Journal of Machine Learning Research 17 (2016) 1-48 Submitted 12/15; Revised 6/16; Published 9/16

Joint Structural Estimation of Multiple Graphical Models

Jing Ma JINMA@UPENN.EDU
Department of Biostatistics and Epidemiology
Perelman School of Medicine
University of Pennsylvania
211 Blockley Hall, 423 Guardian Drive
Philadelphia, PA 19104, USA

George Michailidis GMICHAIL@UFL.EDU
Department of Statistics
University of Florida
205 Griffin-Floyd Hall, P.O. Box 118545
Gainesville, FL 32611, USA

Editor: Nicolai Meinshausen

Abstract

Gaussian graphical models capture dependence relationships between random variables through the pattern of nonzero elements in the corresponding inverse covariance matrices. To date, there has been a large body of literature on both computational methods and analytical results on the estimation of a *single* graphical model. However, in many application domains, one has to estimate several *related* graphical models, a problem that has also received attention in the literature. The available approaches usually assume that all graphical models are *globally* related. On the other hand, in many settings different relationships between subsets of the node sets exist between different graphical models. We develop methodology that *jointly* estimates multiple Gaussian graphical models, assuming that there exists prior information on how they are structurally related. For many applications, such information is available from external data sources. The proposed method consists of first applying neighborhood selection with a group lasso penalty to obtain edge sets of the graphs, and a maximum likelihood refit for estimating the nonzero entries in the inverse covariance matrices. We establish consistency of the proposed method for sparse high-dimensional Gaussian graphical models and examine its performance using simulation experiments. Applications to a climate data set and a breast cancer data set are also discussed.

Keywords: Gaussian graphical model, structured sparsity, group lasso penalty, consistency, edge set recovery

1. Introduction

There has been a large amount of work over the last few years on estimating Gaussian graphical models from high-dimensional data. In this family of models, jointly normally distributed random variables are represented by the nodes of a graph, while its edges reflect conditional dependence relationships amongst nodes that are captured through the nonzero entries of the inverse covariance matrix (or precision matrix) (Lauritzen, 1996; Edwards, 2000). Formally, let X be a p-dimensional multivariate normal random vector where

$$X = (X_1, \dots, X_p) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma).$$

For $1 \leq i \neq j \leq p$, X_i and X_j are said to be conditionally independent given all the remaining variables, if the corresponding entry in the precision matrix $\Omega = \Sigma^{-1}$ is zero. An edge between the nodes X_i and X_j in the graph implies that they are conditionally dependent and corresponds to a nonzero entry in the precision matrix. To identify the graph, one only needs to select the corresponding precision matrix.

Bühlmann and van de Geer (2011, chap. 13) gave an overview of statistical methods developed for estimating a Gaussian graphical model subject to sparsity constraints, an attractive feature that reduces the number of parameters to be estimated and also enhances interpretability of the results. These models have found applications in diverse fields including analysis of omics data (Perroud et al., 2006; Pujana et al., 2007; Putluri et al., 2011), reconstruction of gene regulatory networks (Dehmer and Emmert-Streib, 2008, chap. 6), as well as study of climate networks (Zerenner et al., 2014).

More recently, the focus has shifted from estimating a single graphical model to joint estimation of multiple graphs due to the availability of heterogeneous data (see discussion in Guo et al., 2011). For example, climate models capturing relationships between climate defining variables over a large area share common patterns; i.e. there are *shared common links* and also sharing of *absence of links* between the models (networks at different spatial locations). While separate estimation of individual models without taking the known pattern into consideration ignores the common structure, estimating one single model could mask the differences that could prove critical in understanding local climate features.

Several authors have studied the problem of *jointly* estimating multiple graphical models under different assumptions on how the models are related. Guo et al. (2011) introduced a procedure using a hierarchical penalty on the log-likelihood, whose objective is to estimate the common zeros (absence of edges) in the precision matrix across all graphical models under consideration. Thus, the procedure borrows strength across models through the the non-connected nodes, but does not impose any structure on the connected ones. Danaher et al. (2014) proposed a joint graphical lasso by maximizing the log-likelihood subject to a generalized fused lasso or group lasso penalty, which can be solved efficiently by a standard alternating directions method of multipliers algorithm (Boyd et al., 2011). When employing a group lasso penalty, the underlying assumption is that the various observed graphical models are *perturbations* of a *single* common connectivity pattern across all graphical models, while when using a fused lasso across all models a similar outcome occurs, although more heterogeneity between estimated graphical models can be obtained depending on the tuning of the penalties. The work by Zhu et al. (2014) investigates the joint estimation problem by introducing a truncated ℓ_1 penalty on the pairwise differences between the precision matrices to achieve entry-wise clustering of the network structure over multiple graphs. Peterson et al. (2015) introduced a Bayesian approach that links the estimation of the graphs via a Markov random field prior for common structures. Further, a spike-and-slab prior is placed on the parameters that measure the similarity between graphs, thus relaxing the assumption on sharing a common structure across all graphical models.

Despite recent advances in joint estimation algorithms, theoretical properties of the resulting estimators have not been fully investigated. Guo et al. (2011) represents an exception, wherein asymptotic properties of the resulting estimator are established for consistent recovery of the common zeros across multiple precision matrices, which is the focus of that procedure. Zhu et al. (2014) focused mainly on efficient computational algorithms when the graphs have disjoint subgraphs, with a brief mention of consistency of precision matrices in a special temporal setting; however, no theoretical guarantees are provided for more general settings. Finally, many papers only present algorithms for joint estimation of the Gaussian graphical models under consideration, but no theoretical properties of the estimates (Honorio and Samaras, 2010; Chiquet et al., 2011; Danaher et al., 2014; Mohan et al., 2014).

In this paper, we investigate estimation of multiple graphical models under *complex structural relationships*, assuming that there exists *prior information* on their specification. In many applications, such information is available and may come from prior knowledge in the literature of relationships among different node subsets of the graphical models under consideration, or from clustering of all graphs. The approach allows sharing common sub-graph components between different models and does not require sharing of values for the same element across multiple precision matrices. The proposed method, called the ***J****oint* ***S****tructural* ***E****stimation* ***M****ethod* (JSEM), leverages structured sparsity patterns as illustrated in Section 2 and is a two-step procedure. In the first step, we infer the sparse graphical models by incorporating the available structure through a group lasso penalty. In the second step, we maximize the Gaussian log-likelihood subject to the edge set constraints obtained from the previous step. Numerically, JSEM demonstrates superior performance in controlling both the number of false positive and false negative edges compared to available methods. When applied to joint modeling of climate networks, our results highlight the different roles climate defining factors play at different regions of the United States. In another application to breast cancer gene expression data, the JSEM methodology reveals interesting differences in the molecular network rewiring between the ER+ and ER- classes (see extensive discussion in Section 5.2). Understanding the rewiring of biological networks under different conditions provides deeper insights into biological mechanisms of disease, especially when combined with topology-based pathway enrichment methods as discussed and illustrated in Ma et al. (2016) and Kaushik et al. (2016).

The contributions of this work are three-fold. First, we develop a general framework for the problem of joint estimation of multiple Gaussian graphical models. The method can incorporate detailed structural information regarding relationships between subsets of the graphical models, while in the absence of such information reduces to the group graphical lasso procedure of Danaher et al. (2014). Further, we establish that the JSEM estimator is consistent with a fast rate of convergence in terms of the Frobenius norm for the estimated precision matrices. We also establish rigorously the consistent recovery of the edge sets for JSEM under suitable regularity conditions. Finally, when the externally provided structured sparsity pattern is moderately misspecified, we provide a modified estimator that reduces the number of false positive edges identified due to prior information misspecification, thus further enhancing the applicability of JSEM.

The paper is organized as follows. Section 2 discusses the structural relationships model used in this work and presents the estimation procedure. Section 3 presents the theoretical properties of the proposed method, followed by simulation studies in Section 4 and two real data applications—climate modeling and genomics of breast cancer—are presented in Section 5. We conclude with a discussion in Section 6. Most details of the theoretical analysis and proofs, additional simulation results as well as additional analyses on the applications are relegated to the Appendix.

2. The Joint Structural Estimation Method

Suppose we are interested in estimating K Gaussian graphical models from their corresponding K independent data sets, assuming that the models exhibit complex relationships between their edge sets. The data in the k-th model are organized in an $n_k \times p$ matrix $\mathbf{X}^k = (\mathbf{X}_1^k, \cdots, \mathbf{X}_p^k)$, where each

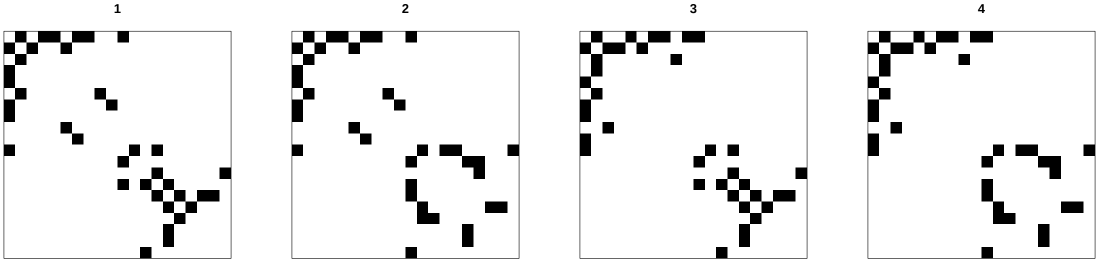


Figure 1: Image plots of the adjacency matrices for four graphical models with vertex set $\{1, \ldots, p\}$. The black color represents presence of an edge. The structured sparsity pattern is encoded in $\mathscr{G} = \cup_{1 \leq i < j \leq p} \mathscr{G}^{ij}$, where $\mathscr{G}^{ij} = \{[1, 3], [2, 4]\}$ for $(i, j) \in \{\lfloor p/2 \rfloor + 1, \ldots, p\}$ and $\mathscr{G}^{ij} = \{[1, 2], [3, 4]\}$ for all other pairs of (i, j).

row represents one observation from $\mathcal{N}(\mathbf{0}, \Sigma_0^k),\ k = 1, \ldots, K$. Throughout the remaining sections, we reserve the notations $\Sigma_0, \Omega_0, \ldots$ to denote the population parameters in the true model and use $\Sigma, \Omega, \ldots$ to denote generic parameters. Without loss of generality, we assume the columns of $\mathbf{X}^k$ are centered and standardized to have mean zero and unit variance. For ease of presentation, it is assumed that the sample size $n_k = n$ for all $k = 1, \ldots, K$, but the modeling framework can easily accommodate unequal sample sizes. Our goal is to estimate jointly $\Omega_0^k = (\Sigma_0^k)^{-1}$ for all k, under the assumption that the K corresponding graphs are related via a structured sparsity pattern $\mathscr{G}$. For example, consider climate models capturing relationships between climate forcing variables defined over a pre-specified spatial domain. Models that belong to the same climate zone may exhibit greater similarity in their graph structures than those from different zones. Thus, one can define $\mathscr{G}$ based on their spatial locations. Figure 1 gives an illustration of the structured sparsity among four graphical models in terms of their adjacency matrices. This pattern indicates that sharing of structures may occur at different subsets of the edge set, which motivates us to develop a joint estimation method that can incorporate such rich and complex structural information.

2.1 Neighborhood Selection

Neighborhood selection was introduced by Meinshausen and Bühlmann (2006) as an efficient method to construct Gaussian graphical models from high-dimensional data. For each node $i = 1, \ldots, p$ in the graphical model, consider the optimal prediction of the random variable X_i as a linear combination of the remaining variables:

$$X_i = \sum_{j \neq i} \theta_{ij} X_j + \varepsilon_i,$$

where $\theta_{ij}\ (j \neq i)$ are the regression coefficients and $\varepsilon_i \perp \{X_j : j \neq i\}$. The matrix $(\theta_{ij})_{1 \leq i,j \leq p}$ is determined by the inverse covariance matrix $\Omega = (\omega_{ij})_{1 \leq i,j \leq p}$. Specifically, it holds that $\theta_{ij} = -\omega_{ij}/\omega_{ii}$, for all $j \neq i$. The set of nonzero coefficients of $\theta_{ij}\ (j \neq i)$ is thus the same as the set of nonzero entries in the row vector of $\omega_{ij}\ (j \neq i)$, which defines the set of neighbors of node i. Using an l_1-penalized regression, Meinshausen and Bühlmann (2006) estimated the neighborhood for each node and combined the estimates to obtain the underlying graph.

2.2 An Illustrative Example

We first illustrate how to extend the idea of neighborhood selection to multiple graphical models using the example in Figure 1. For $k = 1, \dots, K$, let $(\theta^k_{ij})_{p \times p}$ be the matrix of regression coefficients in graph k and $\boldsymbol{\theta}^k_i$ the vector of all θ^k_{ij} $(j \neq i)$ for node $i = 1, \dots, p$. Unless otherwise stated, all vectors are assumed to be column vectors. For node i in a single graph k, neighborhood selection suggests estimating the coefficients $\boldsymbol{\theta}^k_i$ by

$$\min_{\boldsymbol{\theta}^k_i} \frac{1}{n} \|\mathbf{X}^k_i - \mathbf{X}^k_{-i}\boldsymbol{\theta}^k_i\|^2 + 2\lambda \sum_{j \neq i} |\theta^k_{ij}|,$$

where $\mathbf{X}^k_{-i}$ is $\mathbf{X}^k$ with the i-th column removed, $\|\cdot\|$ represents the standard Euclidean norm and λ is the regularization parameter. To achieve joint estimation, consider the following regularized regression problem

$$\min_{\Theta_i} \frac{1}{n} \sum_{k=1}^{K} \|\mathbf{X}^k_i - \mathbf{X}^k_{-i}\boldsymbol{\theta}^k_i\|^2 + 2P_\lambda(\Theta_i), \tag{1}$$

where $K = 4, \Theta_i = (\boldsymbol{\theta}^1_i, \dots, \boldsymbol{\theta}^K_i)$ and $P_\lambda(\Theta_i)$ is a regularization term to be determined next. Note that each column of Θ_i represents the regression coefficients from one graphical model and each row of Θ_i corresponds to the four coefficients at the same (i, j) pair.

The penalty $P_\lambda(\Theta_i)$ is chosen based on information from the structured sparsity pattern $\mathscr{G}$ in Figure 1. For example, for $i = 1$ with grouping $\{[1, 2], [3, 4]\}$,

$$\Theta_1 = \begin{pmatrix} \theta^1_{12} & \theta^2_{12} & \theta^3_{12} & \theta^4_{12} \\ & \vdots & & \vdots \\ \theta^1_{1p} & \theta^2_{1p} & \theta^3_{1p} & \theta^4_{1p} \end{pmatrix}.$$

As indicated by the colors, we can then group the coefficients in the j-th row of Θ_1 $(j = 2, \dots, p)$ as

$$(\underbrace{\theta^1_{1j}, \theta^2_{1j}}_{\boldsymbol{\theta}^{[1,2]}_{1j}}, \underbrace{\theta^3_{1j}, \theta^4_{1j}}_{\boldsymbol{\theta}^{[3,4]}_{1j}})$$

and set $P_\lambda(\Theta_1)$ to be the group lasso penalty

$$\sum_{j=2,\dots,p} \sum_{g=[1,2],[3,4]} \lambda^g_{1j} \|\boldsymbol{\theta}^{[g]}_{1j}\|.$$

The group lasso penalty forces the two coefficients in each group to be zero or nonzero at the same time, leading to the same structure for graphical models belonging to the same group.

The solution $\hat{\Theta}_i$ to (1) for $i = 1, \dots, p$ can then be used for graph selection.

2.3 The General Case

Denote the structured sparsity pattern by $\mathscr{G} = \cup_{1 \leq i < j \leq p} \mathscr{G}^{ij}$, where the union is over all $p(p-1)/2$ pairs of potential edges. Each $\mathscr{G}^{ij}$ is a partition of the set $\{1, 2, \cdots, K\}$ and consists of prior knowledge on the structural similarity for the (i, j)-th pair across models. For example in Figure 1, $\mathscr{G}^{ij} = \{[1, 2], [3, 4]\}$ means that the graphs 1 and 2 exhibit the same structure at (i, j), whereas 3 and 4 behave the same at (i, j). It is possible for all four graphs to have the edge (i, j) or not have the edge (i, j) at the same time, but we do not impose this restriction. Taking the union over all pairs, $\mathscr{G} = \{[1, 2], [3, 4], [1, 3], [2, 4]\}$ in Figure 1. Therefore the pattern $\mathscr{G}$ allows a more flexible structural relationships among multiple graphical models. Further, the sparsity pattern in $\mathscr{G}$ is symmetric as we require $\mathscr{G}^{ji} = \mathscr{G}^{ij}$ for $i < j$.

For $1 \le i < j \le p$ and a group $g \in \mathscr{G}^{ij}$, denote by $\boldsymbol{\theta}_{ij}^{[g]}$ the vector $(\theta_{ij}^k)_{k\in g}$, a concatenation of all regression coefficients from graphs in g. The grouping for the regression coefficients $(\theta_{ij}^1, \dots, \theta_{ij}^K)$ is determined by $\mathscr{G}^{ij}$. Under correctly specified $\mathscr{G}$, all coefficients in the same group should be zero or nonzero simultaneously. For $k = 1, \dots, K$, let $E^k = \{(i, j) : \theta_{ij}^k \neq 0\}$ be the set of undirected edges in graph k and $\mathcal{S}_{E^k}^+ = \{\Omega : \Omega \succ 0 \text{ and } \omega_{ij} = 0 \text{ for all } (i, j) \notin E^k \text{ where } i \neq j\}$.

The *Joint Structural Estimation Method* (JSEM) proceeds with the following two steps.

(I) For $k = 1, \dots, K$, we infer the sparse graphs $\hat{E}^k$ through the following group lasso estimator. For $i = 1, \dots, p$,

$$\min_{\Theta_i} \left\{ \frac{1}{n} \sum_{k=1}^{K} \|\mathbf{X}_i^k - \mathbf{X}_{-i}^k \boldsymbol{\theta}_i^k\|^2 + 2 \sum_{j:j\neq i} \sum_{g\in\mathscr{G}^{ij}} \lambda_{ij}^g \|\boldsymbol{\theta}_{ij}^{[g]}\| \right\}. \tag{2}$$

$\hat{E}^k$ is estimated to be the set

$$\{(i, j) : 1 \le i < j \le p, \hat{\theta}_{ij}^k \neq 0 \text{ OR } \hat{\theta}_{ji}^k \neq 0\}. \tag{3}$$

(II) We refit the model by

$$\min_{\Omega^k \in \mathcal{S}_{\hat{E}^k}^+} \left\{ \operatorname{tr}(\hat{\Sigma}^k \Omega^k) - \log\det(\Omega^k) \right\}, \quad k = 1, \dots, K. \tag{4}$$

Note the grouped variables in (2) are non-overlapping because $\mathscr{G}^{ij}$ partitions the set $\{1, \dots, K\}$ into disjoint subsets. The 'OR' rule defined in (3) can be replaced by the 'AND' rule. The problems in (2) and (4) are both convex and can thus be solved by available convex optimization algorithms. In this work, we use the R-package `grpreg` (Breheny and Huang, 2009) for implementation of the group lasso penalized optimization (2) and the `glasso` (Friedman et al., 2008) one for solving (4). The computational complexity for step (II) is $O(Kp^3)$ using the standard graphical lasso algorithm. Since `grpreg` uses a coordinate descent algorithm, the computational complexity for step (I) can be as fast as $O(nKp^2)$ if the number of graphs K does not exceed the sample size n, or $O(K^2p^2)$ otherwise. Thus, the overall computational complexity of JSEM is $O(Kp^3)$ if $p > K$, and $O(K^2p^2)$ otherwise.

2.4 Choice of Tuning Parameters

Like any other penalty-based method, JSEM requires selection of the tuning parameters λ_{ij}^g for all p regressions in (2). One can customize λ_{ij}^g for each 3-tuple (i, j, g) based on prior knowledge on graph similarity or simply use the same λ for all 3-tuples (i, j, g). In the sequel, we present results based on the latter approach. We recommend choosing the tuning parameters via the Bayesian information criterion (BIC). Specifically, for a given λ, we define BIC for the proposed method as

$$\text{BIC}(\lambda) = \sum_{k=1}^{K} \left\{ \text{tr}(\hat{\Sigma}^k \hat{\Omega}^k_\lambda) - \log\det(\hat{\Omega}^k_\lambda) + \frac{\log(n_k)}{n_k} |\hat{E}^k| \right\},$$

where $\hat{\Omega}^k_\lambda$ $(k = 1, \dots, K)$ are the estimated precision matrices from the data. The optimal tuning parameter is thus $\lambda^* = \text{argmin}_{\lambda \in \mathcal{D}_n} \text{BIC}(\lambda)$, where the set of values $\mathcal{D}_n$ is chosen such that for every $\lambda_j \in \mathcal{D}_n$ $(n_k = n)$:

$$\lambda_j = c_j \left(|g_{\max}| + \sqrt{\log G_0} \right) / \sqrt{n}, \quad c_j = 0.02 * j, \quad j = 1, \dots, 20.$$

Here $|g_{\max}|$ and G_0 refer, respectively, to the maximum size of groups in $\mathscr{G}$ and maximum total number of groups in all regressions. They can be conveniently defined by the input sparsity pattern. In practice, it is also recommended to apply the stability selection procedure (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013) to select graphical models that are both stable and interpretable.

3. Theoretical Results

The JSEM estimator enjoys nice theoretical properties under certain regularity conditions. Specifically, we establish the norm consistency of the estimated precision matrices, as well as the consistent recovery of the edge sets of the various graphical models under consideration based on the structured sparsity pattern $\mathscr{G}$.

3.1 Estimation Consistency

Let $\mathbb{N}^i_{(p-1)K} = \{(j,k) : j \neq i, k = 1, \dots, K\}$ be the variable index set for equation (2) with a fixed node i. Given the structural information $\mathscr{G}$, the grouped variable index set $\{(j,g) : j \neq i, g \in \mathscr{G}^{ij}\}$ defines a partition of $\mathbb{N}^i_{(p-1)K}$. Denote by G_i the cardinality of the set $\{(j,g) : j \neq i, g \in \mathscr{G}^{ij}\}$. Then $1 \leq G_i \leq (p-1)K$. Let $J(\Theta_{0,i}) = \{(j,g) : j \neq i, g \in \mathscr{G}^{ij}, \boldsymbol{\theta}^{[g]}_{0,ij} \neq 0\}$ be the set of nonzero groups in the i-th regression. We assume an overall sparsity at the group level, that is, the size of $J(\Theta_{0,i})$ is $s_i << G_i$. Let

$$G_0 = \max_{i=1,\dots,p} G_i, \quad s_0 = \max_{i=1,\dots,p} s_i, \quad S_0 = \sum_{i=1}^{p} s_i,$$

and also let $|g|$ be the size of the group g with $|g_{\max}| = \max_{g \in \mathscr{G}} |g|$.

Let $\mathbb{M}(p,K)$ represent the set of all $p \times K$ matrices. For $\Delta = (\boldsymbol{\delta}^1, \dots, \boldsymbol{\delta}^K) \in \mathbb{M}(p,K)$ and a group $g \subset \{1, \dots, K\}$, denote by $\boldsymbol{\delta}^{[g]}_j$ the vector composed of all δ^k_j for which $k \in g$. Write $\mathcal{J} = \{J(\Theta_{0,1}), \dots, J(\Theta_{0,p})\}$, the collection of sets of nonzero groups in all p regressions. For any $J \in \mathcal{J}$, denote Δ_J the nonzero matrix in $\mathbb{M}(p,K)$, which has the same coordinates as Δ on J and zero elsewhere. Let J^c denote the complement of the index set J. Write $\underline{0}$ as the zero matrix in $\mathbb{M}(p,K)$. We make the following assumptions.

(A1) For $0 < s < G_0$, there exists $\kappa = \kappa(s) > 0$, such that

$$\min_{J \in \mathcal{J}, |J| \leq s} \min_{\Delta \in \mathcal{F}_J} \frac{\sum_{k=1}^{K} \|\mathbf{X}^k \boldsymbol{\delta}^k\|^2/n}{\|\Delta_J\|_F^2} \geq \kappa^2(s),$$

where for i satisfying $J(\Theta_{0,i}) = J$, $\mathcal{F}_J$ is defined as

$$\mathcal{F}_J = \{\Delta : \Delta \in \mathbb{M}(p, K) \backslash \{\underline{0}\}, \sum_{(j,g) \in J^c} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\| \leq 3 \sum_{(j,g) \in J} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\|\}.$$

(A2) For every $k = 1, \ldots, K$ and $i = 1, \ldots, p$, $\mathrm{Var}(X_i^k) = 1$. Further, there exist constants c_0 and d_0 such that for every k,

$$0 < 1/c_0 \leq \phi_{\min}(\Sigma_0^k) \leq \phi_{\max}(\Sigma_0^k) \leq 1/d_0 < \infty,$$

where $\phi_{\min}(\Sigma_0^k)$ and $\phi_{\max}(\Sigma_0^k)$ are the minimum and maximum eigenvalues of the matrix Σ_0^k, respectively.

Assumption (A1) is a generalization of the Restricted Eigenvalue assumption for the Lasso in Bickel et al. (2009) to the group lasso setting in our problem and requires the super design matrix $\mathrm{diag}(\mathbf{X}^1, \ldots, \mathbf{X}^K)$ to be well conditioned over the restricted set of vectors under consideration. One sufficient condition is that the eigenvalues of the Gram matrix of $\mathrm{diag}(\mathbf{X}^1, \ldots, \mathbf{X}^K)$ is positive when restricted to the subset of sparse vectors with cardinality no greater than $2s$.

The equal variance requirement in assumption (A2) can be easily achieved by appropriate scaling of the data. The second part of the assumption explicitly excludes singular or nearly singular covariance matrices and guarantees that Ω_0^k exists for every model $k = 1, \ldots, K$.

We are now ready to state our first result.

Theorem 1 *Consider* $\hat{\Omega}^k$ $(k = 1, \ldots, K)$ *defined in* (4). *Let Assumption (A1) with* $s = 2s_0$ *and Assumption (A2) be satisfied. For every regression defined in* (2), *choose*

$$\lambda_{ij}^g = \frac{2}{\sqrt{nd_0}} \left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right),$$

with $q > 1$. *Then, with probability at least* $1 - 2pG_0^{1-q}$, *we have*

$$\frac{1}{K} \sum_{k=1}^{K} \|\hat{\Omega}^k - \Omega_0^k\|_F \leq O\left(\sqrt{\frac{S_0}{nK}} \left\{ \sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right\} \right), \qquad (5)$$

where G_0 *is the maximum number of groups in all regressions,* S_0 *is the total number of relevant groups and* $|g_{\max}|$ *is the maximum group size.*

Proof of Theorem 1 is available in Appendix A. Note the rate in (5) improves over estimating each precision matrix separately, as long as the sparsity pattern $\mathscr{G}$ is appropriately specified and nontrivial, i.e. there exists structural similarity among the considered graphical models. Further, the proposed procedure obtains a faster convergence rate than that of Guo et al. (2011) in some scenarios.

For example, if all K graphs share the same structure, then $|g_{\max}| = K$ and $G_0 = p - 1$. Thus, JSEM achieves a convergence rate of the order of

$$O\left(\sqrt{\frac{S_0}{n}}\left\{1 + \frac{\pi}{\sqrt{2}}\sqrt{\frac{q\log(p-1)}{K}}\right\}\right). \tag{6}$$

In contrast, separate estimation of Ω^k is known to be of the order of

$$O\left(\sqrt{\sum_k \|\Omega_0^{k,-}\|_0 \frac{\log p}{nK}}\right),$$

where $\|\Omega_0^{k,-}\|_0$ denotes the number of nonzero off-diagonal entries in Ω_0^k and $\sum_k$ is short-hand notation for $\sum_{k=1}^K$. The joint estimation method by Guo et al. (2011) has the following convergence rate

$$O\left(\sqrt{(p+m)\frac{\log p}{nK}}\right),$$

where $m = |\cup\{k = 1, \ldots, K : \omega_{0,ij}^k \neq 0\}|$. Under correctly specified $\mathscr{G}$, we have $S_0 = m$. Thus, JSEM has a lower estimation error rate than the joint estimation method of Guo et al. (2011). JSEM also outperforms separate estimation if $S_0 \asymp \|\Omega_0^{k,-}\|_0$, where $\asymp$ means that the expressions on both sides are of the same order. On the other hand, the rate in (6) could be worse if the sparsity pattern $\mathscr{G}$ is highly misspecified such that the number of nonzero parameters $S_0 > \sum_k \|\Omega_0^{k,-}\|_0 \geq m$. The issue of sparsity pattern misspecification is addressed in the next section.

3.2 Graph Selection Consistency

To understand how JSEM performs in selecting the edge sets of the graphical models, it suffices to focus on each of the group lasso estimation problems (2), as consistent graph selection relies on consistent variable selection in all p regressions. Unlike the sign consistency in the lasso setting (Zhao and Yu, 2006), variable selection properties with a group lasso penalty are much more complicated because the latter selects whole groups rather than individual variables (see Basu et al., 2015, and the discussion therein). The Basu et al. (2015) paper offers a generalization and introduces the notion of direction consistency for the group lasso. Specifically, for a nonzero vector $\boldsymbol{\xi}$, its direction vector is defined as $D(\boldsymbol{\xi}) = \boldsymbol{\xi}/\|\boldsymbol{\xi}\|$ and $D(\mathbf{0}) = \mathbf{0}$. An estimator $\hat{\Theta}_i$ of (2) is *direction consistent* at rate α_n if for a sequence of positive real numbers $\alpha_n \to 0$,

$$\mathbb{P}(\|D(\hat{\boldsymbol{\theta}}_{ij}^{[g]}) - D(\boldsymbol{\theta}_{0,ij}^{[g]})\| < \alpha_n,\ \forall\,(j,g) \in J(\Theta_{0,i}); \hat{\boldsymbol{\theta}}_{ij}^{[g]} = \mathbf{0},\ \forall\,(j,g) \notin J(\Theta_{0,i})) \to 1,$$

as $n, p \to \infty$. In general, direction consistency does not guarantee sign consistency, especially when there are multiple members within one group. However, if the group is selected, all the members within the group are selected, which is sufficient for joint neighborhood selection for each node and subsequent selection of graphs. Motivated by the above idea, we establish the graph selection consistency property of JSEM in Theorem 2, which can be conveniently modified to adjust for the misspecification in the prior information $\mathscr{G}$. Before we present the main result, we need more notations.

Consider the group lasso estimation problem (2) for node i. For simplicity, we discuss the estimation consistency properties with a common tuning parameter λ for all (j, g). For $k = 1, \ldots, K$, denote $\mathbf{X}_{I_k}^k$ the $n \times |I_k|$ sub-matrix consisting of all relevant variables from the k-th model. In other words, for all $j \in I_k$, there exists a group $g \ni k$ such that $(j, g) \in J(\Theta_{0,i})$. Note the dependency of each index set I_k on i is made implicit here for notational convenience. Further, let $\boldsymbol{\xi}^k \in \mathbb{R}^{|I_k|}$ be a vector indexed by I_k. The following assumption adapts the *Uniform Irrepresentability Condition (IC)* in Basu et al. (2015) to our setting:

(A3) There exists a positive constant η such that for all $\boldsymbol{\xi} = ((\boldsymbol{\xi}^1)^T, \ldots, (\boldsymbol{\xi}^K)^T)^T \in \mathbb{R}^{\sum_k |I_k|}$ with $\max\limits_{(j,g)} \|\boldsymbol{\xi}_j^{[g]}\| \leq 1$ and all $(j, g) \notin J(\Theta_{0,i})$,

$$\left(\sum_{k \in g} \left[(\mathbf{X}_j^k)^T \mathbf{X}_{I_k}^k \{ (\mathbf{X}_{I_k}^k)^T \mathbf{X}_{I_k}^k \}^{-1} \boldsymbol{\xi}^k \right]^2 \right)^{1/2} \leq 1 - \eta. \tag{7}$$

Note the group level constraint (7) is required to hold for all p regressions and is less stringent than the IC for the selection consistency of lasso. In general, it is not easy to verify Assumption (A3). One sufficient condition, as suggested in Zhao and Yu (2006), is that the regression coefficients of $\mathbf{X}_j^k$ on $\mathbf{X}_{I_k}^k$ $(k = 1, \ldots, K)$ have Euclidean norm less than 1 for all $(j, g) \notin J(\Theta_{0,i})$.

Theorem 2 *Let Assumption (A1) with $s = s_0$, (A2) and (A3) be satisfied. Assume further that the sparsity pattern $\mathscr{G}$ is correctly specified. For every regression defined in* (2)*, choose*

$$\lambda \geq \max_{i,(j,g) \notin J(\Theta_{0,i})} \frac{1}{\eta} \frac{1}{\sqrt{nd_0}} \Big(\sqrt{|g|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \Big), \tag{8}$$

$$\alpha_n \geq \max_{i,(j,g) \in J(\Theta_{0,i})} \frac{1}{\kappa(s_0)} \frac{1}{\|\boldsymbol{\theta}_{0,ij}^{[g]}\|} \left\{ \lambda \frac{\sqrt{s_0}}{\kappa(s_0)} + \frac{1}{\sqrt{nd_0}} \Big(\sqrt{|g|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \Big) \right\}, \tag{9}$$

with $q > 1$. Then with probability at least $1 - 4pG_0^{1-q}$, we have simultaneously for all i

1. $\hat{\boldsymbol{\theta}}_{ij}^{[g]} = \mathbf{0}$*, for all $(j, g) \notin J(\Theta_{0,i})$,*
2. $\|\hat{\boldsymbol{\theta}}_{ij}^{[g]} - \boldsymbol{\theta}_{0,ij}^{[g]}\| < \alpha_n \|\boldsymbol{\theta}_{0,ij}^{[g]}\|$*, and hence $\|D(\hat{\boldsymbol{\theta}}_{ij}^{[g]}) - D(\boldsymbol{\theta}_{0,ij}^{[g]})\| < 2\alpha_n$ for all $(j, g) \in J(\Theta_{0,i})$.*

Further, if $\alpha_n < 1$, then

$$\mathbb{P}(\hat{E}^k = E_0^k, \forall\, k = 1, \ldots, K) \geq 1 - 4pG_0^{1-q}.$$

where $\hat{E}^k$ is defined in (3)*.*

Note the choice of λ in (8) is of the same order as the tuning parameter required for estimation consistency in Theorem 1. With the above choice of λ, α_n can be chosen to be of the order of $O(\sqrt{s_0}(\sqrt{|g_{\max}|} + \sqrt{\log G_0})/\sqrt{n})$. A proof of Theorem 2 can be found in Appendix B.

Bach (2008) using a strong irrepresentability assumption also establishes group support recovery. In this work, we take a different route, where a similar strong irrepresentability assumption leads to direction consistency. Then, we leverage the notion of direction consistency to propose *within group thresholding* which allows us to handle successfully moderate misspecification of the group structures, as discussed next. Further, from a technical perspective, we build on the Karush-Kuhn-Tucker (KKT) conditions inversion scheme introduced in Zhao and Yu (2006), and noting that the $\mathrm{sign}(\cdot)$ function in standard lasso KKT conditions is replaced by the $D(\cdot)$ function in the group lasso KKT conditions. Therefore, sign consistency has a natural generalized counterpart when considering optimization over groups.

When $\mathscr{G}$ is misspecified, it is possible that not all the members within a group have nonzero effects. However, the group lasso penalty may fail to exclude members with actual zero effect within the misspecified group, leading to the recovery of spurious edges. The following result implies that the property of direction consistency helps identify influential members within a group, that is, those with noticeable nonzero effects.

Corollary 3 *Let Assumption (A1) with $s = s_0$, (A2) and (A3) be satisfied. For every regression defined in* (2)*, choose λ and α_n as in Theorem 2. Define*

$$\hat{\theta}_{ij}^{k,thr} = \hat{\theta}_{ij}^{k}\mathbf{1}\{\hat{\theta}_{ij}^{k}/\|\hat{\boldsymbol{\theta}}_{ij}^{[g]}\| > 2\alpha_n\},\ \forall\, k \in g,\ \forall\,(j,g) \in J(\Theta_{0,i}),$$

and

$$\hat{E}^{k,thr} = \{(i,j) : 1 \le i < j \le p, \hat{\theta}_{ij}^{k,thr} \neq 0 \text{ OR } \hat{\theta}_{ji}^{k,thr} \neq 0\}.$$

If for all $g \in \mathscr{G}$, $\min_{k\in g} \theta_{0,ij}^{k}/\|\boldsymbol{\theta}_{0,ij}^{[g]}\| > 2\alpha_n$, then

$$\mathbb{P}(\hat{E}^{k,thr} = E_0^k, \forall\, k = 1,\ldots,K) \ge 1 - 4pG_0^{1-q}.$$

The result in Corollary 3 implies immediately that JSEM with an additional thresholding step on the estimated direction vectors $D(\|\hat{\boldsymbol{\theta}}_{ij}^{[g]}\|)$ can be applied to reduce false discoveries and thus improve selection of the edge sets when the structured pattern $\mathscr{G}$ is moderately misspecified (that is, most of the structural relationships specified in $\mathscr{G}$ are reliable). This is illustrated in the third simulation study of Section 4.

4. Performance Evaluation

We present three simulation studies to evaluate the performance of JSEM. Other methods compared include the separate estimation method Glasso, where the *Graphical lasso* by Friedman et al. (2008) is applied to each graphical model separately, joint estimation by Guo et al. (2011), denoted by JEM-G, the Group Graphical Lasso denoted by GGL by Danaher et al. (2014), and the structural pursuit method MGGM by Zhu et al. (2014). Note we choose MGGM over the Fused Graphical Lasso method (Danaher et al., 2014), as the former has been consistently shown to exhibit better performance.

The first study considers a *single common structure* across all graphical models, while the second one features a *more complex structured sparsity pattern*. Our comparisons are based on the overall performance of different methods in terms of their ROC curves, as well as their finite sample performance in identifying the corresponding graphical models. For the latter, we use BIC to select the tuning parameters for all methods; in addition, the maximum likelihood refitting step (4) is added to all joint estimation methods to ensure fair comparison. We point out that the first study is favorable to existing joint estimation methods due to high degree of structural similarity, while the second one with varying degrees of structural similarity is more favorable to the JSEM procedure. Nevertheless, the results show that JSEM outperforms these competing methods in both settings, even when the structured pattern is moderately misspecified.

The third simulation compares JSEM with its thresholded version under misspecified $\mathscr{G}$ using the experimental settings of the first two studies. In this setting, one also needs to select the within group thresholding α_n besides λ. As in previous simulations, we first select λ via BIC without any thresholding. At the optimal λ, we select α_n from the grid of values

$$\alpha_n(c) = c\left(|g_{\max}| + \sqrt{\log G_0}\right)/\sqrt{n}, \quad c \in \{0.1, 0.2, \ldots, 1\},$$

where $|g_{\max}|$ and G_0 are defined by the input sparsity pattern. The optimal α_n^* is selected as the one that minimizes the corresponding BIC.

We refer readers to Appendix C for additional simulation results, including comparison of all joint estimation methods with and without maximum likelihood refitting step (4), and large p settings.

4.1 Simulation Study 1

In our first simulation, we set $K = 5$, with each graphical model being of size $p = 100$. The structured pattern is constructed as follows: we first generate a scale-free network with edge set E_0 as the common structure shared across all graphs, shown in the left panel of Figure 2. To generate the edge set E^k, we randomly select a pair of $(i,j), i < j$ such that $(i,j) \notin E_0$ and add it to E^k. This procedure was repeated $\rho|E_0|$ times for each k, where ρ is a positive number corresponding to the ratio of individual edges to common ones. In this example, we set $\rho = 0.1$ to allow high structural similarity across graphs. Thus, all graphical models have the same degree of sparsity, with 108 or 2.2% of all possible edges present. Note that due to the sparse structure of each graph, the proportion of shared non-edges (common zeros in the adjacency matrices) among all models is 98%.

Given the edge set E^k, we then constructed the inverse covariance matrix with the nonzero off-diagonal entries in Ω^k being uniformly generated from the $[-1, -0.5] \cup [0.5, 1]$ interval. The positive definiteness of Ω^k is guaranteed by setting the diagonal elements to be $|\phi_{\min}(\Omega^k)| + 0.1$. The covariance matrix Σ^k is then determined by

$$\Sigma_{ij}^k = (\Omega^k)_{ij}^{-1}/\sqrt{(\Omega^k)_{ii}^{-1}(\Omega^k)_{jj}^{-1}}.$$

By construction, each Σ^k corresponds to the correlation matrix for the k-th graphical model. The sparsity pattern supplied for JSEM is $\mathscr{G} = \{1, \ldots, K\}$, that is assuming all graphical models share the same structure. Note by setting the parameter $\rho = 0.1$, we have created a situation where about 10% of the information in $\mathscr{G}$ is misspecified for JSEM. This is of interest for us to see whether JSEM is robust to pattern misspecification.

To compare the overall performance of all methods, we generated $n_k = 50$ samples from each $k = 1, \ldots, K$ and computed the average false positive and true positive rates of the estimated precision matrices over a fine grid of tuning parameters from 20 replications. The resulting ROC curves are shown in the right panel of Figure 2. Since both GGL and MGGM require two tuning parameters, one for controlling the *sparsity* of individual graph and the other for controlling the *similarity* across all graphs, we computed the ROC curves over a fine grid of the sparsity parameter while fixing the similarity regularization at four different levels (from low to high similarity), and plotted the one that has the largest value of area under the curve (AUC). The graph $\mathcal{U}$ supplied for MGGM is a complete graph such that each pair of graphical models is included in the fused lasso

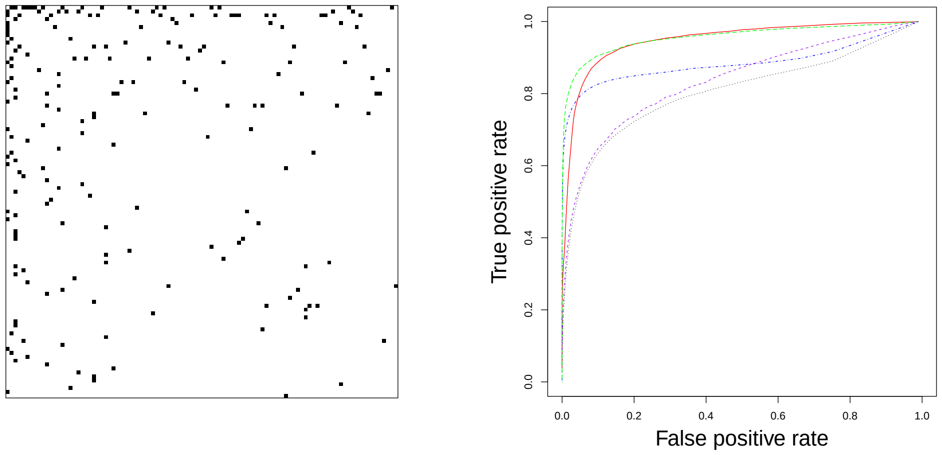


Figure 2: Simulation study 1: left panel shows the image plot of the adjacency matrix corresponding to the shared structure across all graphs. Each black cell indicates presence of an edge. The right panel shows the ROC curves for sample size $n_k = 50$: Glasso (dotted in black), JEM-G (dotdash in blue), GGL (solid in red), MGGM (dashed in purple), JSEM (longdash in green).

penalty. In this example, it turns out that GGL performs the best when there is only regularization on the similarity, i.e. a *group lasso* penalty on the same entry across all K precision matrices, which we expect to exhibit a similar performance to the proposed JSEM. In the right panel of Figure 2, the ROC curve of GGL falls slightly below that of JSEM. In comparison, MGGM does not perform as well despite the flexible penalty. The best curve we got from MGGM shows some advantage over the separate estimation Glasso, but mostly falls below curves from other joint estimation methods. JEM-G performs well and is very competitive compared to GGL and JSEM for very low false positive and high true positive rates, but starts falling behind when the false positive rate is greater than 5%. In this example, JSEM performs the best with the highest ROC curve throughout the domain.

Next, we computed the estimators from different methods with $n_k = 50$ samples for each $k = 1, \ldots, K$, using the tuning parameters selected by BIC. Results are summarized in Table 1, which compares the estimated precision matrices with the population version in the true model based on 50 replications under falsely discovered edges (FP), falsely deleted edges (FN), structural hamming distance (SHD), F_1 score (F1) and Frobenius norm loss (FL). The F_1 score (based on the effectiveness measure in Rijsbergen, 1979) measures the accuracy of a test by summarizing information from both FP and FN, where it reaches its best value at 1 and worst at 0. The results indicate that although GGL is good at controlling false positives, it tends to produce a high number of false negatives. The performance of MGGM is quite the opposite, with relatively small false negatives, but a huge number of false positive edges. In comparison, the proposed method JSEM achieves

Method	FP	FN	SHD	F1	FL
Glasso	35(6)	81(2)	116(5)	0.32(0.02)	0.73($<$ 0.01)
JEM-G	22(4)	40(4)	62(6)	0.69(0.03)	0.28(0.02)
GGL	**17**(6)	73(2)	90(6)	0.44(0.03)	0.29(0.02)
MGGM	286(13)	49(3)	335(13)	0.26(0.01)	0.64(0.02)
JSEM	19(4)	**35**(3)	**54**(6)	**0.73**(0.03)	**0.25**(0.02)

Table 1: Performance of different regularization methods for estimating graphical models in Simulation Study 1: average FP, FN, SHD, F1 and FL (SE) for sample size $n_k = 50$. The best cases are highlighted in bold.

a balance and obtains the highest F_1 score, as well as the lowest Frobenius norm loss. JEM-G performs slightly worse, but still well above the other three methods.

4.2 Simulation Study 2

In our second study, we consider a more structured pattern with $K = 10$ graphs. Each graphical model consists of $p = 50$ variables. Figure 3 shows the heat maps of the 10 adjacency matrices. This structured pattern is constructed as follows: we first generate the adjacency matrices corresponding to five distinct p-dimensional scale-free networks, so that the adjacency matrices in each column of the plot are the same. Next, we replace the connectivity structure of the bottom right diagonal block of size $p/2$ by $p/2$ in each adjacency matrix with that of another two distinct $p/2$-dimensional scale-free networks, so that graphical models in each column exhibit the same connectivity pattern except in the bottom right diagonal block of their adjacency matrices. Note that by replacing the connectivity structure among the second half of the nodes, the relationships between the first half and the second half of the nodes are also altered. In summary, this structured pattern illustrates how different subsets of the edge sets across multiple graphical models can be similar, as well as exhibit differences in their topologies. To the best of our knowledge, such complex relationships have not been studied in the literature. In this setting, the proportion of shared non-edges (common zeros in the precision matrices) among all graphical models is about 60%.

Given the adjacency matrix or equivalently the edge set E^k, we generate the covariance and inverse covariance matrices in the same way as in the first simulation study. The input sparsity pattern $\mathscr{G}$ supplied for JSEM and the graph $\mathcal{U}$ required in MGGM are defined according to the pattern in Figure 3. We also study the effect of misspecification in $\mathscr{G}$ by varying $\rho = 0, 0.2, 0.4, 0.6$, each corresponding to having only $(1 - \rho) * 100\%$ of the information in $\mathscr{G}$ being correct for JSEM.

At each level of pattern misspecification, we generated $n_k = 100$ independent samples for each $k = 1, \ldots, K$ and compared the ROC curves from different methods based on 20 replications in Figure 4. Again, the ROC curves for GGL and MGGM were optimized first with respect to the similarity regularization in terms of AUC. When $\rho = 0$, the results show a superior performance of JSEM, since it effectively incorporates available prior information across the various graphical models. JEM-G also yields a reasonably high ROC curve by taking advantage of the shared non-edges among all models. The performance of MGGM is comparable to that of JEM-G and much better than that of GGL. This is not surprising since MGGM benefits from knowing which pairs of

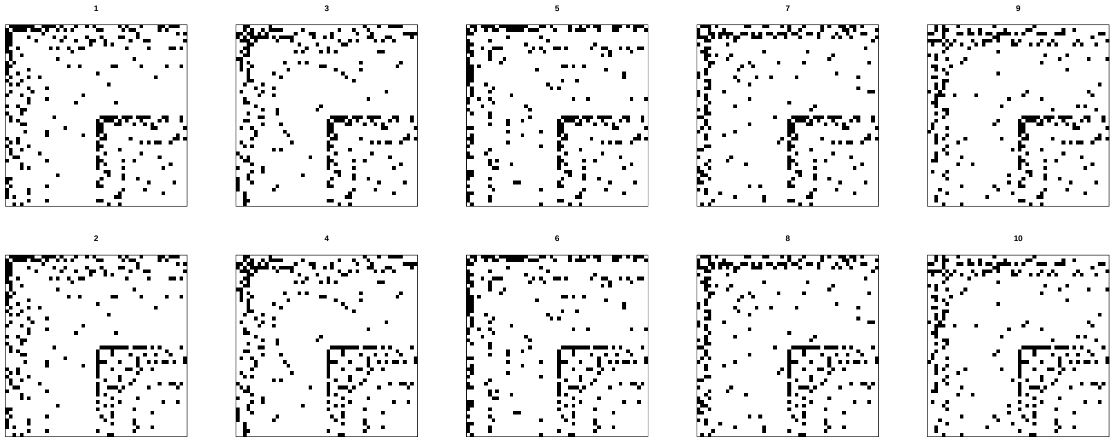


Figure 3: Simulation study 2: image plots of the adjacency matrices from all graphical models. Graphs in the same row share the same connectivity pattern at the bottom right block, whereas graphs in the same column share the same pattern at remaining locations.

graphical models to group. As ρ increases $(0 < \rho \leq 0.4)$, JSEM still performs the best despite the incorrectly specified $\mathscr{G}$, while other methods perform not much better than the separate estimation method Glasso. When $\rho = 0.6$, JSEM starts suffering from the large amount of pattern misspecification as well and performing not much better than separate estimation. Note at such high ρ values, the assumption of the presence of any related structures across graphical models becomes tenuous and therefore one is better off employing a separate estimation method for each graph.

Next, we examined the finite sample performance of different methods in identifying the true graphs and estimating the precision matrices at the optimal choice of tuning parameters. Table 2 shows the deviance measures between the estimated and the true precision matrices based on 50 replications for varying levels of pattern misspecification. For $\rho \leq 0.4$, JSEM achieves a good balance between FP and FN, and yields the highest F_1 score and lowest Frobenius norm loss. JEM-G is also very competitive in controlling false positive edges and comes next in overall performance. MGGM benefits from knowing the grouping structures and has comparable performance to JEM-G. In all cases, GGL achieves low FN, but very high FP, thus resulting in low F_1 scores. When $\rho = 0.6$, the advantage of using a joint estimation method begins to diminish due to the high heterogeneity and separate estimation is recommended.

4.3 Simulation Study 3

Finally, we illustrate how direction consistency helps improve the estimation of graphical models using the previous two experimental settings. Table 3 presents the performance of thresholded JSEM when $\mathscr{G}$ is moderately misspecified with individual to common ratio $\rho = 0.3$, based on 50 replications. Note that we used a larger sample size $n_k = 200$ in both settings to ensure that the Uniform IC required for direction consistency holds. The advantage of thresholding within groups is obvious in both settings, where the thresholded JSEM significantly reduces the number of false

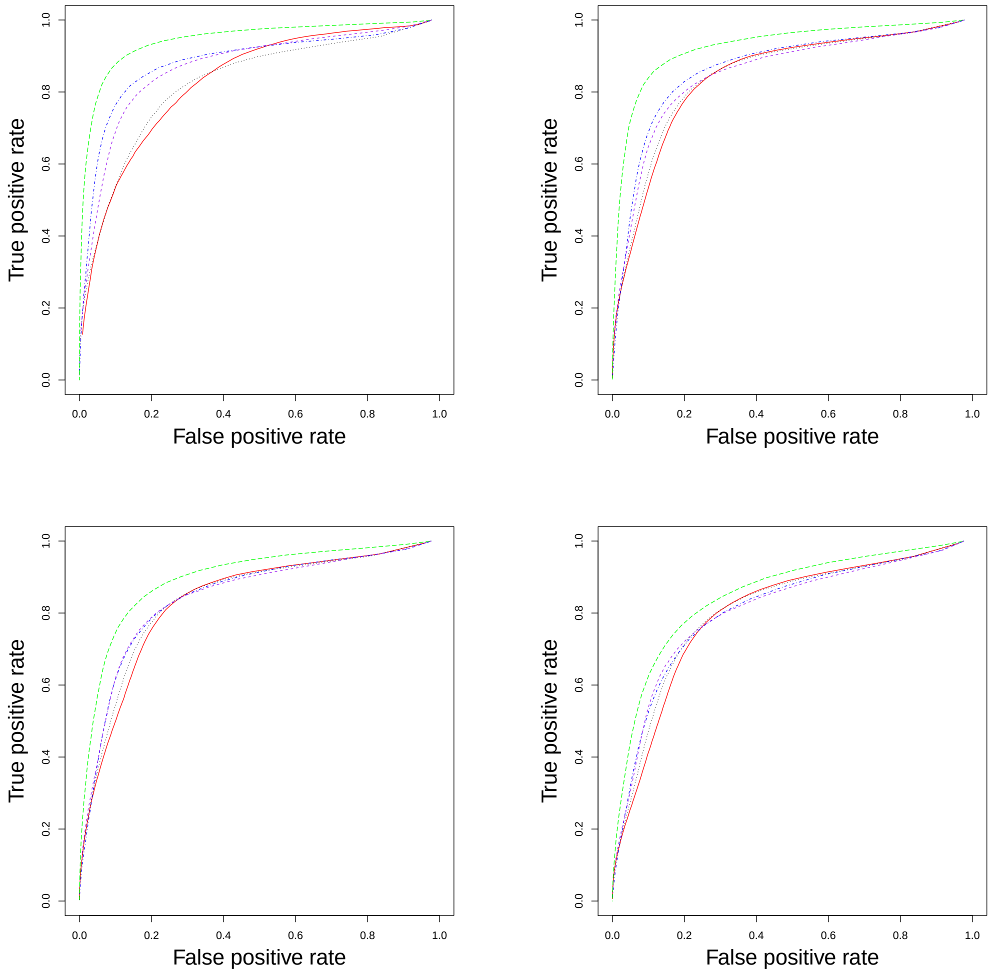


Figure 4: Simulation study 2: ROC curves for sample size $n_k = 100$: Glasso (dotted in black), JEM-G (dotdash in blue), GGL (solid in red), MGGM (dashed in purple), JSEM (longdash in green). The misspecification ratio ρ varies from (left to right): $0, 0.2$ (top row) and $0.4, 0.6$ (bottom row).

positive edges with only a small loss in the presence of false negative edges. One may notice the slight increase in Frobenius norm loss for thresholded JSEM, which is likely due to the increased presence of false negative edges. Nevertheless, the thresholded version of JSEM obtains higher F_1 scores, indicating an overall improvement in the structural estimation of all graphs.

We point out that the JSEM with thresholding procedure is most effective when ρ is moderate to small, such as $\rho < 0.5$ in this example. In other words, one believes most of the structural relationships are fairly reliable. If this is not the case, the numerical work presented strongly suggests that no joint estimation method works well, since the fundamental assumption of structural similarity

ρ	Method	FP	FN	SHD	F1	FL
0	Glasso	154(4)	38(1)	192(4)	0.51(0.01)	0.60(0.005)
	JEM-G	86(3)	**36**(2)	122(3)	0.62(0.01)	0.31(0.01)
	GGL	144(3)	39(1)	184(4)	0.52(0.01)	0.37(0.01)
	MGGM	30(2)	67(1)	97(2)	0.59(0.01)	0.36(0.01)
	JSEM	**21**(2)	42(2)	**63**(3)	**0.75**(0.01)	**0.28**(0.01)
0.2	Glasso	164(3)	**47**(1)	211(4)	0.53(0.01)	0.59(0.005)
	JEM-G	92(3)	57(2)	149(3)	0.59(0.01)	0.35(0.01)
	GGL	155(3)	48(1)	203(3)	0.53(0.01)	0.37(0.01)
	MGGM	94(3)	64(1)	158(4)	0.56(0.01)	0.37(0.01)
	JSEM	**32**(3)	64(2)	**96**(3)	**0.67**(0.01)	**0.32**(0.01)
0.4	Glasso	159(3)	**59**(1)	218(4)	0.55(0.01)	0.57(0.005)
	JEM-G	100(3)	77(2)	177(3)	0.56(0.01)	0.37(0.01)
	GGL	149(3)	61(2)	210(4)	0.55(0.01)	0.37(0.01)
	MGGM	119(3)	65(1)	184(3)	0.58(0.01)	0.37(0.01)
	JSEM	**49**(3)	84(2)	**132**(3)	**0.62**(0.01)	**0.36**(0.01)
0.6	Glasso	176(4)	**73**(2)	249(4)	0.54(0.01)	0.55(0.01)
	JEM-G	94(3)	109(2)	203(3)	0.52(0.01)	0.39(0.01)
	GGL	165(4)	76(2)	241(4)	0.54(0.01)	0.39(0.01)
	MGGM	109(3)	95(2)	204(4)	**0.55**(0.01)	0.39(0.01)
	JSEM	**50**(3)	123(2)	**173**(4)	0.52(0.01)	**0.38**(0.01)

Table 2: Performance of different regularization methods for estimating graphical models in Simulation Study 2: average FP, FN, SHD, F1 and FL (SE) for sample size $n_k = 100$. The best cases are highlighted in bold.

Design	Method	FP	FN	SHD	F1	FL
$K = 5, p = 100$,	JSEM	84(6)	**12**(1)	96(6)	0.71(0.01)	0.16(0.01)
$\mathscr{G} = \{1, 2, 3, 4, 5\}$	ThJSEM	**29**(4)	17(1)	**46**(4)	**0.83**(0.01)	0.16(0.01)
$K = 10, p = 40$,	JSEM	32(2)	**5**(0.7)	37(2)	0.78(0.01)	**0.17**(0.01)
$\mathscr{G}$ as in Figure 3	ThJSEM	**20**(2)	8(0.7)	**28**(2)	**0.82**(0.01)	0.19(0.01)

Table 3: Performance of JSEM and thresholded JSEM with misspecified groups ($\rho = 0.3$): average FP, FN, SHD, F1 and FL (SE) for sample size $n_k = 200$. The better cases are highlighted in bold.

among multiple models is violated. Instead, separate estimation is recommended for handling high heterogeneity among multiple graphical models.

5. Applications

To illustrate the proposed joint estimation method in inferring real-world networks, we applied JSEM to a climate data set to study relationships between climate defining variables at multiple locations in North America, as well as a breast cancer gene expression data extracted from The Cancer Genome Atlas project (TCGA, 2012).

5.1 Application to Climate Modeling

Recent assessments from the Intergovernmental Panel on Climate Change (IPCC, Stocker et al., 2013) indicate multiple lines of evidence for climate change in the past century and these changes have caused significant impacts on natural and human systems. One common approach towards understanding the climate system has been attribution studies of detected changes to internal and external forcing mechanisms (such as solar radiation, greenhouse gases, etc.) using simulated climate models. Lozano et al. (2009) used spatial-temporal modeling to study the attribution of climate defining mechanisms from observed data. In this work, we provide an alternative to learning the complex interactions among climate defining factors exhibited across different climate zones based on observed data.

The data used in this study are monthly measurements from January 2001 to June 2005 on 16 variables including mean temperature (TMP), diurnal temperature range (DTR), maximum and minimum temperature (TMX, TMN), precipitation (PRE), vapor pressure (VAP), cloud cover (CLD), rain days (WET), potential evapotranspiration (PET), frost days (FRS), greenhouse gases (carbon dioxide (CO2), carbon monoxide (CO), methane (CH4), hydrogen (H2)), aerosols (AER) and solar radiation (SOL) from CRU (`http://www.cru.uea.ac.uk/cru/data`), NOAA (`http://www.esrl.noaa.gov/gmd/dv/ftpdata.html`), NASA (`http://disc.sci.gsfc.nasa.gov/aerosols`) and NCDC (`ftp://ftp.ncdc.noaa.gov/pub/data/nsrdb-solar/`). The data are organized as a 2.5 degree latitude by 2.5 degree longitude grid across North America. To avoid complications from any seasonality or autocorrelation of the data, we aggregated the monthly time series into bins of 3-month intervals and took first differences of the quarterly data. The data after differencing were further normalized. Details on the pre-processing steps are included in Appendix D. Next, we randomly selected $K = 27$ locations spanning all types of climate from the 2.5 by 2.5 degree grid of North America (see Figure 5). This gives us an $n \times p$ matrix at each of the 27 locations, corresponding to $n = 17$ observations for the $p = 16$ climate defining variables. At each location, the conditional dependency network is of dimension $p \times p$, which has $16 \times 15/2 = 120$ edges to be inferred.

Our goal is to infer the conditional dependency networks for *all locations simultaneously* based on available spatial information, obtained from the classification of climate zones in Peel et al. (2007). Specifically, we assume that AER and SOL have one common connectivity pattern with other variables in the geographical south of North America and another common pattern in the north. The definition of the south and north is given in Figure 5. Variables on greenhouse gases (CO2, CO, CH4 and H2) are assumed to interact with other variables (except AER and SOL) in the same fashion within each of the four climate groups, that is Mid-latitude Desert, Semiarid Steppe, Humid Subtropical and Humid Continental. The connectivity patterns among all remaining variables are assumed to be the same within each of the six distinct climate zones in Figure 5.

We used BIC on the normalized data to select the tuning parameter λ for the proposed JSEM. At the optimal λ, we applied our method coupled with complementary pairs stability selection (Shah

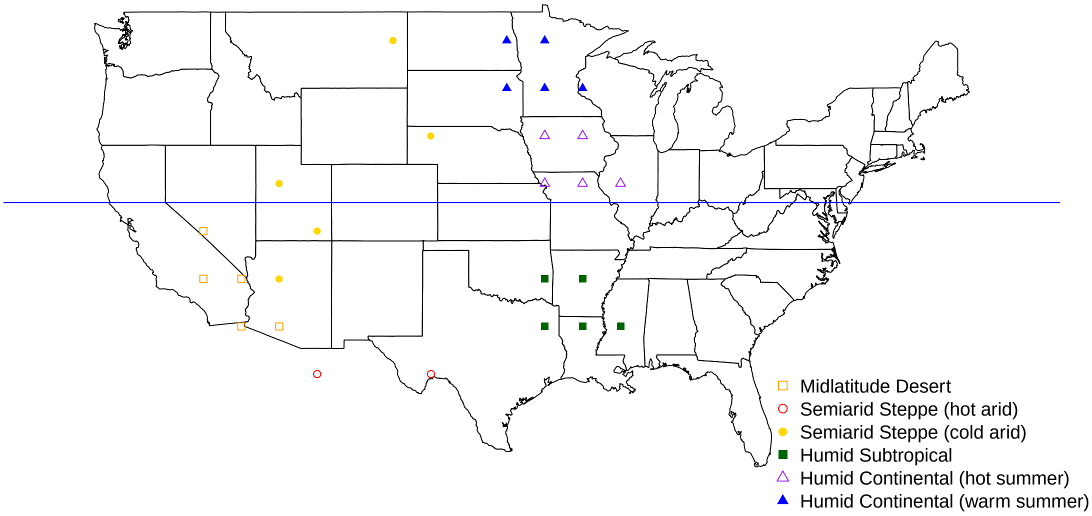

□ Midlatitude Desert
○ Semiarid Steppe (hot arid)
● Semiarid Steppe (cold arid)
■ Humid Subtropical
△ Humid Continental (hot summer)
▲ Humid Continental (warm summer)

Figure 5: The selected 27 locations based on climate classification. The solid line separates the south and north of North America and corresponds to latitude 39 N.

and Samworth, 2013) to identify the interaction networks at the 27 locations. To perform stability selection, we ran our method 50 times on two randomly drawn complementary pairs of sizes 8 and 9, and kept only edges that are selected over 70% of the time.

Due to space limitation, we present in Figure 6 the estimated networks at the six distinct climate zones. Readers are referred to Appendix D for the complete picture of the 27 networks from all the 27 locations under study, as well as more detailed comparisons. Although we do not impose the assumption on sharing of a single common structure across all locations, there are common edges (solid) identified for all climate zones, reflecting key features of climate defining regardless of geographical location. Such relationships are consistent with how the corresponding climate defining variables are defined, as well as how the data are collected (Harris et al., 2014). The Mid-latitude Desert and Semiarid Steppe climate zones share the edge between DTR and CLD, indicating that they are correlated conditional on all other variables. Similar relationships have also been found over drier regions in Zhou et al. (2009). In addition, one can see that the variable FRS interacts mainly with PET at Mid-latitude Desert and Semiarid Steppe climates, whereas it is partially correlated with both PRE and TMN (or TMX) at Continental climates. This can be explained from the distinction between these climate zones. At Humid Continental climate, precipitation is relatively well distributed year-round in most areas and snowfall occurs in all areas. It is thus not difficult to see why precipitation (PRE) and temperature related variables correlate with the number of frost days (FRS). Further, a primary criterion of an area characterized as Mid-latitude Desert or Semiarid Steppe is that it receives precipitation below potential evapotranspiration (PET), which possibly explains why FRS is partially correlated only with PET for Mid-latitude Desert and Semiarid Steppe climate. Finally, we point out that the inferred networks at neighboring climate zones are more similar, such as Semiarid Steppe (hot arid and cold arid), or Humid Continental (hot summer and

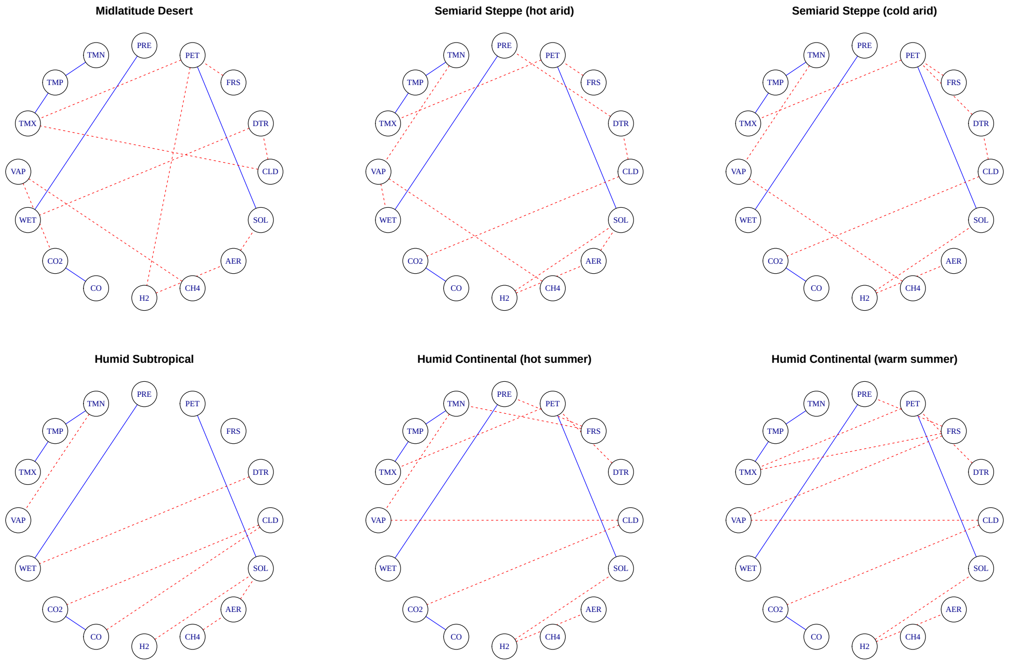

Midlatitude Desert
Semiarid Steppe (hot arid)
Semiarid Steppe (cold arid)
Humid Subtropical
Humid Continental (hot summer)
Humid Continental (warm summer)
PRE
PET
FRS
DTR
CLD
SOL
AER
CH4
H2
CO
CO2
WET
VAP
TMX
TMP
TMN

Figure 6: Estimated climate networks at the six distinct climate zones using JSEM, with edges shared across all locations blue solid and differential edges red dashed.

warm summer), whereas those with dramatically different climate show significantly different connectivity patterns. These common and individual interactions can prove critical in understanding the mechanisms of climate defining, and facilitate decision making in maintaining the best environmental results.

5.2 Application to Breast Cancer

Breast cancer is the most common cancer in women worldwide, with nearly 1.7 million new cases diagnosed in 2012 (second most common cancer overall). This represents about 12% of all new cancer cases and 25% of all cancers in women (Ferlay et al., 2013). Breast cancer is hormone related and this leads to a basic classification of cancer cells. Specifically, a cancer is called estrogen-receptor-positive (or ER+) if it has receptors for estrogen, and hence the cells receive signals from estrogen that could promote their growth. It is estimated that about 80% of all breast cancer cases are ER+ and they are more likely to respond to hormone therapy. Further, ER+ status is associated with better survival rates, especially if the cancer is diagnosed early. On the other hand, the ER- status lacks the estrogen receptor and in general exhibits poorer survival rates. Note that the presence/absence of other hormone receptors (progesterone and HER2) also play an important role in breast cancer tumor classification, therapeutic strategies and survival rates.

The breast cancer data set (TCGA, 2012) contains RNA-seq measurements for 17296 genes from 1033 breast cancer specimens, including ER+, ER- and other unevaluated cases. Due to the overall small sample size, we first reduced the number of variables by focusing on a subset of the genes that are present in the 44 KEGG pathways in Table 4. These pathways correspond to the major signaling and biochemical ones that have been reported in the literature of playing a significant role in all cancer types. This leaves for further consideration 800 genes with 403 samples from the ER+ and 117 from the ER- classes.

The structural similarity between the networks for ER+ and ER- status was defined based on the third column in Table 4, which indicates whether the pathway is significantly enriched when testing ER+ vs ER- status via NetGSA (Ma et al., 2016), and complemented through literature searches. If one pathway is not significantly enriched, then the genes belonging to the pathway are considered to share a common structure under both ER+ and ER- status. However, due to overlaps amongst pathways (since some of their members are assigned to multiple ones in the KEGG database), only genes that did not belong to any of the differential pathways were used to define the common structure. The remaining genes are assumed to have distinct structures under the two conditions.

We then used BIC on the normalized data to select the tuning parameter λ for the proposed JSEM. At the optimal λ, we applied our method coupled with complementary pairs stability selection (Shah and Samworth, 2013) to identify the interaction networks for the ER+ and ER- classes, respectively. Due to the large number of variables, visualization of the estimated networks at the individual gene level is challenging. Instead, we examine the interactions among pathways in Figure 7 to gain insight into their co-regulation behavior. The weighted pathway level network is defined as follows. Let each node in the network represent one pathway, with size proportional to the size of the corresponding pathway. A weighted edge between two pathways P_1 and P_2 is defined as the number of nonzero partial correlations between genes in P_1 and those in P_2 (normalized by the sizes of the two pathways). Links visualized in Figure 7 are the top 5% of the weighted edges, where ranking is based on edge weights. Note pathways that are isolated from all others were removed.

The first thing to note is that structural information provided enables us to estimate a much more dense graph than either separate estimation or an agnostic method like JEM-G (see Figure 12 in Appendix D), which in turn aids biological interpretation. We focus next on the interactions between pathways, as shown in Figure 7. The central role of known cancer related pathways—TGF-β, p53, MAPK and hedgehog—is apparent. Further, we see high degree of interconnections between signaling and biochemical pathways including glycolysis gluconeogenesis, pyrimidine, cysteine and methionine, and tryptophan, which is expected due to the impact of energy metabolism in tumor growth and progression. One surprising finding is that the p53 pathway is connected only in the ER+ class, but we suspect that this may be the case due to the big discrepancy in terms of available samples for the ER+ and ER- classes and the large number of genes present. In summary, the proposed method captures established cross-talk patterns between various signaling and biochemical pathways, which is not the case with competing methods or with separate estimation.

6. Discussion

This work introduces a flexible joint structural estimation method (JSEM) that incorporates *a priori* known structural relationships between multiple graphical models. The proposed method works well in situations where there is a large number of graphical models, but external similarity information is available only for sub-components of the models. In practice, if not all entry-wise structural relationships across multiple graphical models are available, it is recommended to add constraints at mainly edge pairs that are likely to share the same structures instead of providing a highly mis-

Vertex id	Vertex names	KEGG names	Status
1	glycolysis_gluconeogenesis	glycolysis_gluconeogenesis	TRUE
2	citrate_cycle_tca_cycle	citrate_cycle_tca_cycle	FALSE
3	pentose_phosphate	pentose_phosphate_pathway	TRUE
4	fructose_and_mannose	fructose_and_mannose_metabolism	TRUE
5	galactose	galactose_metabolism	TRUE
6	fatty_acid	fatty_acid_metabolism	FALSE
7	oxidative_phosphorylation	oxidative_phosphorylation	FALSE
8	purine	purine_metabolism	TRUE
9	pyrimidine	pyrimidine_metabolism	TRUE
10	glycine_serine_and_threonine	glycine_serine_and_threonine_metabolism	FALSE
11	cysteine_and_methionine	cysteine_and_methionine_metabolism	TRUE
12	valine_leucine_and_isoleucine	valine_leucine_and_isoleucine_degradation	TRUE
13	lysine	lysine_degradation	FALSE
14	arginine_and_proline	arginine_and_proline_metabolism	FALSE
15	tryptophan	tryptophan_metabolism	FALSE
16	beta_alanine	beta_alanine_metabolism	TRUE
17	glutathione	glutathione_metabolism	TRUE
18	starch_and_sucrose	starch_and_sucrose_metabolism	TRUE
19	amino_sugar_and_nucleotide_sugar	amino_sugar_and_nucleotide_sugar_metabolism	FALSE
20	ppar	ppar_signaling_pathway	TRUE
21	mapk	mapk_signaling_pathway	FALSE
22	erbb	erbb_signaling_pathway	TRUE
23	calcium	calcium_signaling_pathway	FALSE
24	chemokine	chemokine_signaling_pathway	TRUE
25	phosphatidylinositol	phosphatidylinositol_signaling_system	FALSE
26	cell_cycle	cell_cycle	TRUE
27	p53	p53_signaling_pathway	TRUE
28	mtor	mtor_signaling_pathway	FALSE
29	wnt	wnt_signaling_pathway	FALSE
30	notch	notch_signaling_pathway	FALSE
31	hedgehog	hedgehog_signaling_pathway	TRUE
32	tgf_beta	tgf_beta_signaling_pathway	TRUE
33	vegf	vegf_signaling_pathway	FALSE
34	toll_like	toll_like_receptor_signaling_pathway	TRUE
35	nod_like	nod_like_receptor_signaling_pathway	TRUE
36	rig_i_like	rig_i_like_receptor_signaling_pathway	FALSE
37	jak_stat	jak_stat_signaling_pathway	TRUE
38	t_cell	t_cell_receptor_signaling_pathway	FALSE
39	b_cell	b_cell_receptor_signaling_pathway	FALSE
40	fc_epsilon_ri	fc_epsilon_ri_signaling_pathway	TRUE
41	neurotrophin	neurotrophin_signaling_pathway	FALSE
42	insulin	insulin_signaling_pathway	FALSE
43	gnrh	gnrh_signaling_pathway	TRUE
44	adipocytokine	adipocytokine_signaling_pathway	TRUE

Table 4: List of simplified vertex (pathway) names, their matching names in KEGG and whether the corresponing pathway is used to define structural similarity

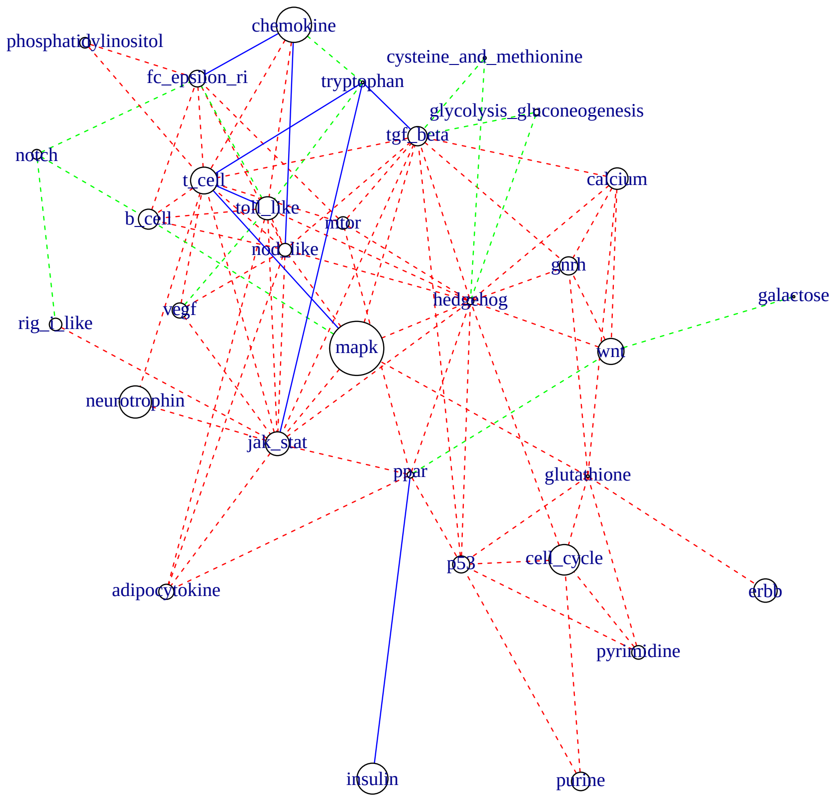


Figure 7: Estimated pathway networks for the ER+ and ER- classes using JSEM, with edges shared across all locations blue solid and differential edges red dashed (ER+) / green dashed (ER-).

specified structured sparsity pattern. On the other hand, if more structural information is available, one may generalize the group lasso penalty to incorporate additional structural constraints.

The theoretical guarantees of JSEM rely on two important, but standard in the literature, assumptions: the restricted eigenvalue assumption (A1) and the uniform IC assumption in (A3). In practice, it might be difficult to verify whether these assumptions are fulfilled, especially the more stringent assumption (A3). For the latter condition, Meinshausen and Yu (2009) observe that the irrepresentability condition (a variant of A3) may be violated in practical settings in the presence of highly correlated variables; nevertheless, the lasso estimates are still ℓ_2 consistent, under (A1).

Acknowledgments

The authors would like to thank two anonymous reviewers for helpful comments and suggestions. The work of GM was supported in part by NSF awards DMS-1228164 and DMS-1545277 and NIH award 7R21GM10171903.

Appendix A. Proof of Theorem 1

To prove the rate of convergence in Theorem 1, we look at three key steps: nodewise regression in subsection A.1, selecting the edge set in A.2 and maximum likelihood refitting in A.3. More information can be found in Appendix E. When it is clear, we shall use $\sum_k$ as a short notation for $\sum_{k=1}^K$.

A.1 Regression

For $j \neq i, g \in \mathscr{G}^{ij}, k \in g$, let $\boldsymbol{\varepsilon}_i^k = \mathbf{X}_i^k - \sum_{j \neq i} \theta_{0,ij}^k \mathbf{X}_j^k$. Let $\langle a, b \rangle$ represent the inner product between two vectors a and b. Denote $\zeta_{ij}^k = \langle \boldsymbol{\varepsilon}_i^k, \mathbf{X}_j^k \rangle / n$ and $\boldsymbol{\zeta}_{ij}^{[g]} = (\zeta_{ij}^k)_{k \in g} \in \mathbb{R}^{|g|}$. Consider the random event $\mathcal{A} = \bigcap_{i,j \neq i, g} \mathcal{A}_{ij}^g$, where $\mathcal{A}_{ij}^g = \{2\|\boldsymbol{\zeta}_{ij}^{[g]}\| \leq \lambda_{ij}^g\}$. By Lemma E.2, if we choose λ_{ij}^g as

$$\lambda_{ij}^g \geq \max_{k \in g} \frac{2}{\sqrt{n\omega_{0,ii}^k}} \left(\sqrt{|g|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right) \tag{10}$$

with $q > 1$, then $\mathbb{P}(\mathcal{A}) \geq 1 - 2pG_0^{1-q}$. We first present the following proposition that establishes oracle bounds for $\hat{\Theta}_i - \Theta_{0,i}$ under the chosen λ_{ij}^g.

Proposition A.1 *For $i = 1, \ldots, p$, consider the problem* (2) *and choose λ_{ij}^g as in* (10)*. Let $\hat{\Theta}_i$ be the solution to problem* (2)*. If Assumption (A1) holds with $\kappa^2 = \kappa^2(s_0)$, then for any solution $\hat{\Theta}_i$ of problem* (2)*, we have on the event $\mathcal{A}$*

$$\sum_{j \neq i, g \in \mathscr{G}^{ij}} \|\hat{\boldsymbol{\theta}}_{ij}^{[g]} - \boldsymbol{\theta}_{0,ij}^{[g]}\| \leq \frac{16}{\kappa^2 \lambda_{\min}} \sum_{(j,g) \in J(\Theta_{0,i})} (\lambda_{ij}^g)^2, \tag{11}$$

$$\mathcal{M}(\hat{\Theta}_i) \leq \frac{64\phi_{\max}}{\kappa^2 \lambda_{\min}^2} \sum_{(j,g) \in J(\Theta_{0,i})} (\lambda_{ij}^g)^2, \tag{12}$$

where $\lambda_{\min} = \min_{i,j \neq i, g \in \mathscr{G}^{ij}} \lambda_{ij}^g$, $\mathcal{M}(\hat{\Theta}_i) = |J(\hat{\Theta}_i)|$ and $\phi_{\max}$ is the maximal eigenvalue of $(\mathbf{X}^k)^T \mathbf{X}^k / n$ for all $k = 1, \cdots, K$. If, in addition, Assumption (A1) holds with $\kappa^2(2s_0)$, then for any solution $\hat{\Theta}_i$ of problem (2) *we have that*

$$\|\hat{\Theta}_i - \Theta_{0,i}\|_F \leq \frac{4\sqrt{10}}{\kappa^2(2s_0)} \frac{\sum_{(j,g) \in J(\Theta_{0,i})} (\lambda_{ij}^g)^2}{\lambda_{\min}\sqrt{s_i}}. \tag{13}$$

By Assumption (A2), $\omega_{0,ii}^k \geq \phi_{\min}(\Omega_0^k) = \phi_{\max}^{-1}(\Sigma_0^k) \geq d_0$ for all i, k. Thus, (10) implies that we can choose $\lambda_{ij}^g = \lambda_{\max}$ as

$$\lambda_{\max} = \frac{2}{\sqrt{nd_0}} \left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right), \tag{14}$$

with $q > 1$ for all 3-tuples (i, j, g). Then we can rewrite the oracle inequalities in (12) and (13) as

$$\mathcal{M}(\hat{\Theta}_i) \le \frac{64\phi_{\max}}{\kappa^2} s_i, \tag{15}$$

$$\|\hat{\Theta}_i - \Theta_{0,i}\|_F \le \frac{8\sqrt{10}}{\kappa^2(2s_0)\sqrt{d_0}} \left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}}\sqrt{q \log G_0} \right) \sqrt{\frac{s_i}{n}}. \tag{16}$$

Detailed proof of Proposition A.1 follows similarly to that of Theorem 3.1 in Lounici et al. (2011) and can be found in Appendix E.

A.2 Selecting Edge Set

Given the estimates $\hat{\Theta}_i$ $(i = 1, \ldots, p)$, define $\hat{E}^k$ as in (3) the estimated set of edges in graph $k = 1, \ldots, K$. For every k, let $\widetilde{\Omega}^k = \operatorname{diag}(\Omega_0^k) + \Omega^k_{0, E_0^k \cap \hat{E}^k}$ and $\widetilde{\Sigma}^k = (\widetilde{\Omega}^k)^{-1}$. Let

$$C_{\text{bias}} = \frac{8\sqrt{10}c_0}{\kappa^2(2s_0)\sqrt{d_0}}.$$

The following corollary is an immediate result of (15) and (16).

Corollary A.1 *Consider $\hat{E}^k$ $(k = 1, \ldots, K)$ selected in* (3). *Suppose all conditions in Theorem 1 are satisfied. Choose $\lambda_{ij}^g = \lambda_{\max}$ as defined in* (14) *with $q > 1$. Then we have on the event $\mathcal{A}$*

$$|\hat{E}^k| \le \frac{64\phi_{\max}}{\kappa^2(s_0)} S_0, \quad k = 1, \ldots, K, \tag{17}$$

and

$$\frac{1}{K}\sum_k \|\widetilde{\Omega}^k - \Omega_0^k\|_F \le \frac{1}{\sqrt{K}}\Big\{\sum_k \|\widetilde{\Omega}^k - \Omega_0^k\|_F^2\Big\}^{1/2} \le C_{\text{bias}}\sqrt{\frac{S_0}{nK}}\left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}}\sqrt{q\log G_0}\right), \tag{18}$$

where G_0 is the maximum number of groups in all p regressions, S_0 is the total number of relevant groups, and $|g_{\max}|$ is the maximum group size.

The bound in (17) says that the cardinality of the estimated set of edges is at most of the order of S_0 and proves essential in controlling the error rate of the maximum likelihood estimate $\hat{\Omega}^k$ in the refitting step. Further, the second inequality in (18) implies

$$\Big\{\sum_k \|\widetilde{\Omega}^k - \Omega_0^k\|_F^2\Big\}^{1/2} \le \tau_1 d_0,$$

provided the sample size n satisfies for $0 < \tau_1 < 1$,

$$n \ge S_0 \left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}}\sqrt{q\log G_0}\right)^2 \left(\frac{C_{\text{bias}}}{\tau_1 d_0}\right)^2.$$

It follows immediately that on the event $\mathcal{A}$, we can bound the spectrum of $\widetilde{\Omega}^k$ $(k = 1, \ldots, K)$ as follows. For a symmetric matrix A, let $\|A\|$ represent the spectral norm of A, which is equal to $\phi_{\max}(A)$. By definition,

$$\phi_{\min}(\widetilde{\Omega}^k) = \min_{v: v^T v = 1} v^T \widetilde{\Omega}^k v = \min_{v: v^T v = 1} \{v^T \Omega_0^k v + v^T(\widetilde{\Omega}^k - \Omega_0^k)v\} \ge \phi_{\min}(\Omega_0^k) - \|\widetilde{\Omega}^k - \Omega_0^k\|.$$

Since $\phi_{\min}(\Omega_0^k) \geq d_0$ by Assumption (A2), we have

$$\begin{aligned}\phi_{\min}(\widetilde{\Omega}^k) &\geq \phi_{\min}(\Omega_0^k) - \|\widetilde{\Omega}^k - \Omega_0^k\| \geq \phi_{\min}(\Omega_0^k) - \|\widetilde{\Omega}^k - \Omega_0^k\|_F \\ &\geq \phi_{\min}(\Omega_0^k) - \Big\{\sum_k \|\widetilde{\Omega}^k - \Omega_0^k\|_F^2\Big\}^{1/2} \geq (1-\tau_1)d_0 > 0, \end{aligned} \tag{19}$$

In addition, we have an upper bound for the maximum eigenvalue of $\widetilde{\Omega}^k$,

$$\begin{aligned}\phi_{\max}(\widetilde{\Omega}^k) &\leq \phi_{\max}(\Omega_0^k) + \|\widetilde{\Omega}^k - \Omega_0^k\| \leq \phi_{\max}(\Omega_0^k) + \|\widetilde{\Omega}^k - \Omega_0^k\|_F \\ &\leq \phi_{\max}(\Omega_0^k) + \Big\{\sum_k \|\widetilde{\Omega}^k - \Omega_0^k\|_F^2\Big\}^{1/2} \leq c_0 + \tau_1 d_0 < \infty. \end{aligned} \tag{20}$$

A.3 Refitting

Let $\hat{\Omega}^k$ $(k = 1, \ldots, K)$ be defined in (4) and

$$r_n = C_{\text{bias}}\sqrt{\frac{S_0}{n}}\left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}}\sqrt{q \log G_0}\right). \tag{21}$$

Proof [of Theorem 1.] In view of Corollary A.1, it suffices to show that

$$\sum_k \|\hat{\Omega}^k - \widetilde{\Omega}^k\|_F^2 \leq O\left(r_n^2\right),$$

since by Cauchy-Schwarz inequality,

$$\frac{1}{K}\sum_k \|\hat{\Omega}^k - \widetilde{\Omega}^k\|_F \leq \frac{1}{\sqrt{K}}\Big\{\sum_k \|\hat{\Omega}^k - \widetilde{\Omega}^k\|_F^2\Big\}^{1/2},$$

and by triangle inequality,

$$\frac{1}{K}\sum_k \|\hat{\Omega}^k - \Omega_0^k\|_F \leq \frac{1}{K}\sum_k \|\hat{\Omega}^k - \widetilde{\Omega}^k\|_F + \frac{1}{K}\sum_k \|\widetilde{\Omega}^k - \Omega_0^k\|_F.$$

For $k = 1, \ldots, K$, let $\Delta^k = \Omega^k - \widetilde{\Omega}^k \in \mathbb{M}(p,p)$ and $\hat{\Delta}^k = \hat{\Omega}^k - \widetilde{\Omega}^k$. Let

$$Q(\Omega) = \sum_k \Big\{\operatorname{tr}(\hat{\Sigma}^k \Omega^k) - \log\det(\Omega^k) - \operatorname{tr}(\hat{\Sigma}^k \widetilde{\Omega}^k) + \log\det(\widetilde{\Omega}^k)\Big\}.$$

Since $(\hat{\Omega}^k)_{k=1}^K$ minimizes $Q(\Omega)$, $(\hat{\Delta}^k)_{k=1}^K$ minimizes $G(\Delta) = Q(\widetilde{\Omega} + \Delta)$. Recall the definition $\mathcal{S}_E^+ = \{\Gamma \in \mathbb{R}^{p\times p} : \Gamma \succ 0 \text{ and } \Gamma_{ij} = 0, \text{ for all } (i,j) \notin E \text{ where } i \neq j\}$. For $k = 1, \ldots, K$, define a sequence of convex sets

$$\mathcal{U}_n(\widetilde{\Omega}^k) = \{\Gamma - \widetilde{\Omega}^k | \Gamma \in \mathcal{S}_{\hat{E}^k}^+\}.$$

The main idea of the proof is as follows. For a sufficiently large $M > 0$, consider the set

$$\mathcal{T}_n = \{(\Delta^1, \ldots, \Delta^K) : \Delta^k \in \mathcal{U}_n(\widetilde{\Omega}^k), \sum_k \|\Delta^k\|_F^2 = M r_n^2\}.$$

Write $\underline{0}_{p\times p}$ the zero matrix in $\mathbb{M}(p,p)$. It is clear that $G(\Delta)$ is a convex function and $G(\hat{\Delta}) \leq G(\underline{0}_{p\times p}) = 0$. Thus if we can show $\inf_{\Delta\in\mathcal{T}_n} G(\Delta) > 0$, the minimizer $\hat{\Delta}$ must be inside the ball defined by $\mathcal{T}_n$. That is $\sum_k \|\hat{\Delta}^k\|_F^2 \leq Mr_n^2$. To see this, note that the convexity of $Q(\Omega)$ implies that $\inf_{\Delta\in\mathcal{T}_n} Q(\widetilde{\Omega}+\Delta) > Q(\widetilde{\Omega}) = 0$. There exists therefore a local minimizer in the ball $\{\widetilde{\Omega}^k + \Delta^k : \sum_k \|\Delta^k\|_F^2 \leq Mr_n^2\}$, or equivalently, $\sum_k \|\hat{\Delta}^k\|_F^2 \leq Mr_n^2$.

In the remainder of the proof, we focus on

$$G(\Delta) = \sum_k \Big\{ \mathrm{tr}(\hat{\Sigma}^k \Delta^k) - \log\det(\widetilde{\Omega}^k + \Delta^k) + \log\det(\widetilde{\Omega}^k) \Big\}.$$

Applying Taylor expansion to the logarithm terms in the above equation, we have

$$\begin{aligned}
&\log\det(\widetilde{\Omega}^k + \Delta^k) - \log\det(\widetilde{\Omega}^k) \\
=&\mathrm{tr}(\widetilde{\Sigma}^k\Delta^k) - \mathrm{vec}(\Delta^k)^T \left\{ \int_0^1 (1-t)(\widetilde{\Omega}^k + t\Delta^k)^{-1} \otimes (\widetilde{\Omega}^k + t\Delta^k)^{-1} dt \right\} \mathrm{vec}(\Delta^k),
\end{aligned}$$

where $\otimes$ is the Kronecker product, and $\mathrm{vec}(\Delta^k)$ is Δ^k vectorized to match the dimensions of the Kronecker product. Therefore, we can rewrite $G(\Delta) = L_1 - L_2 + L_3$, with

$$\begin{aligned}
L_1 &= \sum_k \mathrm{tr}\big\{(\hat{\Sigma}^k - \Sigma_0^k)\Delta^k\big\}, \\
L_2 &= \sum_k \mathrm{tr}\big\{(\widetilde{\Sigma}^k - \Sigma_0^k)\Delta^k\big\}, \\
L_3 &= \sum_k \mathrm{vec}(\Delta^k)^T \left\{ \int_0^1 (1-t)(\widetilde{\Omega}^k + t\Delta^k)^{-1} \otimes (\widetilde{\Omega}^k + t\Delta^k)^{-1} dt \right\} \mathrm{vec}(\Delta^k).
\end{aligned}$$

Next we bound each term separately.

Recall for every k, Σ_0^k and $\hat{\Sigma}^k$ represent the correlation and the sample correlation matrix, respectively. By Lemma 14 of Zhou et al. (2011) [see details on page 3003],

$$\mathbb{P}\Big\{ |\hat{\sigma}_{ij}^k - \sigma_{0,ij}^k| \geq t \Big\} \leq \exp\Big(- \frac{3nt^2}{10\{1 + (\sigma_{0,ij}^k)^2\}} \Big) \leq \exp\Big(- \frac{3nt^2}{20} \Big), \tag{22}$$

for $0 \leq t \leq \{1 + (\sigma_{0,ij}^k)^2\}/2$. Then the union sum inequality and (22) imply that, with probability tending to 1,

$$\max_{k,i\neq j} |\hat{\sigma}_{ij}^k - \sigma_{0,ij}^k| \leq c_1 \sqrt{\frac{1}{nK}} \left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}} \sqrt{q\log G_0} \right),$$

provided that the sample size satisfies

$$n \geq \frac{4c_1^2}{K} \left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}} \sqrt{q\log G_0} \right)^2,$$

where $c_1 > 0$ is a constant. Write $\Delta^k = \Delta^{k,+} + \Delta^{k,-}$ such that $\Delta^{k,+} = \mathrm{diag}(\Delta^k)$ and $\Delta^{k,-}$ consists of the off-diagonal entries of Δ^k. Then

$$|L_1| \leq \sum_k \sum_{i\neq j} |\hat{\sigma}_{ij}^k - \sigma_{0,ij}^k||\Delta_{ij}^k| \leq c_1 \sqrt{\frac{1}{nK}} \left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}} \sqrt{q\log G_0} \right) \sum_k \|\Delta^{k,-}\|_1.$$

By Cauchy-Schwarz inequality and the definition of $\Delta^k \in \mathcal{U}_n(\widetilde{\Omega}^k)$,

$$\sum_k \|\Delta^{k,-}\|_1 \leq \sum_k (2|\hat{E}^k|)^{1/2} \|\Delta^{k,-}\|_F \leq \max_k (2|\hat{E}^k|)^{1/2} \sqrt{K} \Big(\sum_k \|\Delta^k\|_F^2 \Big)^{1/2}.$$

Using the bound of $\hat{E}^k$ in (17) and the definition of r_n, we obtain

$$\begin{aligned} |L_1| &\leq c_1 \sqrt{\frac{1}{n}} \left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right) \frac{8\sqrt{2\phi_{\max} S_0}}{\kappa(s_0)} \Big(\sum_k \|\Delta^k\|_F^2 \Big)^{1/2} \\ &= \frac{8\sqrt{2} c_1 \sqrt{\phi_{\max}}}{C_{\text{bias}} \kappa(s_0)} r_n \Big(M r_n^2 \Big)^{1/2} = \frac{8\sqrt{2} c_1 \sqrt{\phi_{\max}}}{C_{\text{bias}} \kappa(s_0)} \sqrt{M} r_n^2, \end{aligned} \tag{23}$$

where the first equality in (23) follows from the definition of r_n in (21).

Using results from (19) and (18) together with Cauchy-Schwarz inequality, the second term L_2 can be bounded by

$$\begin{aligned} |L_2| &\leq \sum_k |\langle \widetilde{\Sigma}^k - \Sigma_0^k, \Delta^k \rangle| \leq \sum_k \|\widetilde{\Sigma}^k - \Sigma_0^k\|_F \|\Delta^k\|_F \leq \sum_k \|\Delta^k\|_F \frac{\|\widetilde{\Omega}^k - \Omega_0^k\|_F}{\phi_{\min}(\widetilde{\Omega}^k) \phi_{\min}(\Omega_0^k)} \\ &\leq \frac{1}{(1-\tau_1) {d_0}^2} \Big(\sum_k \|\Delta^k\|_F^2 \Big)^{1/2} \Big(\sum_k \|\widetilde{\Omega}^k - \Omega_0^k\|_F^2 \Big)^{1/2} \leq \frac{\sqrt{M} r_n^2}{(1-\tau_1) {d_0}^2}, \end{aligned} \tag{24}$$

where the last inequality in (24) comes from the rotation invariant property of the Frobenius norm.

Finally we bound L_3. Suppose for a small constant $0 < \tau_2 < 1$ such that $\tau_1 + \tau_2 < 1$, the sample size n satisfies

$$n \geq M S_0 \left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} \right)^2 \left(\frac{C_{\text{bias}}}{\tau_2 d_0} \right)^2,$$

then $\sqrt{M} r_n \leq \tau_2 d_0$. By (20), $\phi_{\max}(\widetilde{\Omega}^k)$ is bounded above by $c_0 + \tau_1 d_0$. Therefore for $\Delta \in \mathcal{T}_n$,

$$\begin{aligned} \phi_{\max}(\widetilde{\Omega}^k + \Delta^k) &\leq c_0 + \tau_1 d_0 + \|\Delta^k\| \leq c_0 + \tau_1 d_0 + \|\Delta^k\|_F \\ &\leq c_0 + \tau_1 d_0 + \Big(\sum_k \|\Delta^k\|_F^2 \Big)^{1/2} \leq c_0 + (\tau_1 + \tau_2) d_0, \\ \phi_{\min}(\widetilde{\Omega}^k + \Delta^k) &\geq (1-\tau_1) d_0 - \|\Delta^k\| \geq (1-\tau_1) d_0 - \|\Delta^k\|_F \\ &\geq (1-\tau_1) d_0 - \Big(\sum_k \|\Delta^k\|_F^2 \Big)^{1/2} \geq (1 - \tau_1 - \tau_2) d_0 > 0. \end{aligned}$$

For $\widetilde{\Omega}^k$ and Δ^k defined above, Zhou et al. (2011) showed that $\widetilde{\Omega}^k + t\Delta^k \succ 0, t \in [0,1]$, for all $k = 1, \ldots, K$ on the event $\mathcal{A}$. Thus, following similar arguments as in Rothman et al. (2008, page 502), we have

$$\begin{aligned} |L_3| &\geq \frac{1}{2} \sum_k \phi_{\min}^2 (\widetilde{\Omega}^k + \Delta^k)^{-1} \|\Delta^k\|_F^2 = \frac{1}{2} \sum_k \phi_{\max}^{-2} (\widetilde{\Omega}^k + \Delta^k) \|\Delta^k\|_F^2 \\ &\geq \frac{M r_n^2}{2(c_0 + \tau_1 d_0 + \tau_2 d_0)^2}. \end{aligned}$$

Combining the above three bounds, we thus have

$$\begin{aligned} G(\Delta) &\geq |L_3| - |L_1| - |L_2| \\ &\geq \frac{Mr_n^2}{2(c_0+\tau_1 d_0+\tau_2 d_0)^2} - \frac{8\sqrt{2}c_1\sqrt{\phi_{\max}}}{C_{\text{bias}}\kappa(s_0)}\sqrt{M}r_n^2 - \frac{\sqrt{M}r_n^2}{(1-\tau_1){d_0}^2} \\ &\geq Mr_n^2\left\{\frac{1}{2(c_0+\tau_1 d_0+\tau_2 d_0)^2} - \frac{8c_1\sqrt{2\phi_{\max}}}{C_{\text{bias}}\kappa(s_0)}\frac{1}{\sqrt{M}} - \frac{1}{(1-\tau_1){d_0}^2\sqrt{M}}\right\} > 0, \end{aligned}$$

for M sufficiently large. ■

Appendix B. Proof of Theorem 2

Consider the group lasso estimator $\hat{\Theta}_i$ defined in (2). Since the problem (2) is a special case of the generic group lasso in Basu et al. (2015), we adapt their results in Theorem 4.1 to our design.

Proof Let $\mathcal{X}_i$ be the block diagonal matrix composed of all variables but $\mathbf{X}_i^k$ $(k = 1, \ldots, K)$, that is

$$\mathcal{X}_i = \begin{pmatrix} \mathbf{X}_{-i}^1 & & \\ & \ddots & \\ & & \mathbf{X}_{-i}^K \end{pmatrix}.$$

After rearranging the columns of $\mathcal{X}_i$, we assume without loss of generality $\mathcal{X}_i = (\mathcal{X}_{i,(1)}, \mathcal{X}_{i,(2)})$ such that

$$\mathcal{X}_{i,(1)} = \text{diag}(\mathbf{X}_{I_1}^1, \ldots, \mathbf{X}_{I_K}^K)$$

is the sub-matrix consisting of all relevant variables. Denote the Gram matrix

$$C = \frac{1}{n}\mathcal{X}_i^T\mathcal{X}_i = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

with $C_{11} = \mathcal{X}_{i,(1)}^T\mathcal{X}_{i,(1)}/n$ and $C_{22} = \mathcal{X}_{i,(2)}^T\mathcal{X}_{i,(2)}/n$. C_{12} and C_{21} are also defined accordingly. Note due to the block diagonal structure of $\mathcal{X}_{i,(1)}$, C_{11} is also block diagonal.

Now consider interchanging the columns of $\mathcal{X}_i$ such that

$$\tilde{\mathcal{X}}_i = \mathcal{X}_i\text{diag}(R_1, R_2) = (\mathcal{X}_{i,(1)}R_1, \mathcal{X}_{i,(2)}R_2) = (\tilde{\mathcal{X}}_{i,(1)}, \tilde{\mathcal{X}}_{i,(2)}),$$

where the columns of $\tilde{\mathcal{X}}_{i,(1)}$ and $\tilde{\mathcal{X}}_{i,(2)}$ are ordered in groups of variables. Here R_l is the product of elementary column switching matrices and satisfies $R_l^{-1} = R_l^T$ $(l = 1, 2)$. Note $R_1 \in \mathbb{M}(\sum_k |I_k|, \sum_k |I_k|)$. Based on $\tilde{\mathcal{X}}_i$, we can define $\tilde{C}_{11}, \tilde{C}_{21}$ and $\tilde{C}_{22}$ similarly as above. The advantage of using $\tilde{\mathcal{X}}_i$ as the design matrix is that it orders the variables based on the grouping structures, and is in the form of the generic group lasso design in Basu et al. (2015). It is thus more straightforward to adapt their results using $\tilde{\mathcal{X}}_i$. Moreover, since each group of variables (j, g) corresponds to regression coefficients at the same (i, j) position across different models in g, the matrix $\tilde{C}_{11}$ is in fact a block matrix, whose diagonal blocks are all identity matrices. To see this, consider

$g = \{k_1, k_2\}$, the columns of $\tilde{\mathcal{X}}_{i,(1)}$ that correspond to the group (j, g) is

$$\mathbf{X}_j^g = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \vdots & \vdots \\ \mathbf{X}_j^{k_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_j^{k_2} \\ \vdots & \vdots \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Hence the (j, g)-th diagonal block $(\tilde{C}_{11})_{[j,g]} = (\mathbf{X}_j^g)^T \mathbf{X}_j^g / n = I_2$.

With the above notations, the Uniform IC in Assumption (A3) is equivalent to saying for all $\boldsymbol{\xi} = ((\boldsymbol{\xi}^1)^T, \dots, (\boldsymbol{\xi}^K)^T)^T \in \mathbb{R}^{\sum_k |I_k|}$ with $\max\limits_{(j,g)\in J(\Theta_{0,i})} \|\boldsymbol{\xi}_j^{[g]}\| \le 1$ and all $(j, g) \notin J(\Theta_{0,i})$

$$\| \left[\tilde{C}_{21}(\tilde{C}_{11})^{-1}\tilde{\boldsymbol{\xi}}\right]_{[j,g]} \| \le 1 - \eta,$$

where $\tilde{\boldsymbol{\xi}} = R_1^T \boldsymbol{\xi}$.

It remains to select λ and α_n to ensure that the direction consistency results hold simultaneously for all i with probability tending to 1. For any $(j, g) \in J(\Theta_{0,i})$, denote $(\tilde{C}_{11})^{-1}_{[j,g]}$ the diagonal block in $\tilde{C}_{11}^{-1}$ corresponding to the group (j, g). By Theorem 4.1 of Basu et al. (2015), it suffices to find the upper bounds for $\|\tilde{C}_{11}^{-1}\|, \|(\tilde{C}_{11})^{-1}_{[j,g]}\|, \|(\tilde{C}_{22})_{[j,g]}\|$ and substitute the constant variance σ with the appropriate bound for $\mathrm{Var}(X_i^k | X_{-i}^k) = 1/\omega_{0,ii}^k$ $(k = 1, \dots, K)$.

By definition and the fact that the columns of $\mathbf{X}^k$ are centered and standardized to have mean zero and unit variance, $(\tilde{C}_{11})_{[j,g]}$ is the identity matrix of size $|g| \times |g|$. It follows that

$$1 = \phi_{\min}^{-1}((\tilde{C}_{11})_{[j,g]}) \le \phi_{\max}((\tilde{C}_{11})^{-1}_{[j,g]}) = \|(\tilde{C}_{11})^{-1}_{[j,g]}\| \le \|(\tilde{C}_{11})^{-1}\|, \tag{25}$$

where the last step is obtained by applying Courant minimax principle since $0 \prec (\tilde{C}_{11})^{-1}_{[j,g]} \preceq (\tilde{C}_{11})^{-1}$. Similarly, for any $(j, g) \notin J(\Theta_{0,i})$, $(\tilde{C}_{22})_{[j,g]}$ is the identity matrix and

$$\|(\tilde{C}_{22})_{[j,g]}\| = 1. \tag{26}$$

Moreover, the variance for the random design in our problem

$$\mathrm{Var}(X_i^k | X_{-i}^k) = 1/\omega_{0,ii}^k \le 1/d_0, \ \forall\, k, \tag{27}$$

by Assumption (A2).

It remains to find an upper bound for $\|\tilde{C}_{11}^{-1}\|$. Under Assumption (A1) with $s = s_0$, if we set $\Delta \in \mathcal{F}$ such that $\boldsymbol{\delta}_j^{[g]} = \mathbf{0}$ for any $(j, g) \notin J(\Theta_{0,i})$, then

$$\frac{\sum_k \|\mathbf{X}^k \boldsymbol{\delta}^k\|^2 / n}{\|\Delta_{J(\Theta_{0,i})}\|_F^2} = \frac{\boldsymbol{\xi}^T C_{11} \boldsymbol{\xi}}{\boldsymbol{\xi}^T \boldsymbol{\xi}},$$

where $\boldsymbol{\xi} = ((\boldsymbol{\xi}^1)^T, \dots, (\boldsymbol{\xi}^K)^T)^T \in \mathbb{R}^{\sum_k |I_k|}$ such that each $\boldsymbol{\xi}^k$ corresponds to the nonzero part of $\boldsymbol{\delta}^k$. If we choose Δ such that $\boldsymbol{\xi}$ is the eigenvector corresponding to the smallest eigenvalue of C_{11}, then

$$\kappa^2(s_0) \leq \frac{\sum_k \|\mathbf{X}^k \boldsymbol{\delta}^k\|^2/n}{\|\Delta_{J(\Theta_{0,i})}\|_F^2} = \frac{\boldsymbol{\xi}^T C_{11} \boldsymbol{\xi}}{\boldsymbol{\xi}^T \boldsymbol{\xi}} = \phi_{\min}(C_{11}).$$

Since $R_1^{-1} = R_1^T$, C_{11} and $\tilde{C}_{11}$ are similar (there exists a non-singular matrix P such that $P^{-1} C_{11} P = \tilde{C}_{11}$) and thus share the same set of eigenvalues. Therefore $\phi_{\min}(\tilde{C}_{11}) \geq \kappa^2(s_0)$ and

$$\|\tilde{C}_{11}^{-1}\| \leq \kappa^{-2}(s_0). \tag{28}$$

Combining the upper bounds in (25), (26), (27) and (28), Theorem 4.1 of Basu et al. (2015) implies that if we select λ and α_n as in (8) and (9), respectively, the direction consistency results follow by considering the union bound on all probabilities made across $i = 1, \dots, p$.

Further, if $\alpha_n < 1$, the direction consistency property of $\hat{\Theta}_i$ implies exact recovery of all nonzero entries in the inverse covariance matrices, provided that the sparsity pattern $\mathscr{G}$ is correctly specified. In other words, the set in (3) estimates correctly the true edge set E_0^k for all k.

The probability statement $1 - 4pG_0^{1-q}$ follows from considering the union bound of the above result over all p regressions.

This completes the proof. ■

Appendix C. Additional Simulation Results

C.1 Performance with and without maximum likelihood refitting

In the main paper, we have compared the performance of different methods in estimating multiple Gaussian graphical models under optimally chosen tuning parameters with the results shown in Tables 1 and 2. All joint estimation methods were evaluated by adding the maximum likelihood refitting Step (II) for fair comparisons. To confirm that this is indeed the case, we present in the following additional simulation results for cases evaluated without the maximum likelihood refitting step.

Table 5 presents the complete table of deviance measures for various methods considered in simulation study 1. These methods include the separate estimation method Glasso, where the *Graphical lasso* by Friedman et al. (2008) is applied to each graphical model separately, joint estimation by Guo et al. (2011), denoted by JEM-G, the Group Graphical Lasso denoted by GGL by Danaher et al. (2014), and the structural pursuit method MGGM by Zhu et al. (2014). For the latter three methods, we also present deviance measures for which the maximum likelihood refitting Step (II) is not included, denoted respectively by JEM-G1, GGL1 and MGGM1. For the proposed two-step method JSEM, deviance measures based on Step I only is presented under the name JSEM1. It is clear from Table 5 that the refitting step generally does not introduce more errors in terms of structural estimation, but can significantly reduce the estimation errors in Frobenius norm.

Table 6 presents the performance of different regularization methods in estimating multiple inverse covariance matrices in simulation study 2. Here we observe similar pattern as that in Table 5, which confirms again that contribution from the maximum likelihood refitting step is mainly in reducing the Frobenius norm loss.

Method	FP	FN	SHD	F1	FL
Glasso	35(6)	81(2)	116(5)	0.32(0.02)	0.73(< 0.01)
JEM-G1	22(4)	40(4)	62(6)	0.69(0.03)	0.28(0.03)
JEM-G	22(4)	40(4)	62(6)	0.69(0.03)	0.28(0.02)
GGL1	18(7)	73(2)	91(7)	0.44(0.03)	0.70(0.01)
GGL	**17**(6)	73(2)	90(6)	0.44(0.03)	0.29(0.02)
MGGM1	291(14)	47(3)	339(14)	0.26(0.01)	0.69(0.02)
MGGM	286(13)	49(3)	335(13)	0.26(0.01)	0.64(0.02)
JSEM1	20(4)	**34**(3)	**54**(6)	**0.73**(0.03)	0.71(0.04)
JSEM	19(4)	35(3)	**54**(6)	**0.73**(0.03)	**0.25**(0.02)

Table 5: Performance of different regularization methods for estimating graphical models in Simulation Study 1: average FP, FN, SHD, F1 and FL (SE) for sample size $n_k = 50$. JEM-G1, GGL1, MGGM1 and JSEM1 correspond to respective method without the maximum likelihood refitting step. The best cases are highlighted in bold.

C.2 Performance as a function of p and n

Table 7 presents the performance of JSEM for $p = 500$ and $p = 1000$ with sample sizes n varying from 100, 200 to 500. The simulation setup is similar to that in Simulation Study 1: at each p, there are $K = 5$ graphical models sharing a *single* common structure. Individual structures with $\rho = 0.1$ are added to each graph separately such that about 10% of the edges in each graph are unique to themselves. It is clear that as the sample size n increases, the performance of JSEM also improves with smaller structural hamming distances (SHD), higher F_1 score (F1) and smaller Frobenius norm loss (FL). In particular, the number of falsely rejected edges (FN) has decreased significantly.

Appendix D. Real Data Analysis

D.1 Climate data sources and pre-processing

The data we use in this study come from multiple sources and are collected under different resolutions for varying lengths of time periods. Specifically, the sources we consider include:

(1) CRU: Climate Research Unit provides monthly climatology data (`http://www.cru.uea.ac.uk/cru/data`) for 10 surface variables including mean temperature (TMP), diurnal temperature range (DTR), maximum and minimum temperature (TMX, TMN), precipitation (PRE), vapor pressure (VAP), cloud cover (CLD), rainday counts (WET), potential evapotranspiration (PET) and frost days (FRS) from 1901 to 2013 at the 0.5 degree latitude and longitude resolution. Note these high-resolution gridded data sets are constructed using not only directly observed data, but also derived and estimated values with well-known formulae wherever the observed data are not available (see details in Harris et al., 2014).

(2) NASA: The Goddard Earth Sciences Data and Information Services Center (GES DISC) from the National Aeronautics and Space Administration (NASA) has collected aerosol measurements using Moderate Resolution Imaging Spectroradiometer (MODIS) on satellites. The

ρ	Method	FP	FN	SHD	F1	FL
0	Glasso	154(4)	38(1)	192(4)	0.51(0.01)	0.60(0.005)
	JEM-G1	87(2)	**36**(2)	123(3)	0.62(0.01)	0.41(0.01)
	JEM-G	86(3)	**36**(2)	122(3)	0.62(0.01)	0.31(0.01)
	GGL1	152(3)	38(1)	191(4)	0.51(0.01)	0.60(0.01)
	GGL	144(3)	39(1)	184(4)	0.52(0.01)	0.37(0.01)
	MGGM1	30(2)	67(1)	97(2)	0.59(0.01)	0.37(0.01)
	MGGM	30(2)	67(1)	97(2)	0.59(0.01)	0.36(0.01)
	JSEM1	22(2)	42(2)	64(3)	0.75(0.01)	0.68(0.01)
	JSEM	**21**(2)	42(2)	**63**(3)	**0.75**(0.01)	**0.28**(0.01)
0.2	Glasso	164(3)	**47**(1)	211(4)	0.53(0.01)	0.59(0.005)
	JEM-G1	94(3)	57(2)	151(3)	0.59(0.01)	0.44(0.01)
	JEM-G	92(3)	57(2)	149(3)	0.59(0.01)	0.35(0.01)
	GGL1	163(3)	**47**(1)	210(4)	0.53(0.01)	0.59(0.005)
	GGL	155(3)	48(1)	203(3)	0.53(0.01)	0.37(0.01)
	MGGM1	98(4)	63(1)	161(4)	0.56(0.01)	0.38(0.01)
	MGGM	94(3)	64(1)	158(4)	0.56(0.01)	0.37(0.01)
	JSEM1	33(3)	64(2)	97(3)	0.67(0.01)	0.77(0.01)
	JSEM	**32**(3)	64(2)	**96**(3)	**0.67**(0.01)	**0.32**(0.01)
0.4	Glasso	159(3)	**59**(1)	218(4)	0.55(0.01)	0.57(0.005)
	JEM-G1	101(3)	77(2)	178(3)	0.56(0.01)	0.45(0.01)
	JEM-G	100(3)	77(2)	177(3)	0.56(0.01)	0.37(0.01)
	GGL1	158(3)	60(1)	218(4)	0.55(0.01)	0.57(0.005)
	GGL	149(3)	61(2)	210(4)	0.55(0.01)	0.37(0.01)
	MGGM1	122(3)	65(1)	187(3)	0.57(0.01)	0.38(0.01)
	MGGM	119(3)	65(1)	184(3)	0.58(0.01)	0.37(0.01)
	JSEM1	50(3)	83(2)	133(3)	**0.62**(0.01)	0.84(0.01)
	JSEM	**49**(3)	84(2)	**132**(3)	**0.62**(0.01)	**0.36**(0.01)
0.6	Glasso	176(4)	**73**(2)	249(4)	0.54(0.01)	0.55(0.01)
	JEM-G1	95(3)	109(2)	204(4)	0.52(0.01)	0.45(0.01)
	JEM-G	94(3)	109(2)	203(3)	0.52(0.01)	0.39(0.01)
	GGL1	174(4)	74(2)	248(4)	0.54(0.01)	0.55(0.01)
	GGL	165(4)	76(2)	241(4)	0.54(0.01)	0.39(0.01)
	MGGM1	113(3)	94(2)	207(4)	**0.55**(0.01)	0.39(0.01)
	MGGM	109(3)	95(2)	204(4)	**0.55**(0.01)	0.39(0.01)
	JSEM1	52(3)	122(2)	174(4)	0.53(0.01)	0.89(0.01)
	JSEM	**50**(3)	123(2)	**173**(4)	0.52(0.01)	**0.38**(0.01)

Table 6: Performance of different regularization methods for estimating graphical models in Simulation Study 2: average FP, FN, SHD, F1 and FL (SE) for sample size $n_k = 100$. JEM-G1, GGL1, MGGM1 and JSEM1 correspond to respective method without the maximum likelihood refitting step. The best cases are highlighted in bold.

p	n	FP	FN	SHD	F1	FL
500	100	20(4)	179(8)	200(9)	0.79(0.01)	0.21(0.01)
	200	40(5)	50(3)	90(6)	0.92(0.005)	0.14(0.01)
	500	49(3)	36(1)	85(3)	0.92(0.002)	0.09(0.003)
1000	100	17(4)	643(14)	661(15)	0.58(0.01)	0.23(0.005)
	200	34(5)	187(9)	221(10)	0.89(0.005)	0.14(0.005)
	500	77(6)	80(2)	157(5)	0.93(0.002)	0.10(0.003)

Table 7: Performance of JSEM as a function of p and n: average FP, FN, SHD, F1 and FL (SE). The setup is similar to that in Simulation Study 1.

data set obtained from Terra satellite consists of monthly average aerosol optical depth (AER) at the 1 degree latitude by 1 degree longitude resolution from March 2000 to August 2014.

(3) NCDC: The National Solar Radiation Database (NSRDB) 1991-2010 (a collaborative project between The National Renewable Energy Laboratory (NREL) and the National Climatic Data Center (NCDC)) provides statistical summaries for solar data (ftp://ftp.ncdc.noaa.gov/pub/data/nsrdb-solar/) from 860 different locations across the United States. The locations are recorded using their latitude, longitude and altitude. We used measurements for global horizontal radiation (SOL) at 242 class I stations that have high-quality data.

(4) NOAA: The climate data center of National Oceanic and Atmospheric Administration (NOAA) has archived the trace gases data, including carbon dioxide (CO2), carbon monoxide (CO), methane (CH4) and hydrogen (H2), from 170 worldwide stations (http://www.esrl.noaa.gov/gmd/dv/ftpdata.html). These data sets consist of measurements spanning different time periods, with CO2 ranging from 1968 to 2013 (the longest) and H2 from 1992 to 2005 (the shortest). In addition, they come with relatively low resolution compared to other variables due to the limited number of stations.

To ensure compatibility and consistency among multiple data sources, we performed the following pre-processing:

(1) Normalization: We first transformed each data set into monthly observations in a standard format including longitude, latitude, altitude (when available), date, variable, value, unit, and source. We focus on a 54-month time period from January 2001 to June 2005 where data for all variables are available.

(2) Interpolation and smoothing: We interpolated the monthly data from NCDC and NOAA onto a common 2.5 by 2.5 degree grid for North America using thin plate splines. Since the data from CRU and NASA were provided for a finer resolution grid, thin plate splines were used to first interpolate the data onto a grid of the same resolution as the source data. Then we performed spatial averaging to get data on the common 2.5 by 2.5 degree grid.

(3) Seasonality and autocorrelation: We reduced the short-term autocorrelation by aggregating the time series for each variable at each location into bins of 3-month intervals and taking

first differences on the quarterly data. The resulting data, consisting of 17 measurements, are assumed to be independent samples for the corresponding variable at the specified location.

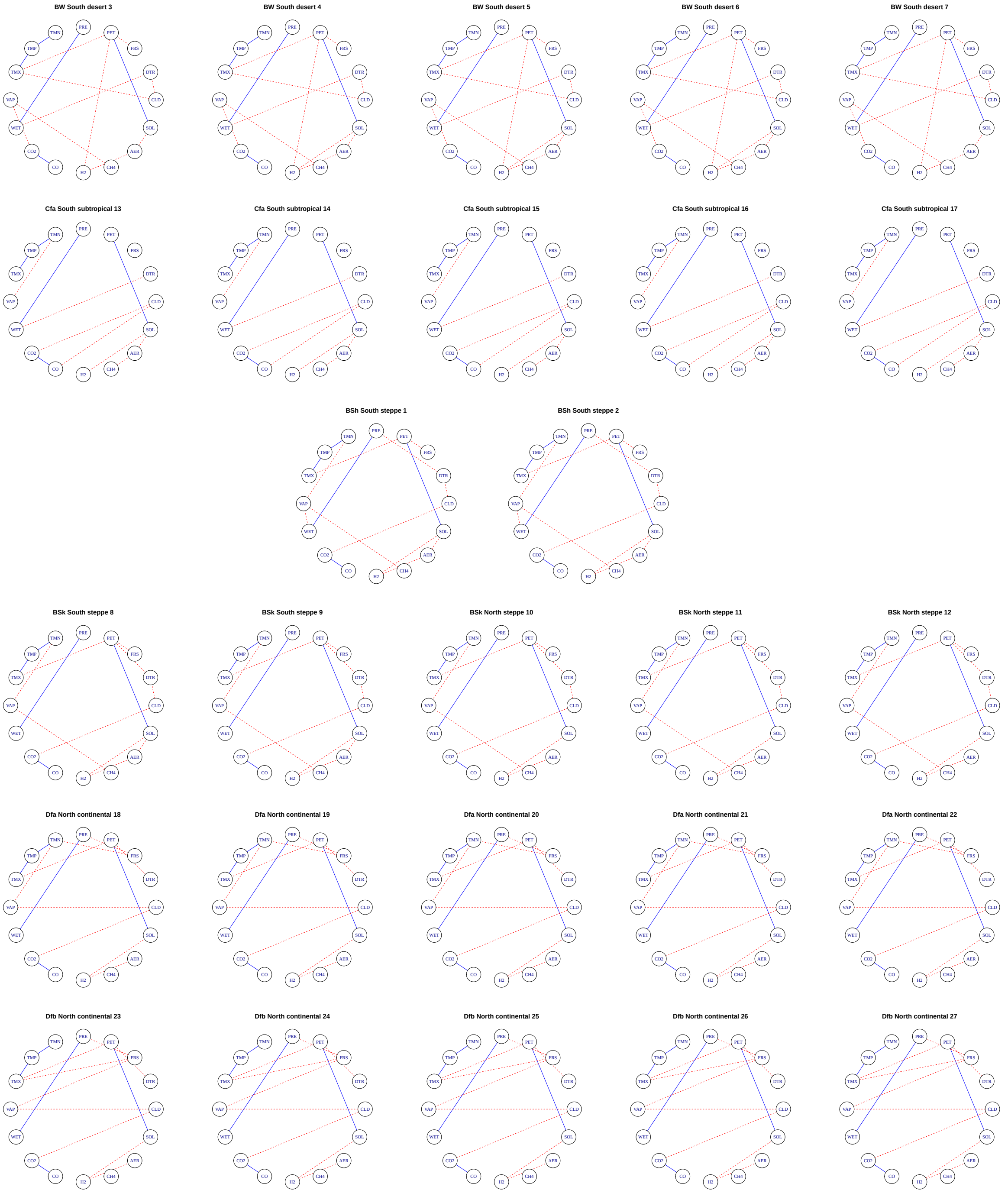
The final data are organized as an $n \times p$ matrix at each of the 27 locations considered, where $n = 17$ and $p = 16$.

D.2 Additional results in climate modeling

The inferred networks at the six distinct climate zones using JSEM are presented in Section 5. The estimated networks at the 27 distinct locations are also presented in Figure 8 for reference. For notational convenience, we have renamed the climate zones such that BW for Midlatitude Desert, Cfa for Humid Subtropical, Bsh for Semiarid Steppe (hot arid), Bsk for Semiarid Steppe (cold arid), Dfa for Humid Continental (hot summer), and Dfb for Humid Continental (warm summer).

The networks in Figure 8 are ordered such that those in the same row belong to the same climate zone; further, networks in the third and forth rows represent those from the Semiarid Steppe group, whereas networks in the last two rows are all from the Humid Continental group. Such an ordering respects how the structural information is defined and helps visualize similarities across networks. Indeed, by comparing networks at locations from geographical south, that is networks entitled 'South', we notice that the interactions between AER, SOL and the remaining variables are very similar. For example, almost all of them share the edges AER—SOL and SOL—H2, except at two locations 'BW South desert 3' and 'BW South desert 7'. In contrast, networks from the geographical north all share the edges AER—H2 and SOL—H2. Further, the interactions between variables on greenhouse gases (CO2, CO, CH4 and H2) and others have four distinct patterns at the four distinct climate groups. For example, greenhouse gases interact with VAP for the desert group, whereas they interact with CLD in the subtropical group. The partial correlation between CLD and greenhouse gases at subtropical climate makes sense because such humid areas are more likely to be cloudy, thereby influencing the concentration of CO2 (Graham et al., 2003). Finally, variables excluding AER, SOL and those on greenhouse gases show distinct interaction patterns with others at the six distinct climate zones. In particular, one can see that the variable FRS interacts mainly with PET at Desert and Steppe climate, whereas it is partially correlated with both PRE and TMN (or TMX) at Continental climate. This can be explained from the distinction between these climate zones. At Humid Continental climate, precipitation is relatively well distributed year-round in most areas and snowfall occurs in all areas. It is thus not difficult to see why precipitation (PRE) and temperature related variables correlate with the number of frost days (FRS). Further, a primary criterion of an area being Midlatitude Desert or Semiarid Steppe is that it receives precipitation below potential evapotranspiration (PET), which possibly explains why FRS is partially correlated only with PET for Desert and Steppe climate. We also point out that networks at adjacent climate zones are very similar. For example, networks at 'Bsh Steppe' and 'Bsk Steppe' share similar topologies.

As a comparison, we also applied other joint estimation methods JEM-G, GGL and MGGM on the same data set. For each of the three methods considered here, we used BIC on the normalized data to select the optimal tuning parameters and coupled each method with complementary pairs stability selection (Shah and Samworth, 2013) to infer the related climate networks. As in the case of JSEM, we run each method 50 times on two randomly drawn complementary pairs of size 8 and 9 and kept only edges that are selected above a certain threshold. The selection probability used for JSEM is 70%. However, as the second simulation study indicates JEM-G and GGL tend to produce higher false positives, especially GGL, we increased the probability threshold for JEM-G and GGL



BW South desert 3

BW South desert 4

BW South desert 5

BW South desert 6

BW South desert 7

Cfa South subtropical 13

Cfa South subtropical 14

Cfa South subtropical 15

Cfa South subtropical 16

Cfa South subtropical 17

BSh South steppe 1

BSh South steppe 2

BSk South steppe 8

BSk South steppe 9

BSk North steppe 10

BSk North steppe 11

BSk North steppe 12

Dfa North continental 18

Dfa North continental 19

Dfa North continental 20

Dfa North continental 21

Dfa North continental 22

Dfb North continental 23

Dfb North continental 24

Dfb North continental 25

Dfb North continental 26

Dfb North continental 27

Figure 8: Estimated climate networks at the 27 locations using JSEM, with edges shared across all locations solid and differential edges dashed.

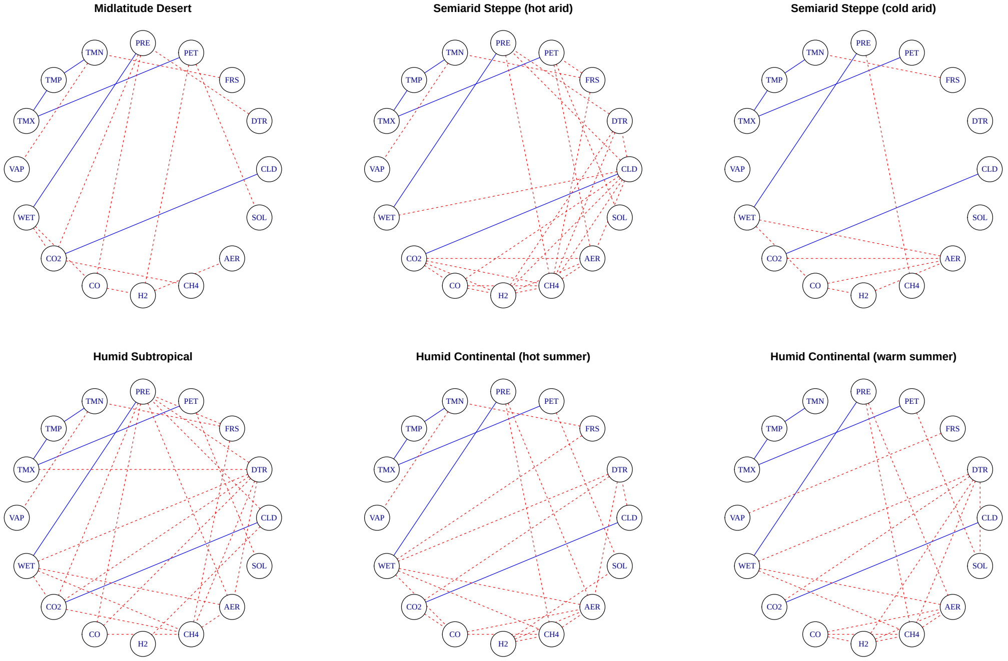

Midlatitude Desert
Semiarid Steppe (hot arid)
Semiarid Steppe (cold arid)
Humid Subtropical
Humid Continental (hot summer)
Humid Continental (warm summer)
PRE
PET
FRS
DTR
CLD
SOL
AER
CH4
H2
CO
CO2
WET
VAP
TMX
TMP
TMN

Figure 9: Estimated climate networks at the six distinct climate zones using JEM-G, with edges shared across all locations solid and differential edges dashed.

to 90% and 100%, respectively. On the other hand, we reduced the threshold for MGGM to 50% due to the relatively few edges recovered. The results are shown in Figure 9, 10 and 11.

One can see clearly that the estimated networks using the three methods exhibit quite different connectivity patterns from those inferred from JSEM. In particular, the results from GGL seem to suggest strong conditional dependence among a subset of variables, which distinguishes itself from JEM-G and JSEM. The estimated networks using MGGM, though sparse, bear certain similarity to those recovered using GGL. On the other hand, the results from JEM-G and JSEM are more similar. For example, common edges identified using JEM-G, such as TMN—TMP, TMP—TMX, PRE—WET, also show up under JSEM. The common edge between CLD and CO2 is found at all locations except Midlatitude Desert under JSEM, whereas the edge between PET and SOL identified using JSEM exists everywhere except at Semiarid Steppe (cold arid) under JEM-G. Note although JEM-G does not require external information on the structural relationships across graphs, the inferred networks respect roughly the spatial pattern of all climate zones. For instance, Humid Continental (hot summer and cool summer) are more similar.

D.3 Additional results in analysis of breast cancer

We have presented the inferred pathway level networks under both ER+ and ER- status using JSEM in the main paper. As a comparison, we applied JEM-G with tuning parameters selected via BIC to the same normalized and processed data set. To ensure a stable estimation, we further

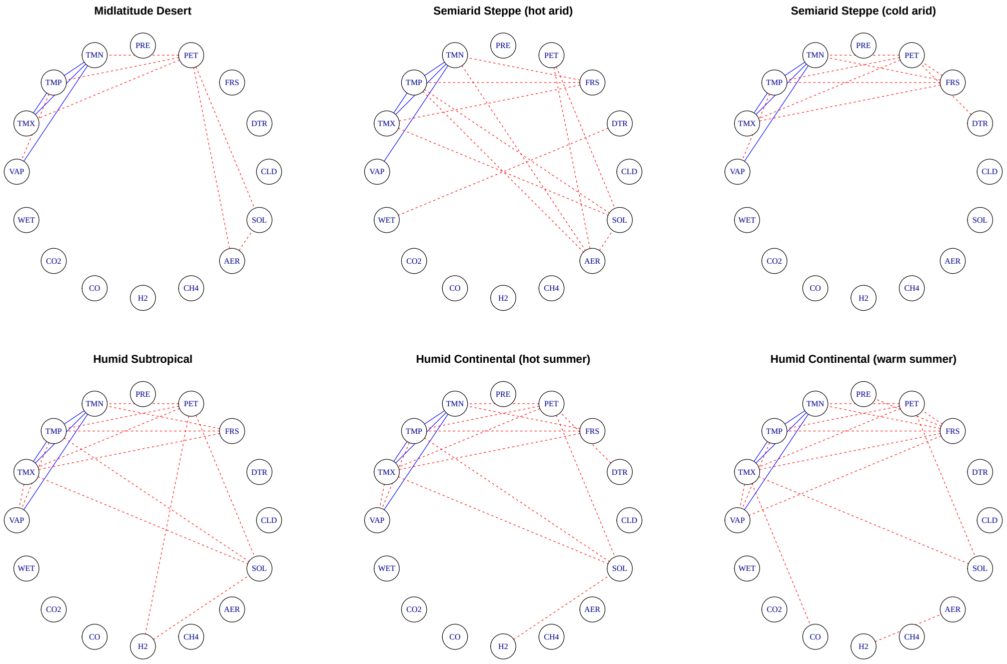

Midlatitude Desert
Semiarid Steppe (hot arid)
Semiarid Steppe (cold arid)
Humid Subtropical
Humid Continental (hot summer)
Humid Continental (warm summer)
PRE
PET
FRS
DTR
CLD
SOL
AER
CH4
H2
CO
CO2
WET
VAP
TMX
TMP
TMN

Figure 10: Estimated climate networks at the six distinct climate zones using GGL, with edges shared across all locations solid and differential edges dashed.

coupled JEM-G with complementary pairs stability selection (Shah and Samworth, 2013). Figure 12 shows the pathway level interactions estimated from JEM-G at selection frequency 70%, after removing isolated pathways. One striking difference between Figure 12 and Figure 7 is that Figure 12 sees more edges shared across the two classes (in blue). This is partly due to how JEM-G is implemented directly via the sample covariance matrices and partly to JEM-G being an agnostic method.

We also present estimated gene-level networks (with isolated genes removed) using both JEM-G and JSEM in Figure 13. Similar to what we observe in the pathway level network comparison, the JEM-G recovered gene networks show more edges shared between the two classes. In comparison, JSEM recovers more differential edges for the ER+ class. Apart from the differences, we also observe some similarities between the estimated gene networks. For example, both methods identify a small hub around the gene SFRP1, indicating their potential in regulating the underlying biological process.

Appendix E. Additional Technical Details

We include here some additional lemmas and proofs necessary for establishing the theoretical results in Section 3.

The first lemma is borrowed from Basu et al. (2015, Lemma A.2). We state the result here for completeness. Please refer to their paper for proof of the lemma.

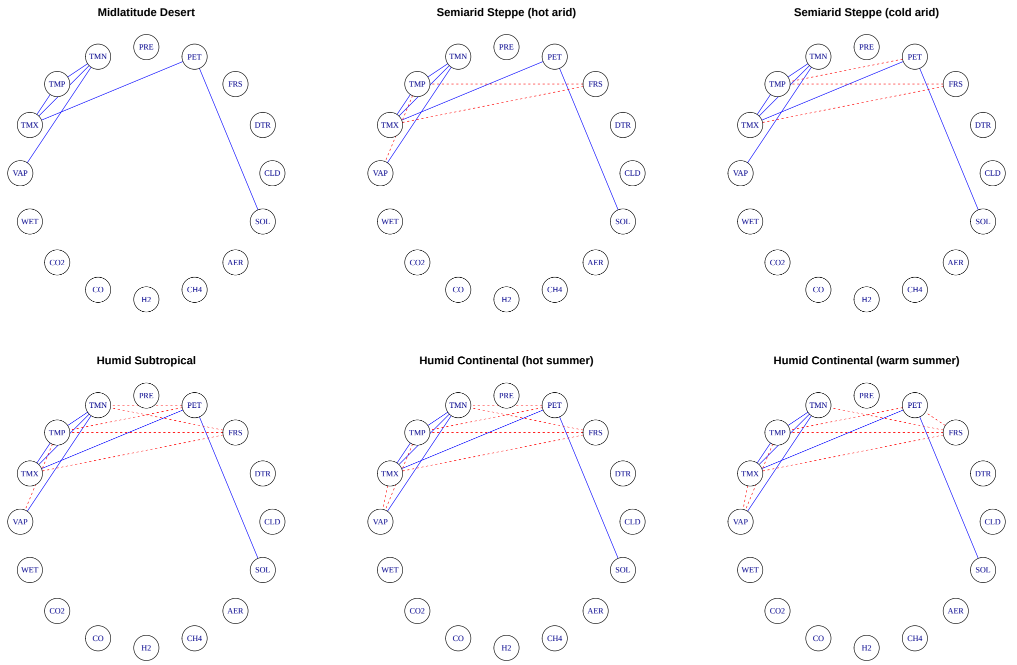

Midlatitude Desert
Semiarid Steppe (hot arid)
Semiarid Steppe (cold arid)
Humid Subtropical
Humid Continental (hot summer)
Humid Continental (warm summer)
PRE
PET
FRS
DTR
CLD
SOL
AER
CH4
H2
CO
CO2
WET
VAP
TMX
TMP
TMN

Figure 11: Estimated climate networks at the six distinct climate zones using MGGM, with edges shared across all locations solid and differential edges dashed.

Lemma E.1 *Let $Z_{k\times 1} \sim \mathcal{N}(\mathbf{0}, \Sigma)$. Then for any $t > 0$, the following inequalities hold:*

$$P\big(\big|\|Z\| - E\|Z\|\big| > t\big) \le 2\exp\left(-\frac{2t^2}{\pi^2\|\Sigma\|}\right), \quad E\|Z\| \le \sqrt{k}\sqrt{\|\Sigma\|}.$$

The next lemma provides a concentration bound for the random event $\mathcal{A}$ used in the proof of Theorem 1.

Lemma E.2 *Consider the random event $\mathcal{A} = \bigcap_{i,j\neq i,g} \mathcal{A}_{ij}^g$, where $\mathcal{A}_{ij}^g = \{2\|\boldsymbol{\zeta}_{ij}^{[g]}\| \le \lambda_{ij}^g\}$ and $\boldsymbol{\zeta}_{ij}^{[g]}$ is defined in Section A.1 of the Appendix. For each combination of $(i, j \neq i, g)$, choose*

$$\lambda_{ij}^g \ge \max_{k\in g} \frac{2}{\sqrt{n\omega_{0,ii}^k}}\left(\sqrt{|g|} + \frac{\pi}{\sqrt{2}}\sqrt{q\log G_0}\right). \tag{29}$$

where $q > 1$ and G_0 is the maximum number of groups in all regressions. Then

$$\mathbb{P}(\mathcal{A}) \ge 1 - 2pG_0^{1-q}.$$

Proof By Bonferroni inequality, $\mathbb{P}(\mathcal{A}^c) \le \sum_{i,j\neq i,g} \mathbb{P}(\{\mathcal{A}_{ij}^g\}^c)$. For any 3-tuple of $(i, j \neq i, g)$, it suffices to find an upper bound for $\mathbb{P}(\{\mathcal{A}_{ij}^g\}^c)$. Denote $\boldsymbol{\Psi}_j^k = (\mathbf{X}_j^k)^T\mathbf{X}_j^k/n$ and $\boldsymbol{\Phi}_j^k = \mathbf{X}_j^k(\mathbf{X}_j^k)^T/n$,

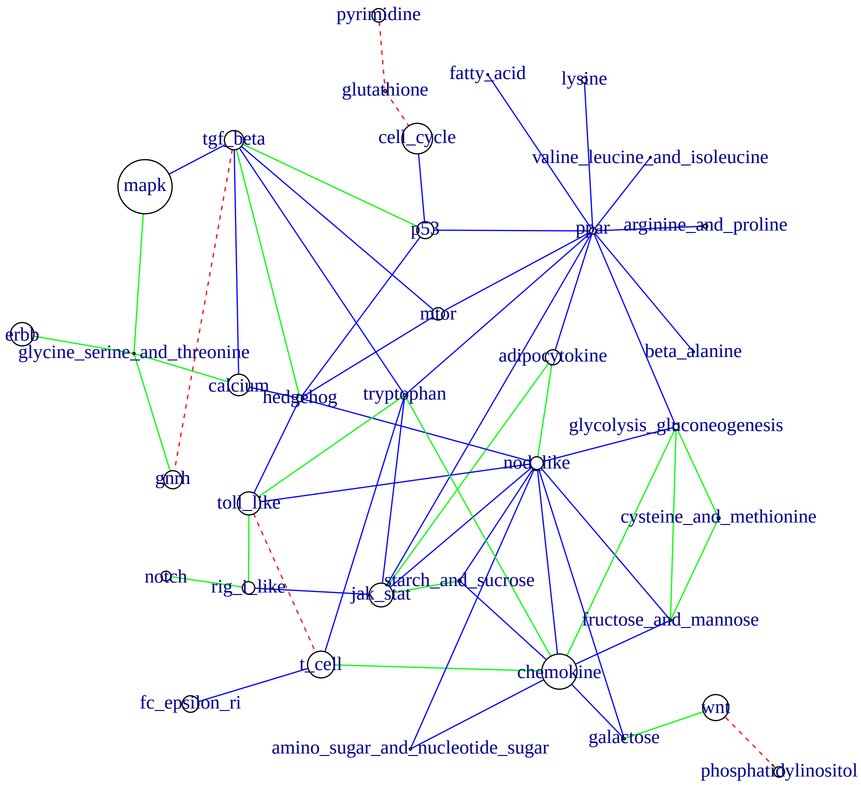


Figure 12: Estimated pathway networks for the ER+ and ER- classes using JEM-G, with edges shared blue solid and differential edges red dashed (ER+) / green dashed (ER-).

both of rank 1. The eigendecomposition of $\mathbf{\Phi}_j^k$ is $\mathbf{\Phi}_j^k = \mathbf{Q}^k\mathbf{V}^k(\mathbf{Q}^k)^T$, where $\mathbf{Q}^k$ is the orthogonal matrix whose columns are the eigenvectors of $\mathbf{\Phi}_j^k$ and $\mathbf{V}^k$ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues. It is clear that the only non-zero eigenvalue of $\mathbf{\Phi}_j^k$ is given by $\gamma_j^k = \|\mathbf{X}_j^k\|^2/n = 1$. Let Q_1^k be the eigenvector corresponding to γ_j^k. Therefore

$$\|\boldsymbol{\zeta}_{ij}^{[g]}\|^2 = \sum_{k\in g}\left(\zeta_{ij}^k\right)^2 = \sum_{k\in g}\frac{1}{n^2}(\boldsymbol{\varepsilon}_i^k)^T\mathbf{X}_j^k(\mathbf{X}_j^k)^T\boldsymbol{\varepsilon}_i^k = \frac{1}{n}\sum_{k\in g}(\boldsymbol{\varepsilon}_i^k)^T\mathbf{Q}^k\mathbf{V}^k(\mathbf{Q}^k)^T\boldsymbol{\varepsilon}_i^k,$$
$$= \frac{1}{n}\sum_{k\in g}(\boldsymbol{\varepsilon}_i^k)^T Q_1^k\gamma_j^k(Q_1^k)^T\boldsymbol{\varepsilon}_i^k = \frac{1}{n}\|Z^{[g]}\|^2,$$

where $Z^{[g]} = (Z^k)_{k\in g}$ with $Z^k = (Q_1^k)^T\boldsymbol{\varepsilon}_i^k$. By definition of $\boldsymbol{\varepsilon}_i^k$, $\mathrm{Var}(Z^k) = 1/\omega_{0,ii}^k$ and $\mathrm{Var}(Z^{[g]})$ is a diagonal matrix with the diagonal $(1/\omega_{0,ii}^k)_{k\in g}$. Note that the independence of Z^k and $Z^{k'}$ $(k \neq k')$ comes from the fact that $\boldsymbol{\varepsilon}_i^k$ and $\boldsymbol{\varepsilon}_i^{k'}$ are independent. Therefore

$$\mathbb{P}(\{\mathcal{A}_{ij}^g\}^c) = \mathbb{P}(\|Z^{[g]}\|/\sqrt{n} > \lambda_{ij}^g/2) = \mathbb{P}(\|Z^{[g]}\| - E\|Z^{[g]}\| > \sqrt{n}\lambda_{ij}^g/2 - E\|Z^{[g]}\|).$$

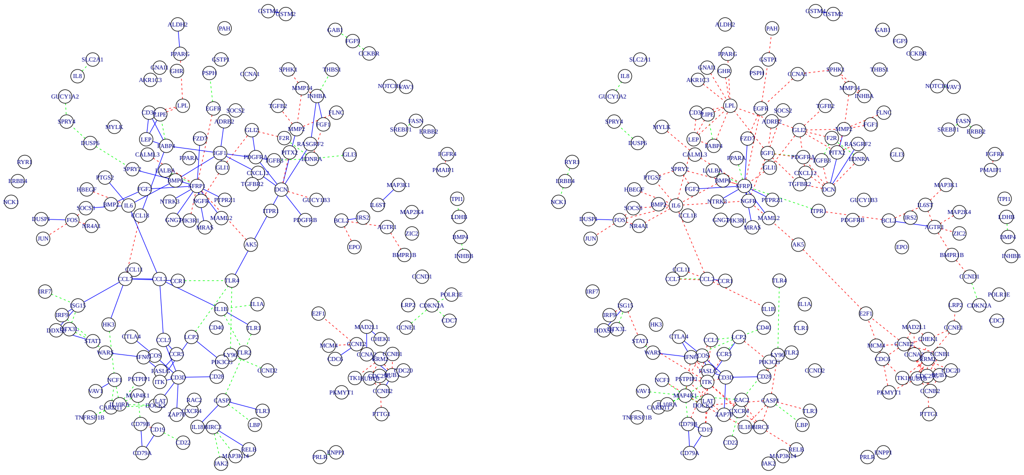


Figure 13: Estimated gene networks combined using JEM-G (left) and JSEM (right), with edges shared between ER+ and ER- blue solid and differential edges red dashed (ER+) / green dashed (ER-).

Applying Lemma E.1,

$$\begin{aligned}\mathbb{P}(\{\mathcal{A}_{ij}^g\}^c) &\leq \mathbb{P}(|\|Z^{[g]}\| - E\|Z^{[g]}\|| > \sqrt{n}\lambda_{ij}^g/2 - E\|Z^{[g]}\|) \\ &\leq 2\exp\left\{-\frac{2}{\pi^2\|\mathrm{Var}(Z^{[g]})\|}\left(\frac{\sqrt{n}\lambda_{ij}^g}{2} - E\|Z^{[g]}\|\right)^2\right\}.\end{aligned}$$

Choose λ_{ij}^g such that the right-hand side of above inequality is less than $2G_0^{-q}$ for some positive parameter q. Then

$$\lambda_{ij}^g \geq \frac{2}{\sqrt{n}}\left(E\|Z^{[g]}\| + \frac{\pi}{\sqrt{2}}\sqrt{q\log G_0}\sqrt{\|\mathrm{Var}(Z^{[g]})\|}\right),$$

and is satisfied if

$$\lambda_{ij}^g \geq \max_{k\in g}\frac{2}{\sqrt{n\omega_{0,ii}^k}}\left(\sqrt{|g|} + \frac{\pi}{\sqrt{2}}\sqrt{q\log G_0}\right),$$

by Lemma E.1. With the above choice of λ_{ij}^g,

$$\mathbb{P}(\mathcal{A}^c) \leq \sum_{i=1}^p\sum_{j\neq i}\sum_{g\in\mathscr{G}^{ij}}\mathbb{P}(\{\mathcal{A}_{ij}^g\}^c) \leq 2pG_0^{1-q},$$

or equivalently, $\mathbb{P}(\mathcal{A}) \geq 1 - 2pG_0^{1-q}$. ■

Proof of Proposition A.1

Proof For all $\Theta_i \in \mathbb{M}(p-1, K)$, using a similar argument to that in Lemma 3.1 of Lounici et al. (2011), it is straightforward to verify the following:

$$\sum_{k=1}^{K} \frac{1}{n} \|\mathbf{X}_{-i}^k(\hat{\boldsymbol{\theta}}_i^k - \boldsymbol{\theta}_{0,i}^k)\|^2 + \sum_{j \neq i} \sum_{g \in \mathcal{G}^{ij}} \lambda_{ij}^g \|\hat{\boldsymbol{\theta}}_{ij}^{[g]} - \boldsymbol{\theta}_{ij}^{[g]}\|$$

$$\leq \sum_{k=1}^{K} \frac{1}{n} \|\mathbf{X}_{-i}^k(\boldsymbol{\theta}_i^k - \boldsymbol{\theta}_{0,i}^k)\|^2 + 4 \sum_{(j,g) \in J(\Theta_i)} \lambda_{ij}^g \min\left(\|\boldsymbol{\theta}_{ij}^{[g]}\|, \|\hat{\boldsymbol{\theta}}_{ij}^{[g]} - \boldsymbol{\theta}_{ij}^{[g]}\|\right), \tag{30}$$

$$\Big\{\sum_{k \in g} \langle n^{-1}\mathbf{X}_j^k, \mathbf{X}_{-i}^k(\hat{\boldsymbol{\theta}}_i^k - \boldsymbol{\theta}_{0,i}^k)\rangle^2\Big\}^{1/2} \leq \frac{3\lambda_{ij}^g}{2}, \tag{31}$$

$$\mathcal{M}(\hat{\Theta}_i) \leq \frac{4\phi_{\max}}{\lambda_{\min}^2} \sum_{k=1}^{K} \frac{1}{n} \|\mathbf{X}_{-i}^k(\hat{\boldsymbol{\theta}}_i^k - \boldsymbol{\theta}_{0,i}^k)\|^2, \tag{32}$$

where $\lambda_{\min}$ and $\phi_{\max}$ are defined in Proposition A.1.

Let $\Delta = (\boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^K)$ be a matrix in $\mathbb{M}(p, K)$ such that $\delta_j^k = \hat{\theta}_{ij}^k - \theta_{0,ij}^k$ for $j \neq i$ and $\delta_i^k = 0$ for all k. We would like to first find an upper bound for B^2, where

$$B^2 := \sum_k \frac{1}{n} \|\mathbf{X}_{-i}^k(\hat{\boldsymbol{\theta}}_i^k - \boldsymbol{\theta}_{0,i}^k)\|^2 = \sum_k \frac{1}{n} \|\mathbf{X}^k \boldsymbol{\delta}^k\|^2.$$

On the event $\mathcal{A}$, we have

$$\sum_{j \neq i} \sum_{g \in \mathscr{G}^{ij}} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\| \leq B^2 + \sum_{j \neq i} \sum_{g \in \mathscr{G}^{ij}} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\| \leq 4 \sum_{(j,g) \in J(\Theta_{0,i})} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\|, \tag{33}$$

where the second inequality follows from setting $\Theta_i = \Theta_{0,i}$ in (30). Therefore

$$\sum_{(j,g) \in J(\Theta_{0,i})^c} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\| \leq 3 \sum_{(j,g) \in J(\Theta_{0,i})} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\|,$$

which implies that $\Delta \in \mathcal{F}$, the restricted set defined in Assumption (A1). Under Assumption (A1) with $\kappa = \kappa(s_0)$, one has

$$B^2 \geq \kappa^2 \|\Delta_J\|_F^2 = \kappa^2 \sum_{(j,g) \in J(\Theta_{0,i})} \|\boldsymbol{\delta}_j^{[g]}\|^2. \tag{34}$$

Combing (33) and the Cauchy-Schwarz inequality, we obtain

$$B^2 \leq 4 \sum_{(j,g) \in J(\Theta_{0,i})} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\| \leq 4\Big\{\sum_{(j,g) \in J(\Theta_{0,i})} (\lambda_{ij}^g)^2\Big\}^{1/2} \Big(\sum_{(j,g) \in J(\Theta_{0,i})} \|\boldsymbol{\delta}_j^{[g]}\|^2\Big)^{1/2} \tag{35}$$

$$\leq 4\Big\{\sum_{(j,g) \in J(\Theta_{0,i})} (\lambda_{ij}^g)^2\Big\}^{1/2} \frac{B}{\kappa}, \tag{36}$$

where the last inequality in (36) comes from (34). Canceling out the extra B in (36), we get

$$B^2 = \sum_k \frac{1}{n} \|\mathbf{X}_{-i}^k(\hat{\boldsymbol{\theta}}_i^k - \boldsymbol{\theta}_{0,i}^k)\|^2 \leq \frac{16}{\kappa^2} \sum_{(j,g)\in J(\Theta_{0,i})} (\lambda_{ij}^g)^2. \tag{37}$$

To show the inequality in (11), we note by (33), the Cauchy-Schwarz inequality, (34) and (37),

$$\begin{aligned}
\sum_{j\neq i} \sum_{g\in\mathscr{G}^{ij}} \|\boldsymbol{\delta}_j^{[g]}\| &\leq \frac{1}{\lambda_{\min}} \sum_{j\neq i} \sum_{g\in\mathscr{G}^{ij}} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\| \leq \frac{4}{\lambda_{\min}} \sum_{(j,g)\in J(\Theta_{0,i})} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\| \\
&\leq \frac{4}{\lambda_{\min}} \Big\{ \sum_{(j,g)\in J(\Theta_{0,i})} \|\boldsymbol{\delta}_j^{[g]}\|^2 \Big\}^{1/2} \Big\{ \sum_{(j,g)\in J(\Theta_{0,i})} (\lambda_{ij}^g)^2 \Big\}^{1/2} \\
&\leq \frac{4}{\lambda_{\min}} \frac{B}{\kappa} \Big\{ \sum_{(j,g)\in J(\Theta_{0,i})} (\lambda_{ij}^g)^2 \Big\}^{1/2} \\
&\leq \frac{16}{\kappa^2 \lambda_{\min}} \sum_{(j,g)\in J(\Theta_{0,i})} (\lambda_{ij}^g)^2.
\end{aligned}$$

(12) follows readily from (32) and (37)

$$\mathcal{M}(\hat{\Theta}_i) \leq \frac{4\phi_{\max}}{\lambda_{\min}^2} B^2 \leq \frac{64\phi_{\max}}{\kappa^2 \lambda_{\min}^2} \sum_{(j,g)\in J(\Theta_{0,i})} (\lambda_{ij}^g)^2.$$

Finally, we prove (13). Let $J_0 = J(\Theta_{0,i})$ and J_1 denote the set of indices in J_0^c corresponding to the s_i largest values of $\lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\|$. The dependence of J_0 and J_1 on i is made implicit here for clarity. Let $J_{01} = J_0 \cup J_1$. So $|J_{01}| \leq 2s_i$. Let (j_ℓ, g_ℓ) be the index of the ℓth largest element of the set $\{\lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\| : (j,g) \in J_0^c\}$. Then

$$\lambda_{ij_\ell}^{g_\ell} \|\Delta_{ij_\ell}^{[g_\ell]}\| \leq \sum_{(j,g)\in J_0^c} \frac{\lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\|}{\ell}.$$

Combining with the fact that $\Delta \in \mathcal{F}$, we have on the event $\mathcal{A}$,

$$\begin{aligned}
\sum_{(j,g)\in J_{01}^c} \Big(\lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\|\Big)^2 \leq \sum_{(j,g)\in J_0^c} \Big(\lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\|\Big)^2 &\leq \sum_{\ell=s_i+1}^{\infty} \frac{\Big(\sum_{(j,g)\in J_0^c} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\|\Big)^2}{\ell^2} \\
&\leq \frac{1}{s_i} \Big(\sum_{(j,g)\in J_0^c} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\| \Big)^2 \\
&\leq \frac{9}{s_i} \Big(\sum_{(j,g)\in J_0} \lambda_{ij}^g \|\boldsymbol{\delta}_j^{[g]}\| \Big)^2 \\
&\leq \frac{9}{s_i} \sum_{(j,g)\in J_0} (\lambda_{ij}^g)^2 \|\Delta_{J_0}\|_F^2 \qquad (38) \\
&\leq \frac{9}{s_i} \sum_{(j,g)\in J_0} (\lambda_{ij}^g)^2 \|\Delta_{J_{01}}\|_F^2,
\end{aligned}$$

where (38) comes from the Cauchy-Schwarz inequality. It follows immediately that

$$\lambda_{\min}^2 \sum_{(j,g)\in J_{01}^c} \|\boldsymbol{\delta}_j^{[g]}\|^2 \le \frac{9}{s_i} \sum_{(j,g)\in J_0} (\lambda_{ij}^g)^2 \|\Delta_{J_{01}}\|_F^2.$$

Hence

$$\begin{aligned}
\|\hat{\Theta}_i - \Theta_{0,i}\|_F^2 =& \sum_{j\neq i} \sum_{g\in \mathscr{G}^{ij}} \|\boldsymbol{\delta}_j^{[g]}\|^2 = \|\Delta_{J_{01}}\|_F^2 + \|\Delta_{J_{01}^c}\|_F^2 \\
\le& \|\Delta_{J_{01}}\|_F^2 + \frac{9}{s_i \lambda_{\min}^2} \sum_{(j,g)\in J_0} (\lambda_{ij}^g)^2 \|\Delta_{J_{01}}\|_F^2 \\
\le& \frac{10}{s_i \lambda_{\min}^2} \sum_{(j,g)\in J_0} (\lambda_{ij}^g)^2 \|\Delta_{J_{01}}\|_F^2. \qquad (39)
\end{aligned}$$

Now we bound $\|\Delta_{J_{01}}\|_F$. Note (35) implies that

$$B^2 \le 4\Big\{ \sum_{(j,g)\in J_0} (\lambda_{ij}^g)^2 \Big\}^{1/2} \|\Delta_{J_0}\|_F \le 4\Big\{ \sum_{(j,g)\in J_0} (\lambda_{ij}^g)^2 \Big\}^{1/2} \|\Delta_{J_{01}}\|_F.$$

Further we have $B^2 \ge \kappa^2(2s_0)\|\Delta_{J_{01}}\|_F^2$ under Assumption (A1) with $s = 2s_0$. So

$$\|\Delta_{J_{01}}\|_F^2 \le \frac{B^2}{\kappa^2(2s_0)} \le \frac{4}{\kappa^2(2s_0)} \Big\{ \sum_{(j,g)\in J_0} (\lambda_{ij}^g)^2 \Big\}^{1/2} \|\Delta_{J_{01}}\|_F,$$

which implies

$$\|\Delta_{J_{01}}\|_F \le \frac{4}{\kappa^2(2s_0)} \Big\{ \sum_{(j,g)\in J_0} (\lambda_{ij}^g)^2 \Big\}^{1/2}. \qquad (40)$$

Plugging the bound in (40) into (39), we obtain

$$\|\hat{\Theta}_i - \Theta_{0,i}\|_F^2 \le \left\{ \frac{4\sqrt{10}}{\kappa^2(2s_0)} \right\}^2 \left\{ \frac{\sum_{(j,g)\in J_0} (\lambda_{ij}^g)^2}{\lambda_{\min}\sqrt{s_i}} \right\}^2,$$

or equivalently

$$\|\hat{\Theta}_i - \Theta_{0,i}\|_F \le \frac{4\sqrt{10}}{\kappa^2(2s_0)} \frac{\sum_{(j,g)\in J(\Theta_{0,i})} (\lambda_{ij}^g)^2}{\lambda_{\min}\sqrt{s_i}}.$$

■

Proof of Corollary A.1

Proof By definition, $\omega^k_{0,ij} = -\theta^k_{0,ij}\omega^k_{0,ii}$ for all $j \neq i$ and $k = 1, \ldots, K$. Further, under Assumption (A2), $\omega^k_{0,ii} \leq \phi_{\max}(\Omega^k_0) = \phi^{-1}_{\min}(\Sigma^k_0) \leq c_0$ for all i, k . Therefore

$$\begin{aligned}\sum_k \|\widetilde{\Omega}^k - \Omega^k_0\|^2_F &= \sum_k \sum_{i=1}^p \sum_{j \in J(\Theta_{0,i}) \cap J(\hat{\Theta}_i)^c} (\theta^k_{0,ij}\omega^k_{0,ii})^2 \\ &= \sum_{i=1}^p \sum_{j \in J(\Theta_{0,i}) \cap J(\hat{\Theta}_i)^c} \sum_{g \in \mathscr{G}^{ij}} \sum_{k \in g} (\theta^k_{0,ij}\omega^k_{0,ii})^2 \\ &\leq {c_0}^2 \sum_{i=1}^p \sum_{j \in J(\Theta_{0,i}) \cap J(\hat{\Theta}_i)^c} \sum_{g \in \mathscr{G}^{ij}} \|\boldsymbol{\theta}^{[g]}_{0,ij}\|^2 \\ &\leq {c_0}^2 \sum_{i=1}^p \sum_{j \neq i} \sum_{g \in \mathscr{G}^{ij}} \|\boldsymbol{\theta}^{[g]}_{0,ij} - \hat{\boldsymbol{\theta}}^{[g]}_{ij}\|^2.\end{aligned}$$

Under Assumption (A1) with $s = 2s_0$, applying Proposition A.1 with $\lambda^g_{ij} = \lambda_{\max}$ in (14),

$$\sum_{j \neq i} \sum_{g \in \mathscr{G}^{ij}} \|\boldsymbol{\theta}^{[g]}_{0,ij} - \hat{\boldsymbol{\theta}}^{[g]}_{ij}\|^2 \leq \left\{ \frac{4\sqrt{10}}{\kappa^2(2s_0)} \lambda_{\max} \right\}^2 s_i.$$

Therefore,

$$\sum_k \|\widetilde{\Omega}^k - \Omega^k_0\|^2_F \leq \left\{ \frac{4\sqrt{10}c_0}{\kappa^2(2s_0)} \lambda_{\max} \right\}^2 \sum_{i=1}^p s_i = \left\{ \frac{4\sqrt{10}c_0}{\kappa^2(2s_0)} \lambda_{\max} \right\}^2 S_0.$$

It follows immediately that

$$\begin{aligned}\frac{1}{K}\sum_k \|\widetilde{\Omega}^k - \Omega^k_0\|_F &\leq \frac{1}{\sqrt{K}}\Big\{\sum_k \|\widetilde{\Omega}^k - \Omega^k_0\|^2_F\Big\}^{1/2} \leq \frac{4\sqrt{10}c_0}{\kappa^2(2s_0)}\lambda_{\max}\sqrt{\frac{S_0}{K}} \\ &\leq C_{\text{bias}}\sqrt{\frac{S_0}{nK}}\left(\sqrt{|g_{\max}|} + \frac{\pi}{\sqrt{2}}\sqrt{q \log G_0}\right).\end{aligned}$$

To bound the size of the estimated edge set $\hat{E}^k$, we notice if there exists (i, j, k) such that $\hat{\boldsymbol{\theta}}^k_{ij} \neq 0$, then $\hat{\boldsymbol{\theta}}^{[g]}_{ij} \neq \mathbf{0}$, where $g \ni k$. Hence $\mathcal{M}(\hat{\boldsymbol{\theta}}^k_i) \leq \mathcal{M}(\hat{\Theta}_i)$ for all k. By (12), the upper bound for $\hat{E}^k$ is thus

$$|\hat{E}^k| \leq \sum_{i=1}^p \mathcal{M}(\hat{\boldsymbol{\theta}}^k_i) \leq \sum_{i=1}^p \frac{64\phi_{\max}}{\kappa^2(s_0)\lambda^2_{\min}} \sum_{(j,g) \in J(\Theta_{0,i})} (\lambda^g_{ij})^2 = \frac{64\phi_{\max}}{\kappa^2(s_0)} \sum_{i=1}^p s_i \leq \frac{64\phi_{\max}}{\kappa^2(s_0)} S_0.$$

■

References

Francis Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

Sumanta Basu, Ali Shojaie, and George Michailidis. Network granger causality with inherent grouping structure. *Journal of Machine Learning Research*, 16:417–453, 2015.

Peter J. Bickel, Ya'Acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

Patrick Breheny and Jian Huang. Penalized methods for bi-level variable selection. *Statistics and Its Interface*, 2(3):369–380, 2009.

Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data*. Springer-Verlag Berlin Heidelberg, 2011.

Julien Chiquet, Yves Grandvalet, and Christophe Ambroise. Inferring multiple graphical structures. *Statistics and Computing*, 21(4):537–553, 2011.

Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

Matthias Dehmer and Frank Emmert-Streib. *Analysis of Microarray Data: A Network-Based Approach*. John Wiley & Sons, 2008.

David Edwards. *Introduction to Graphical Modelling*. Springer New York, 2000.

Jacques Ferlay, Isabelle Soerjomataram, M Ervik, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald Maxwell Parkin, David Forman, and Freddie Bray. *GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]*. Lyon, France: International Agency for Research on Cancer, 2013.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Eric A. Graham, Stephen S. Mulkey, Kaoru Kitajima, Nathan G. Phillips, and S. Joseph Wright. Cloud cover limits net co2 uptake and growth of a rainforest tree during tropical rainy seasons. *Proceedings of the National Academy of Sciences*, 100(2):572–576, 2003.

Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

I. Harris, P.D. Jones, T.J. Osborn, and D.H. Lister. Updated high-resolution grids of monthly climatic observations–the cru ts3. 10 dataset. *International Journal of Climatology*, 34(3):623–642, 2014.

Jean Honorio and Dimitris Samaras. Multi-task learning of gaussian graphical models. In *International Conference on Machine Learning (ICML)*, pages 447–454, 2010.

Akash K. Kaushik, Ali Shojaie, Katrin Panzitt, Rajni Sonavane, Harene Venghatakrishnan, Mohan Manikkam, Alexander Zaslavsky, Vasanta Putluri, Vihas T. Vasu, Yiqing Zhang, et al. Inhibition of the hexosamine biosynthetic pathway promotes castration-resistant prostate cancer. *Nature Communications*, 7, 2016.

Steffen L Lauritzen. *Graphical Models*. Oxford University Press, New York, 1996.

Karim Lounici, Massimiliano Pontil, Sara van de Geer, and Alexandre B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Annals of Statistics*, 39(4):2164–2204, 2011.

Aurelie C Lozano, Hongfei Li, Alexandru Niculescu-Mizil, Yan Liu, Claudia Perlich, Jonathan Hosking, and Naoki Abe. Spatial-temporal causal modeling for climate change attribution. In *SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 587–596. ACM, 2009.

Jing Ma, Ali Shojaie, and George Michailidis. Network-based pathway enrichment analysis with incomplete network information. *Bioinformatics*, page btw410, 2016.

Nicolai Meinshausen and Peter Bühlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, pages 246–270, 2009.

Karthik Mohan, Palma London, Maryam Fazel, Daniela Witten, and Su-In Lee. Node-based learning of multiple gaussian graphical models. *Journal of Machine Learning Research*, 15(1):445–488, 2014.

Murray C. Peel, Brian L. Finlayson, and Thomas A. McMahon. Updated world map of the köppen-geiger climate classification. *Hydrology and Earth System Sciences Discussions Discussions*, 4 (2):439–473, 2007.

Bertrand Perroud, Jinoo Lee, Nelly Valkova, Amy Dhirapong, Pei-Yin Lin, Oliver Fiehn, Dietmar Kültz, and Robert H. Weiss. Pathway analysis of kidney cancer using proteomics and metabolic profiling. *Molecular Cancer*, 5:64, 2006.

Christine Peterson, Francesco Stingo, and Marina Vannucci. Bayesian inference of multiple gaussian graphical models. *Journal of the American Statistical Association*, 110(509):159–174, 2015.

Miguel Angel Pujana, Jing-Dong J Han, Lea M Starita, Kristen N Stevens, Muneesh Tewari, Jin Sook Ahn, Gad Rennert, Víctor Moreno, Tomas Kirchhoff, Bert Gold, et al. Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nature Genetics*, 39(11): 1338–1349, 2007.

Nagireddy Putluri, Ali Shojaie, Vihas T Vasu, Shaiju K Vareed, Srilatha Nalluri, Vasanta Putluri, Gagan Singh Thangjam, Katrin Panzitt, Christopher T Tallman, Charles Butler, et al. Metabolomic profiling reveals potential markers and bioprocesses altered in bladder cancer progression. *Cancer Research*, 71(24):7376–7386, 2011.

C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

Adam J. Rothman, Peter J. Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Rajen D. Shah and Richard J. Samworth. Variable selection with error control: Another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):55–80, 2013.

Thomas F. Stocker, Dahe Qin, Gian-Kasper Plattner, Melinda Tignor, Simon K. Allen, Judith Boschung, Alexander Nauels, Yu Xia, Vincent Bex, Pauline M. Midgley, et al. Climate change 2013: The physical science basis. contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change. Technical report, Intergovernmental Panel on Climate Change, 2013.

TCGA. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

Tanja Zerenner, Petra Friederichs, Klaus Lehnertz, and Andreas Hense. A gaussian graphical model approach to climate networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 24(2): 023103, 2014.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

Liming Zhou, Aiguo Dai, Yongjiu Dai, Russell S Vose, Cheng-Zhi Zou, Yuhong Tian, and Haishan Chen. Spatial dependence of diurnal temperature range trends on precipitation from 1950 to 2004. *Climate Dynamics*, 32(2):429–440, 2009.

Shuheng Zhou, Philipp Rutimann, Min Xu, and Peter Bühlmann. High-dimensional covariance estimation based on gaussian graphical models. *Journal of Machine Learning Research*, 12: 2975–3026, 2011.

Yunzhang Zhu, Xiaotong Shen, and Wei Pan. Structural pursuit over multiple undirected graphs. *Journal of the American Statistical Association*, 109(508):1683–1696, 2014.