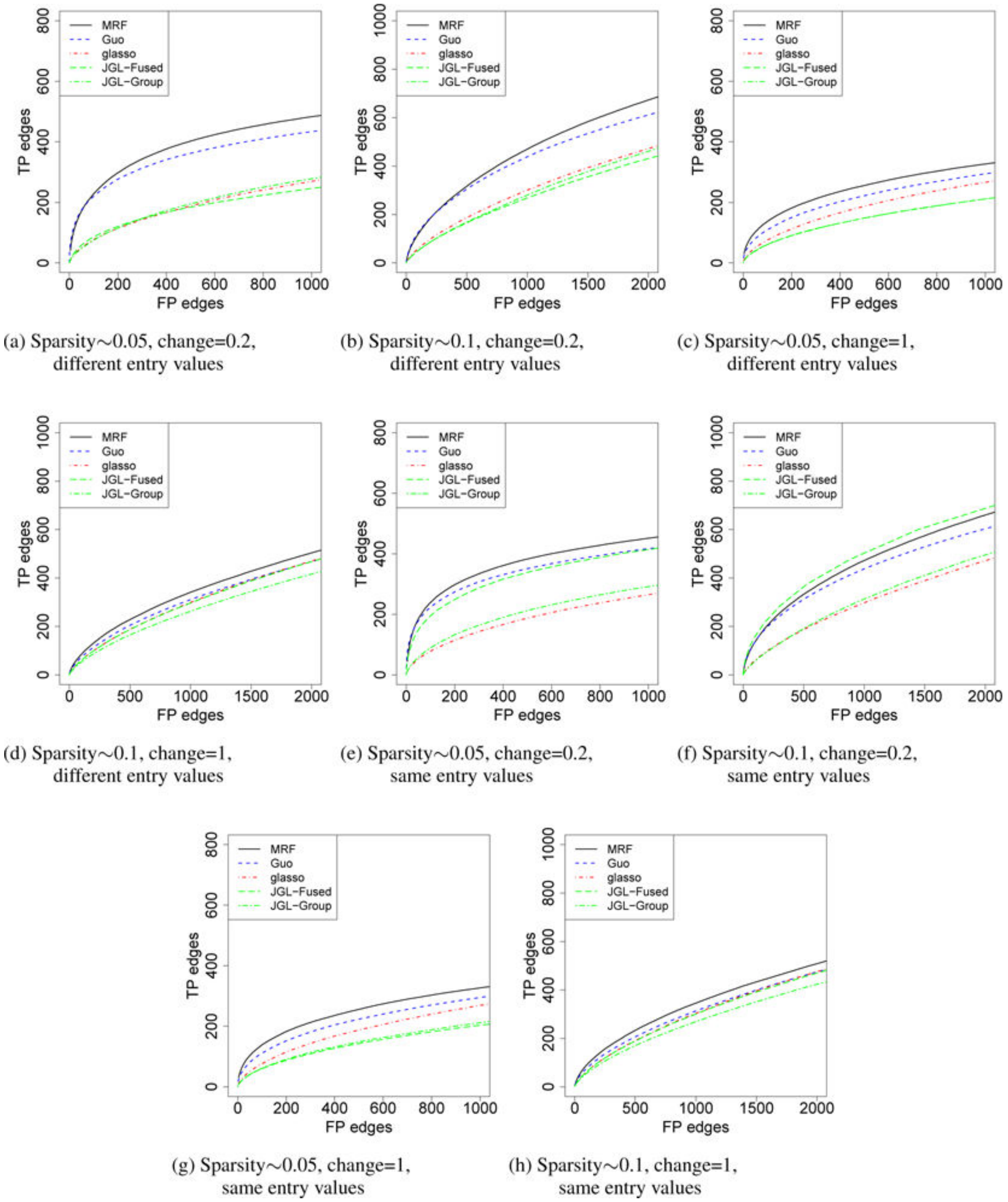
Wang H, et al. Bayesian graphical lasso models and efficient posterior computation. Bayesian Analysis. 2012; 7:867–886.

Willsey AJ, Sanders SJ, Li M, Dong S, Tebbenkamp AT, Muhle RA, Reilly SK, Lin L, Fertuzinhos S, Miller JA, et al. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. Cell. 2013; 155:997–1007. [PubMed: 24267886]

Yuan M, Lin Y. Model selection and estimation in the gaussian graphical model. Biometrika. 2007; 94:19–35.

Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Statistical applications in genetics and molecular biology. 2005; 4

Zhang CH, Huang J. The sparsity and bias of the lasso selection in high-dimensional linear regression. The Annals of Statistics. 2008:1567–1594.



(a) Sparsity∼0.05, change=0.2, different entry values

(b) Sparsity∼0.1, change=0.2, different entry values

(c) Sparsity∼0.05, change=1, different entry values

(d) Sparsity∼0.1, change=1, different entry values

(e) Sparsity∼0.05, change=0.2, same entry values

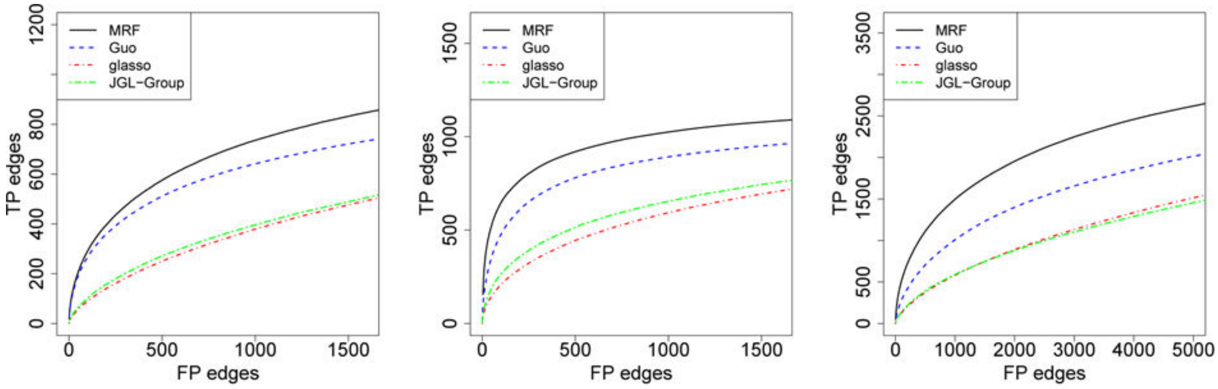
(f) Sparsity∼0.1, change=0.2, same entry values

(g) Sparsity∼0.05, change=1, same entry values

(h) Sparsity∼0.1, change=1, same entry values

Figure 1.

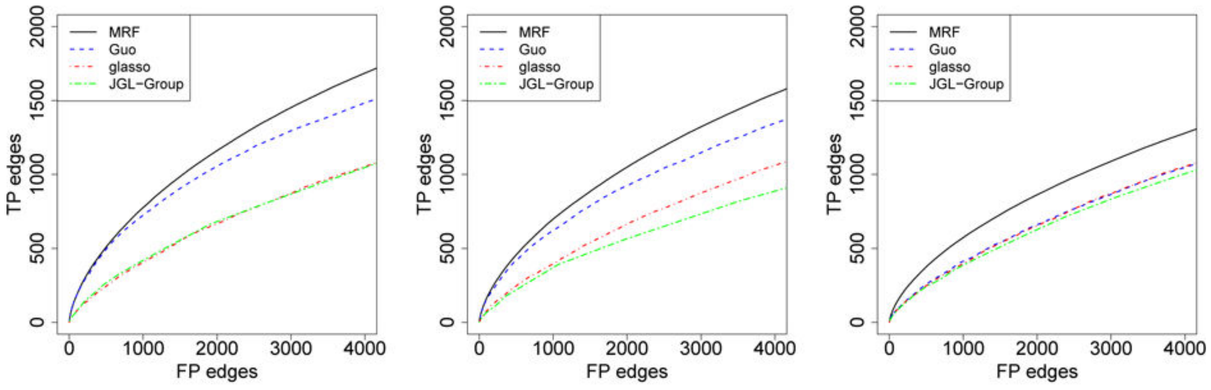
Comparisons of different models for the estimation of three graphs. For the shared edges, the corresponding entries in the precision matrices take the same (“same entry values”) or different (“different entry values”) non-zero values. The x-axis was truncated to be slightly larger than the total number of true positive edges. The curves represent the average of 100 independent runs.



(a) $n = 100$, $|T| = 10$ (b) $n = 200$, $|T| = 10$ (c) $n = 100$, $|T| = 30$

Figure 2.

Comparisons of different models for the estimation of mutiple graphs with temporal dependency. The x-axis was truncated to be slightly larger than the total number of true positive edges. The curves represent the average of 100 independent runs.



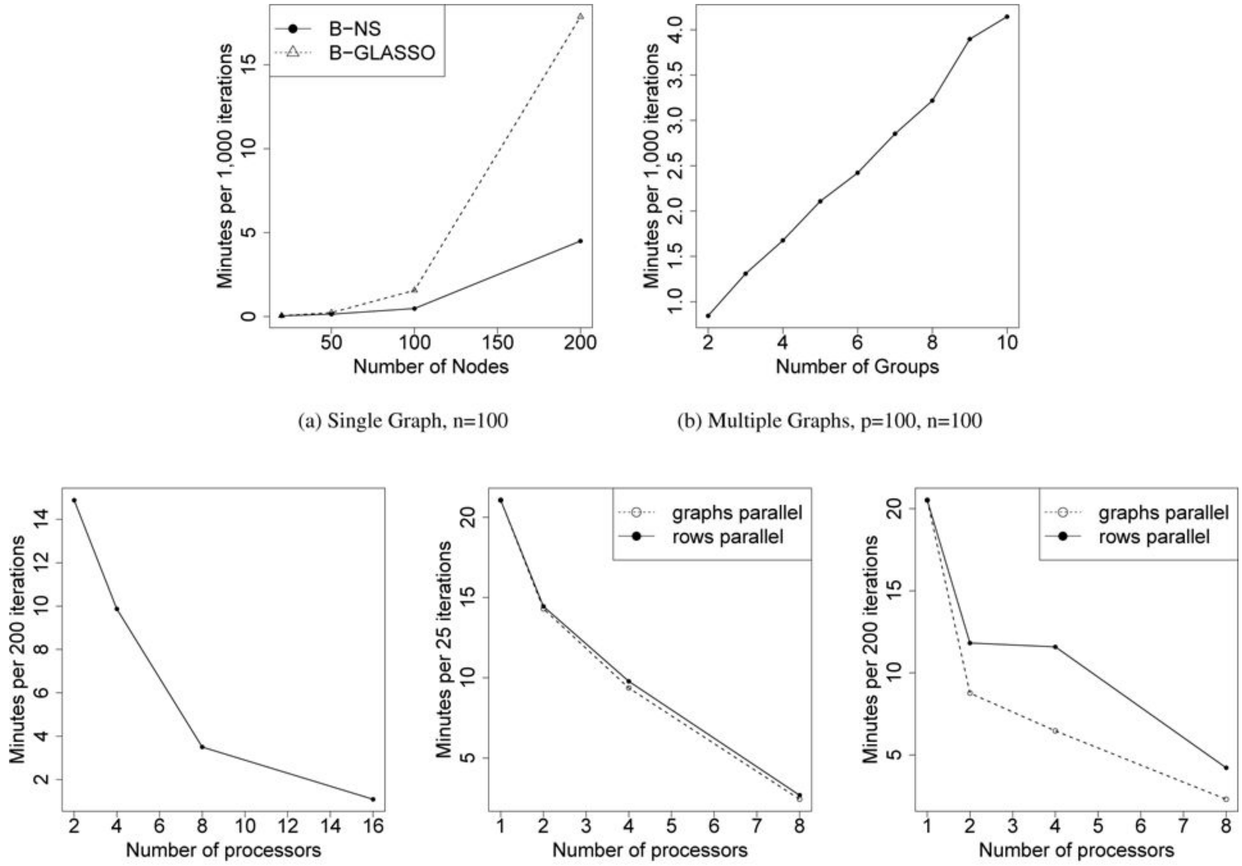
(a) perturbation=0.1

(b) perturbation=0.2

(c) perturbation=0.5

Figure 3.

Comparisons of different models for the estimation of mutiple graphs with temporal and spatial dependency. The x-axis was truncated to be slightly larger than the total number of true positive edges. The curves represent the average of 100 independent runs.



(a) Single Graph, n=100

(b) Multiple Graphs, p=100, n=100

(c) Single Graph, p=500, n=200

(d) 8 Graphs, p=500, n=200

(e) 16 Graphs, p=200, n=100

Figure 4.

Comparing the running time. (a) Single graph with increasing number of nodes, we compared our method (B-NS) with Bayesian Graphical Lasso (B-GLASSO) (Wang et al., 2012); (b) Multiple graphs with increasing number of graphs; (c–e) we implemented parallel computing.

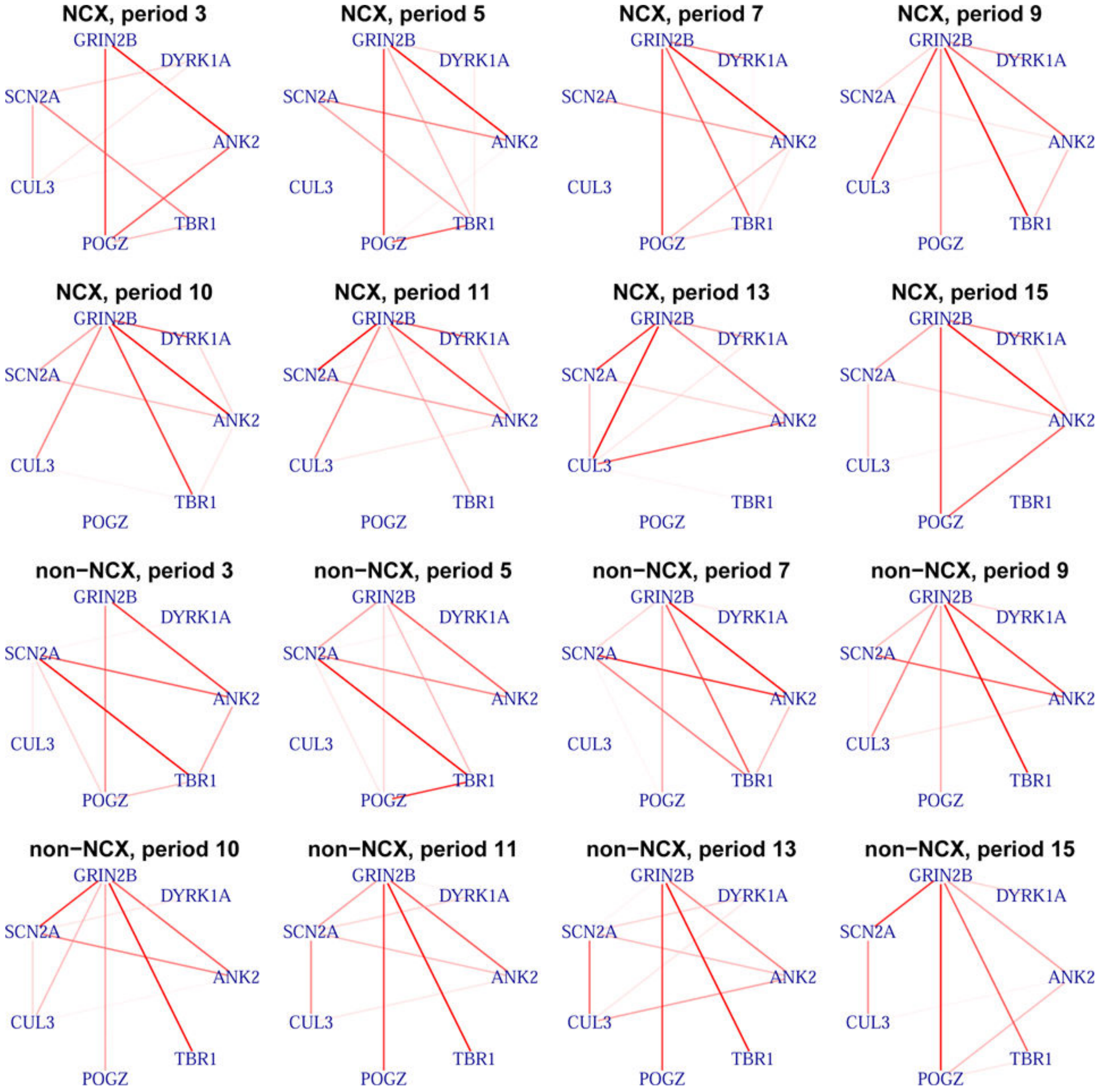
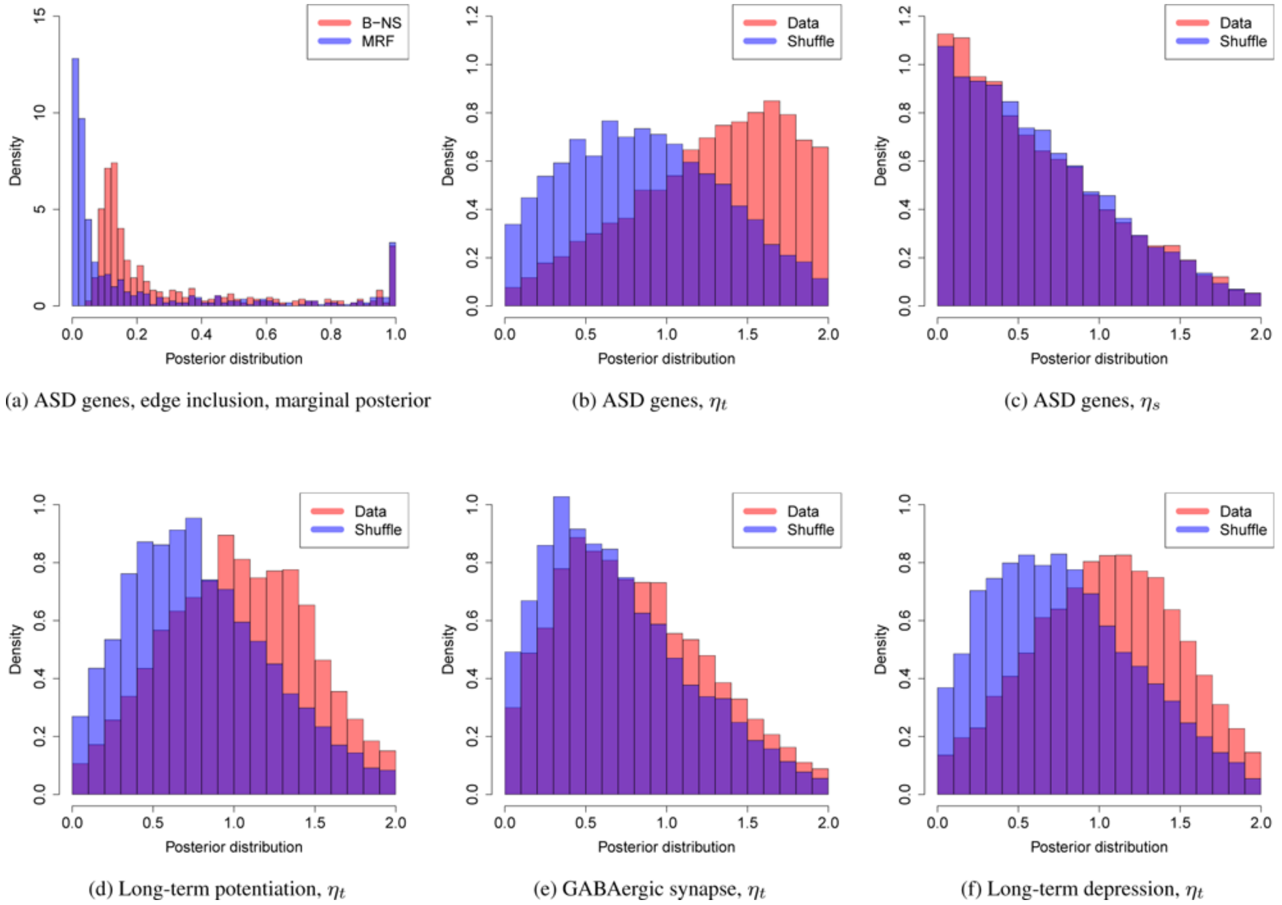


Figure 5.

The estimated graphs for the ASD genes. Period 10 corresponds to early childhood (1 years ≤ age ≤ 6 years), which is the typical period of autism onset. Some periods are skipped and are shown in the Supplementary Materials. The rank of the marginal probabilities is represented by the color gradient of the edges. NCX: neocortical regions; non-NCX: non-neocortical regions.



(a) ASD genes, edge inclusion, marginal posterior

(b) ASD genes, η_t

(c) ASD genes, η_s

(d) Long-term potentiation, η_t

(e) GABAergic synapse, η_t

(f) Long-term depression, η_t

Figure 6.

The posterior distributions. (a) The posterior distribution of the edge inclusion marginal probabilities, ASD genes; (b, c) The posterior distribution of the MRF model parameters, ASD genes; (d–f) The posterior distribution of η_t, the three pathways.