Joint Estimation of Multiple Graphical Models

Jian Guo, Elizaveta Levina, George Michailidis and Ji Zhu
Department of Statistics, University of Michigan, Ann Arbor

October 27, 2009

Abstract

Gaussian graphical models explore dependence relationships between random variables, through estimation of the corresponding inverse covariance (precision) matrices. The objective of this paper is to develop an estimator for such models appropriate for heterogeneous data; specifically, data obtained from different categories that share some common structure, but also exhibit differences. An example of such a data structure is gene networks corresponding to different subtypes of a certain disease. In this setting, estimating a single graphical model would mask the underlying heterogeneity, while estimating separate models for each category ignores the common structure. We propose a method which jointly estimates several graphical models corresponding to the different categories present in the data. The method aims to preserve the common structure, while allowing for differences between the categories. This is achieved through a hierarchical penalty that targets the removal of common zeros in the precision matrices across categories. We establish the asymptotic consistency and sparsistency of the proposed estimator in the high-dimensional case, and illustrate its superior performance on a number of simulated networks. An application to learning semantic connections between terms from webpages collected from computer science departments is also included.

KEY WORDS: Covariance matrix, graphical models, hierarchical penalty, high-dimensional data, networks.

1 Introduction

Graphical models represent the relationships between a set of random variables through their joint distribution. Generally, the variables correspond to the nodes of the graph, while edges represent their marginal or conditional dependencies. The study of graphical models has attracted a lot of attention both in the statistical and computer science literatures; see, for example, the books by Lauritzen (1996) and Pearl (2000). They have proved useful in a variety of contexts, including causal inference and estimation of networks. Special members of this family of models include Bayesian networks that correspond to a directed acyclic graph, and Gaussian models that assume the joint distribution to be Gaussian. In the latter case, because the Gaussian distribution is fully characterized by its first two moments, the entire dependence structure can be determined from the covariance matrix, where off-diagonal elements are proportional to marginal correlations, or, more commonly, from the precision matrix (inverse covariance matrix), where the off-diagonal elements are proportional to partial correlations. Specifically, variables j and j' are conditionally independent given all other variables (there is no edge between j and j' in the graph), if and only if the (j, j')-th element in the precision matrix is zero; thus the problem of estimating a Gaussian graphical model is equivalent to estimating a precision matrix.

The literature on estimating a precision matrix goes back to Dempster (1972), who advocated the estimation of a sparse dependence structure, i.e., setting some elements of the precision matrix to zero. Edwards (2000) gave an extensive review of early work in this area. A standard approach is the backward stepwise selection method, which starts from removing the least significant edges from a fully connected graph, and continues removing edges until all remaining edges9 are significant according to an individual partial correlation test. This procedure does not account for multiple testing; a conservative simultaneous testing procedure was proposed by Drton and Perlman (2004).

More recently, the focus has shifted to using regularization for sparse estimation of the precision matrix and the corresponding graphical model. For example, Meinshausen and Buhlmann (2006) proposed to select edges for each node in the graph by regressing the variable on all other variables using the ℓ_1 penalized regression (lasso). This method reduces to solving p separate regression problems, and does not provide an estimate of the matrix itself. A penalized maximum likelihood approach using the ℓ_1 penalty has been considered by Yuan and Lin (2007), d'Aspremont et al. (2008), Friedman et al. (2008) and Rothman et al. (2008), who have all proposed different algorithms for computing this estimator. This approach produces a sparse estimate of the precision matrix, which can then be used to infer a graph, and has been referred to as Glasso (Friedman et al., 2008) or SPICE (Rothman et al., 2008). Theoretical properties of the ℓ_1-penalized maximum likelihood estimator in the large p scenario were derived by Rothman et al. (2008), who showed that the rate of convergence in the Frobenius norm is $O_P(\sqrt{q(\log p)/n})$, where q is the total number of nonzero elements in the precision matrix. Fan et al. (2009) and Lam and Fan (2009) extended this penalized maximum likelihood approach to general non-convex penalties, such as SCAD (Fan and Li, 2001), while Lam and Fan (2009) also established a "sparsistency" property of the penalized likelihood estimator, implying that it estimates true zeros correctly with probability tending to 1. Alternative penalized estimators based on the pseudo-likelihood instead of the likelihood have been recently proposed by Rocha et al. (2008) and Peng et al. (2009); the latter paper also established consistency in terms of both estimation and model selection.

The focus so far in the literature has been on estimating a single Gaussian graphical model. However, in many applications it is more realistic to fit a *collection* of such models, due to the heterogeneity of the data involved. For example, consider gene networks describing different subtypes of the same cancer: there are some shared pathways across different subtypes, and there are also links that are unique to a particular subtype. Another example

from text mining, which is discussed in detail in Section 5, is word relationships inferred from webpages. In our example, the webpages are collected from university computer science departments, and the different categories correspond to faculty, student, course, etc. In such cases, borrowing strength across different categories by jointly estimating these models could reveal a common structure and reduce the variance of the estimates, especially when the number of samples is relatively small. To accomplish this joint estimation, we propose a method that links the estimation of separate graphical models through a hierarchical penalty. Its main advantage is the ability to discover a common structure and jointly estimate common links across graphs, which leads to improvements over fitting separate models, since it borrows information from other related graphs. While in this paper we focus on continuous data, this methodology can be extended to graphical models with categorical variables; fitting such models to a single graph has been considered by Kolar and Xing (2008); Ravikumar et al. (2009).

The remainder of the paper is organized as follows. Section 2 introduces the new method and addresses algorithmic issues. Theoretical results on consistency and sparsistency are presented in Section 3. The performance of the proposed method on synthetic and real data is demonstrated in Sections 4 and 5, respectively.

2 Methodology

Suppose we have a heterogeneous data set with p variables and K categories. The k-th category contains n_k observations $(\boldsymbol{x}_1^{(k)}, \ldots, \boldsymbol{x}_{n_k}^{(k)})^{\mathsf{T}}$, where each $\boldsymbol{x}_i^{(k)} = (x_{i,1}^{(k)}, \ldots, x_{i,p}^{(k)})$ is a p-dimensional row vector. Without loss of generality, we assume the observations in the same category are centered along each variable, i.e., $\sum_{i=1}^{n_k} x_{i,j}^{(k)} = 0$ for all $1 \leq j \leq p$ and $1 \leq k \leq K$. We further assume that $\boldsymbol{x}_1^{(k)}, \ldots, \boldsymbol{x}_{n_k}^{(k)}$ are independent and identically distributed, sampled from a p-variate Gaussian distribution with mean zero (without loss of generality since we center the data) and covariance matrix $\boldsymbol{\Sigma}^{(k)}$. Let $\boldsymbol{\Omega}^{(k)} = (\boldsymbol{\Sigma}^{(k)})^{-1} = (\omega_{j,j'}^{(k)})_{p \times p}$. The log-

likelihood of the observations in the k-th category is given by

$$l(\mathbf{\Omega}^{(k)}) = -\frac{n_k}{2}\log(2\pi) + \frac{n_k}{2}\Big[\log\{\det(\mathbf{\Omega}^{(k)})\} - \text{trace}(\widehat{\mathbf{\Sigma}}^{(k)}\mathbf{\Omega}^{(k)})\Big], \tag{1}$$

where $\widehat{\mathbf{\Sigma}}^{(k)}$ is the sample covariance matrix for the k-th category, and $\det(\cdot)$ and $\text{trace}(\cdot)$ are the determinant and the trace of a matrix, respectively.

The most direct way to deal with such heterogeneous data is to estimate K individual graphical models. We can compute a separate ℓ_1-regularized estimator for each category k, $1 \leq k \leq K$, by solving

$$\min_{\mathbf{\Omega}^{(k)}} \ \text{trace}(\widehat{\mathbf{\Sigma}}^{(k)}\mathbf{\Omega}^{(k)}) - \log\{\det(\mathbf{\Omega}^{(k)})\} + \lambda_k \sum_{j \neq j'} |\omega_{j,j'}^{(k)}|, \tag{2}$$

where the minimum is taken over symmetric positive definite matrices. The ℓ_1 penalty shrinks some of the off-diagonal elements in $\mathbf{\Omega}^{(k)}$ to zero and the tuning parameter λ_k controls the degree of the sparsity in the estimated precision matrix. Problem (2) can be efficiently solved by existing algorithms such as glasso (Friedman et al., 2008). We will refer to this approach as the separate estimation method and use it as a benchmark to compare with the joint estimation method we propose next.

2.1 The Joint Estimation Method

To improve estimation in cases where graphical models for different categories may share some common structure, we propose a joint estimation method. First, we reparameterize each $\omega_{j,j'}^{(k)}$ as

$$\omega_{j,j'}^{(k)} = \theta_{j,j'}\gamma_{j,j'}^{(k)}, \quad 1 \leq j \neq j' \leq p\,, 1 \leq k \leq K\ . \tag{3}$$

An analogous parameterization in a dimension reduction setting was used in Michailidis and de Leeuw (2001). For identifiability purposes, we restrict $\theta_{j,j'} > 0$, $1 \leq j \neq j' \leq p$. To preserve symmetry, we also require $\theta_{j,j'} = \theta_{j',j}$ and $\gamma_{j,j'}^{(k)} = \gamma_{j',j}^{(k)}$, $1 \leq j \neq j' \leq p$ and $1 \leq k \leq K$. This decomposition treats $\{\omega_{j,j'}^{(1)}, \ldots, \omega_{j,j'}^{(K)}\}$ as a group, with the common factor

$\theta_{j,j'}$ controlling the presence of the link between nodes j and j' in any of the categories, and $\gamma_{j,j'}^{(k)}$ reflects the differences between categories. Let $\boldsymbol{\Theta} = (\theta_{j,j'})_{p\times p}$ and $\boldsymbol{\Gamma}^{(k)} = (\gamma_{j,j'}^{(k)})_{p\times p}$. To estimate this model, we propose the following penalized criterion:

$$\begin{aligned}
\min_{\boldsymbol{\Theta},\{\boldsymbol{\Gamma}^{(k)}\}_{k=1}^K} \quad & \sum_{k=1}^K \Big[\mathrm{trace}(\widehat{\boldsymbol{\Sigma}}^{(k)}\boldsymbol{\Omega}^{(k)}) - \log\{\det(\boldsymbol{\Omega}^{(k)})\}\Big] + \eta_1 \sum_{j\neq j'} \theta_{j,j'} + \eta_2 \sum_{j\neq j'}\sum_{k=1}^K |\gamma_{j,j'}^{(k)}| \\
\text{subject to} \quad & \omega_{j,j'}^{(k)} = \theta_{j,j'}\gamma_{j,j'}^{(k)}, \quad \theta_{j,j'} > 0, \quad 1 \le j, j' \le p \\
& \theta_{j,j'} = \theta_{j',j}, \quad \gamma_{j,j'}^{(k)} = \gamma_{j',j}^{(k)}, \quad 1 \le j \neq j' \le p; 1 \le k \le K \\
& \theta_{j,j} = 1, \quad \gamma_{j,j}^{(k)} = \omega_{j,j}^{(k)} \quad , 1 \le j \le p; 1 \le k \le K, \qquad (4)
\end{aligned}$$

where η_1 and η_2 are two tuning parameters. The first one, η_1, controls the sparsity of the common factors $\theta_{j,j'}$'s and it can effectively remove the common zero elements across $\boldsymbol{\Omega}^{(1)}, \ldots, \boldsymbol{\Omega}^{(K)}$; i.e., if $\theta_{j,j'}$ is shrunk to zero, there will be no link between nodes j and j' in any of the K graphs. If $\theta_{j,j'}$ is not zero, some of the $\gamma_{j,j'}^{(k)}$'s (and hence some of the $\omega_{j,j'}^{(k)}$'s) can still be set to zero by the second penalty. This allows graphs belonging to different categories to have different structures. Note that this decomposition has also been used by Zhou and Zhu (2007) for group variable selection in regression problems.

To simplify the model, the two tuning parameters η_1 and η_2 in (4) can be replaced by a single tuning parameter, by letting $\eta = \eta_1\eta_2$. It turns out that solving (4) is equivalent to solving

$$\begin{aligned}
\min_{\boldsymbol{\Theta},\{\boldsymbol{\Gamma}^{(k)}\}_{k=1}^K} \quad & \sum_{k=1}^K \Big[\mathrm{trace}(\widehat{\boldsymbol{\Sigma}}^{(k)}\boldsymbol{\Omega}^{(k)}) - \log\{\det(\boldsymbol{\Omega}^{(k)})\}\Big] + \sum_{j\neq j'} \theta_{j,j'} + \eta \sum_{j\neq j'}\sum_{k=1}^K |\gamma_{j,j'}^{(k)}| \\
\text{subject to} \quad & \omega_{j,j'}^{(k)} = \theta_{j,j'}\gamma_{j,j'}^{(k)}, \quad \theta_{j,j'} > 0 \\
& \theta_{j,j'} = \theta_{j',j}, \quad \gamma_{j,j'}^{(k)} = \gamma_{j',j}^{(k)}, \quad 1 \le j \neq j' \le p; 1 \le k \le K \\
& \theta_{j,j} = 1, \quad \gamma_{j,j}^{(k)} = \omega_{j,j}^{(k)} \quad , 1 \le j \le p; 1 \le k \le K. \qquad (5)
\end{aligned}$$

Let $Q_{\eta_1,\eta_2}^{**}(\boldsymbol{\Theta}, \{\boldsymbol{\Gamma}^{(k)}\}_{k=1}^K)$ be the criterion we optimize in (4) and let $Q_{\eta}^{*}(\boldsymbol{\Theta}, \{\boldsymbol{\Gamma}^{(k)}\}_{k=1}^K)$ be the corresponding criterion in (5). For two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same size, we denote their

element-wise product by $\boldsymbol{A}\cdot\boldsymbol{B}$. Then, the equivalence between (4) and (5) can be formalized as follows.

Lemma 1 *Let* $(\widehat{\boldsymbol{\Theta}}^{*}, \{\widehat{\boldsymbol{\Gamma}}^{(k)*}\}_{k=1}^{K}))$ *be a local minimizer of* $Q_{\eta}^{*}(\boldsymbol{\Theta}, \{\boldsymbol{\Gamma}^{(k)}\}_{k=1}^{K})$*. Then, there exists a local minimizer of* $Q_{\eta_1,\eta_2}^{**}(\boldsymbol{\Theta}, \{\boldsymbol{\Gamma}^{(k)}\}_{k=1}^{K})$*, denoted as* $(\widehat{\boldsymbol{\Theta}}^{**}, \{\widehat{\boldsymbol{\Gamma}}^{(k)**}\}_{k=1}^{K}))$*, such that* $\widehat{\boldsymbol{\Theta}}^{**} \cdot \widehat{\boldsymbol{\Gamma}}^{(k)**} = \widehat{\boldsymbol{\Theta}}^{*} \cdot \widehat{\boldsymbol{\Gamma}}^{(k)*}$ *for all* $1 \leq k \leq K$*. Similarly, if* $(\widehat{\boldsymbol{\Theta}}^{**}, \{\widehat{\boldsymbol{\Gamma}}^{(k)**}\}_{k=1}^{K}))$ *is a local minimizer of* $Q_{\eta_1,\eta_2}^{**}(\boldsymbol{\Theta}, \{\boldsymbol{\Gamma}^{(k)}\}_{k=1}^{K})$*, then there exists a local minimizer of* $Q_{\eta}^{*}(\boldsymbol{\Theta}, \{\boldsymbol{\Gamma}^{(k)}\}_{k=1}^{K})$*, denoted as* $(\widehat{\boldsymbol{\Theta}}^{*}, \{\widehat{\boldsymbol{\Gamma}}^{(k)*}\}_{k=1}^{K}))$*, such that* $\widehat{\boldsymbol{\Theta}}^{**} \cdot \widehat{\boldsymbol{\Gamma}}^{(k)**} = \widehat{\boldsymbol{\Theta}}^{*} \cdot \widehat{\boldsymbol{\Gamma}}^{(k)*}$ *for all* $1 \leq k \leq K$*.*

The proof follows closely the proof of the Lemma in Zhou and Zhu (2007) and is omitted. This result implies that in practice, instead of tuning two parameters η_1 and η_2, we only need to tune one parameter η, which reduces the overall computational cost.

2.2 The Algorithm

First we reformulate the problem (5) in a more convenient form for computational purposes.

Lemma 2 *Let* $\{\widehat{\boldsymbol{\Omega}}^{(k)}\}_{k=1}^{K}$ *be a local minimizer of*

$$\min_{\{\boldsymbol{\Omega}^{(k)}\}_{k=1}^{K}} \sum_{k=1}^{K} \Big[\mathrm{trace}(\widehat{\boldsymbol{\Sigma}}^{(k)}\boldsymbol{\Omega}^{(k)}) - \log\{\det(\boldsymbol{\Omega}^{(k)})\}\Big] + \lambda \sum_{j\neq j'} \sqrt{\sum_{k=1}^{K} |\omega_{j,j'}^{(k)}|}, \tag{6}$$

where $\lambda = 2\sqrt{\eta}$*. Then, there exists a local minimizer of* (5)*,* $(\widehat{\boldsymbol{\Theta}}, \{\widehat{\boldsymbol{\Gamma}}^{(k)}\}_{k=1}^{K})$*, such that* $\widehat{\boldsymbol{\Omega}}^{(k)} = \widehat{\boldsymbol{\Theta}} \cdot \widehat{\boldsymbol{\Gamma}}^{(k)}$*, for all* $1 \leq k \leq K$*. On the other hand, if* $(\widehat{\boldsymbol{\Theta}}, \{\widehat{\boldsymbol{\Gamma}}^{(k)}\}_{k=1}^{K})$ *is a local minimizer of* (5)*, then there also exists a local minimizer of* (6)*,* $\{\widehat{\boldsymbol{\Omega}}^{(k)}\}_{k=1}^{K}$*, such that* $\widehat{\boldsymbol{\Omega}}^{(k)} = \widehat{\boldsymbol{\Theta}} \cdot \widehat{\boldsymbol{\Gamma}}^{(k)}$*, for all* $1 \leq k \leq K$*.*

The proof is given in the Appendix. To optimize (6) we use an iterative approach based on local linear approximation (LLA) (Zou and Li, 2008). Specifically, letting $(\omega_{j,j'}^{(k)})^{(t)}$ denote the estimates from the previous iteration t, we approximate

$$\sqrt{\sum_{k=1}^{K} |\omega_{j,j'}^{(k)}|} \approx \frac{\sum_{k=1}^{K} |\omega_{j,j'}^{(k)}|}{\sqrt{\sum_{k=1}^{K} |(\omega_{j,j'}^{(k)})^{(t)}|}}. \tag{7}$$

Thus, at the $(t+1)$-th iteration, problem (6) is decomposed into K individual optimization problems:

$$(\boldsymbol{\Omega}^{(k)})^{(t+1)} \;=\; \arg\min_{\boldsymbol{\Omega}^{(k)}} \; \mathrm{trace}(\widehat{\boldsymbol{\Sigma}}^{(k)}\boldsymbol{\Omega}^{(k)}) - \log\{\det(\boldsymbol{\Omega}^{(k)})\} + \lambda \sum_{j\neq j'} \tau_{j,j'}^{(k)} |\omega_{j,j'}^{(k)}|, \tag{8}$$

where $\tau_{j,j'}^{(k)} = \left(\sum_{k=1}^{K} |(\omega_{j,j'}^{(k)})^{(t)}|\right)^{-1/2}$. Note that criterion (8) is exactly the sparse precision matrix estimation problem with weighted ℓ_1 penalty; the solution can be efficiently computed using the Glasso algorithm of Friedman et al. (2008). For numerical stability, we threshold $\left(\sum_{k=1}^{K} |(\omega_{j,j'}^{(k)})^{(t)}|\right)^{1/2}$ at 10^{-10}. In summary, the proposed algorithm for solving (6) is:

Step 0. Initialize $\widehat{\boldsymbol{\Omega}}^{(k)} = (\widehat{\boldsymbol{\Sigma}}^{(k)} + \nu \boldsymbol{I}_p)^{-1}$ for all $1 \leq k \leq K$, where $\boldsymbol{I}_p$ is the identity matrix and the constant ν is chosen to guarantee $\widehat{\boldsymbol{\Sigma}}^{(k)} + \nu \boldsymbol{I}_p$ is positive definite;

Step 1. Update $\widehat{\boldsymbol{\Omega}}^{(k)}$ by (8) for all $1 \leq k \leq K$ using glasso;

Step 2. Repeat Step 1 until convergence is achieved.

2.3 Model Selection

The tuning parameter λ in (6) controls the sparsity of the resulting estimator. We select the optimal tuning parameter by minimizing a Bayesian information criterion (BIC), which balances the goodness of fit and the model complexity. Specifically, we define BIC for the proposed joint estimation method as

$$BIC(\lambda) = \sum_{k=1}^{K} \Big[\mathrm{trace}(\widehat{\boldsymbol{\Sigma}}^{(k)}\widehat{\boldsymbol{\Omega}}_{\lambda}^{(k)}) - \log\{\det(\widehat{\boldsymbol{\Omega}}_{\lambda}^{(k)})\} + \log(n_k) df_k\Big], \tag{9}$$

where $\widehat{\boldsymbol{\Omega}}_{\lambda}^{(1)}, \ldots, \widehat{\boldsymbol{\Omega}}_{\lambda}^{(K)}$ are the estimates from (6) with tuning parameter λ and the degrees of freedom are defined as $df_k = \#\{(j,j') : j < j', \widehat{\omega}_{j,j'}^{(k)} \neq 0\}$.

3 Asymptotic Properties

Next, we derive the asymptotic properties of the joint estimation method, including consistency, as well as sparsistency, when both p and n go to infinity and the tuning parameter

goes to 0 at a certain rate. First, we introduce the necessary notation and state certain regularity conditions on the true precision matrices $\{\mathbf{\Omega}_0^{(1)}, \ldots, \mathbf{\Omega}_0^{(K)}\}$, where $\mathbf{\Omega}_0^{(k)} = (\omega_{0,j,j'}^{(k)})_{p\times p}$, $1 \le k \le K$.

Let $T_k = \{(j, j') : j \neq j', \omega_{j,j'}^{(k)} \neq 0\}$ be the set of indices of all nonzero off-diagonal elements in $\mathbf{\Omega}^{(k)}$, and let $T = T_1 \cup \cdots \cup T_K$. Let $q_k = |T_k|$ and $q = |T|$ be the cardinalities of T_k and T, respectively. We assume that the following regularity conditions hold:

(A) There exist constants τ_1, τ_2 such that for all $p \ge 1$ and $1 \le k \le K$,

$$0 < \tau_1 < \phi_{\min}(\mathbf{\Omega}_0^{(k)}) \le \phi_{\max}(\mathbf{\Omega}_0^{(k)}) < \tau_2 < \infty$$

where $\phi_{\min}$ and $\phi_{\max}$ indicate the minimal and maximal eigenvalues;

(B) There exists a constant $\tau_3 > 0$ such that

$$\min_{1\le k\le K} \min_{(j,j')\in T_k} |\omega_{0,j,j'}^{(k)}| \ge \tau_3 \ .$$

Condition (A) is a standard one, also used in Bickel and Levina (2008) and Rothman et al. (2008), that guarantees that the inverse exists and is well conditioned. Condition (B) ensures that non-zero elements are bounded away from 0.

Theorem 1 *(Consistency) Suppose conditions (A) and (B) hold,* $(p+q)(\log p)/n = o(1)$ *and* $\Lambda_1\sqrt{(\log p)/n} \le \lambda \le \Lambda_2\sqrt{(1+p/q)(\log p)/n}$ *for some positive constants* Λ_1 *and* Λ_2. *Then. there exists a local minimizer* $\{\widehat{\mathbf{\Omega}}^{(k)}\}_{k=1}^K$ *of* (6), *such that*

$$\sum_{k=1}^K \|\widehat{\mathbf{\Omega}}^{(k)} - \mathbf{\Omega}_0^{(k)}\|_F = O_P\Big(\sqrt{\frac{(p+q)\log p}{n}}\Big)$$

Theorem 2 *(Sparsistency) Suppose all conditions in Theorem 1 hold. We further assume* $\sum_{k=1}^K \|\widehat{\mathbf{\Omega}}^{(k)} - \mathbf{\Omega}_0^{(k)}\|^2 = O_P(\eta_n)$, *where* $\eta_n \to 0$ *and* $[(\log p)/n]^{1/2} + \eta_n^{1/2} = O(\lambda)$. *Then with probability tending to 1, the local minimizer* $\{\widehat{\mathbf{\Omega}}^{(k)}\}_{k=1}^K$ *in Theorem 1 satisfies* $\widehat{\omega}_{j,j'}^{(k)} = 0$ *for all* $(j, j') \in T_k^c$, $1 \le k \le K$.

This theorem is analogous to Theorem 2 in Lam and Fan (2009). Note that consistency requires both an upper and a lower bound on λ, whereas sparsistency requires a lower bound. To make the bounds compatible, we need

$$\sqrt{\frac{\log p}{n}} + \sqrt{\eta_n} = O\Big(\sqrt{\frac{(1+p/q)(\log p)}{n}}\Big) \tag{10}$$

Since η_n is the rate of convergence in the operator norm, we can bound it using the fact that $\|M\|_F^2/p \leq \|M\|^2 \leq \|M\|_F^2$. This leads to two extreme cases:

"Worst-case" scenario: $\sum_k \|\widehat{\mathbf{\Omega}}^{(k)} - \mathbf{\Omega}_0^{(k)}\|$ has the same rate as $\sum_k \|\widehat{\mathbf{\Omega}}^{(k)} - \mathbf{\Omega}_0^{(k)}\|_F$ and thus $\eta_n = O((p+q)(\log p)/n)$. The two bounds are compatible only when $q = O(1)$.

"Best-case" scenario: $\sum_k \|\widehat{\mathbf{\Omega}}^{(k)} - \mathbf{\Omega}_0^{(k)}\|$ has the same rate as $\sum_k \|\widehat{\mathbf{\Omega}}^{(k)} - \mathbf{\Omega}_0^{(k)}\|_F/\sqrt{p}$. Then, $\eta_n = O((1+q/p)(\log p)/n)$ and we have both consistency and sparsistency as long as $q = O(p)$.

4 Numerical Evaluation

In this section, we assess the performance of the joint estimation method on three types of simulated networks: a chain, a nearest neighbor, and a scale-free (power law) network. In all cases, we set $p = 100$ and $K = 3$. For $k = 1, \ldots, K$, we generate $n_k = 100$ independently and identically distributed observations from a multivariate normal distribution $N(\mathbf{0}, (\mathbf{\Omega}^{(k)})^{-1})$, where $\mathbf{\Omega}^{(k)}$ is the precision matrix of the k-th category.

We compare the joint estimation method to the method that estimates each category separately via (2). A number of metrics are used to assess performance, including ROC curves, average entropy loss (EL), average Frobenius loss (FL), average false positive (FP) and average false negative (FN) rates, and the average rate of mis-identified common zeros among the categories (CZ). For the ROC curve, we plot sensitivity (the average proportion of correctly detected links) against the average false positive rate over a range of values of

the tuning parameter λ. The average entropy loss (EL) and average Frobenius loss (FL) are defined as:

$$\begin{aligned} EL &= \frac{1}{K}\sum_{k=1}^{K}\left[\text{trace}((\mathbf{\Omega}^{(k)})^{-1}\widehat{\mathbf{\Omega}}^{(k)}) - \log(\det((\mathbf{\Omega}^{(k)})^{-1}\widehat{\mathbf{\Omega}}^{(k)}))\right] - p \ , \\ FL &= \frac{1}{K}\sum_{k=1}^{K}\|\mathbf{\Omega}^{(k)} - \widehat{\mathbf{\Omega}}^{(k)}\|_F^2/\|\mathbf{\Omega}^{(k)}\|_F^2 \ . \end{aligned}$$

The average false positive rate gives the proportion of false discoveries (true zeros estimated as non-zero), the average false negative rate gives the proportion of off-diagonal nonzero elements estimated as zero, and the common zeros error rate gives the proportion of common zeros across $\mathbf{\Omega}^{(1)}, \ldots, \mathbf{\Omega}^{(K)}$ estimated as non-zero. The respective formal definitions are:

$$\begin{aligned} FP &= \frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{1\le j<j'\le p}\mathrm{I}(\omega_{j,j'}^{(k)} = 0, \widehat{\omega}_{j,j'}^{(k)} \neq 0)}{\sum_{1\le j<j'\le p}\mathrm{I}(\omega_{j,j'}^{(k)} = 0)} \ , \\ FN &= \frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{1\le j<j'\le p}\mathrm{I}(\omega_{j,j'}^{(k)} \neq 0, \widehat{\omega}_{j,j'}^{(k)} = 0)}{\sum_{1\le j<j'\le p}\mathrm{I}(\omega_{j,j'}^{(k)} \neq 0)} \ , \\ CZ &= \frac{\sum_{1\le j<j'\le p}\mathrm{I}(\sum_{k=1}^{K}|\omega_{j,j'}^{(k)}| = 0, \sum_{k=1}^{K}|\widehat{\omega}_{j,j'}^{(k)}| \neq 0)}{\sum_{1\le j<j'\le p}\mathrm{I}(\sum_{k=1}^{K}|\omega_{j,j'}^{(k)}| = 0)} \ . \end{aligned}$$

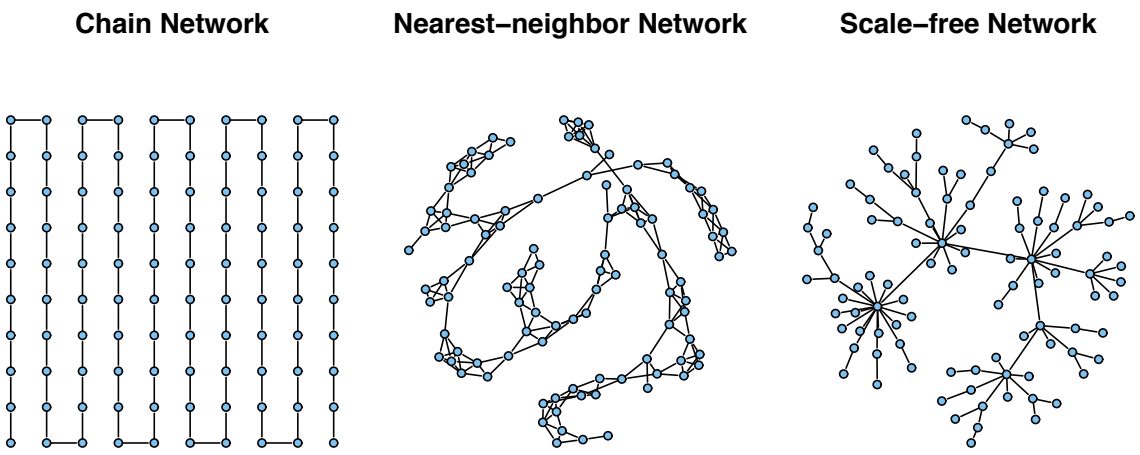


Figure 1: The common links (present in all categories) in the three simulated networks.

Example 1: Chain Networks

In this example, we consider a chain network, which corresponds to a tridiagonal precision matrix. The covariance matrices $\mathbf{\Sigma}^{(k)}$ are constructed as follows: let the (j, j')-th element $\sigma_{j,j'}^{(k)} = \exp\{-|s_j - s_{j'}|/2\}$, where $s_1 < s_2 < \cdots < s_p$ and

$$s_j - s_{j-1} \sim \text{Uniform}(0.5, 1),\ j = 2, \ldots, p$$

Further, let $\mathbf{\Omega}^{(k)} = \left(\mathbf{\Sigma}^{(k)}\right)^{-1}$. The K precision matrices generated by this procedure share the same pattern of zeros (common structure), but the values of their non-zero off-diagonal elements may be different. Figure 1 (left panel) shows the common link structure across the K categories. Further, we add heterogeneity to the common structure by creating additional individual links as follows: for each $\mathbf{\Omega}^{(k)}$, $1 \leq k \leq K$, we randomly pick a pair of symmetric zero elements and replace them with a value uniformly generated from the $[-1, -0.5] \cup [0.5, 1]$ interval. This procedure is repeated ρM times, where M is the number of off-diagonal nonzero elements in the lower triangular part of $\mathbf{\Omega}^{(k)}$ and ρ is the ratio of the number of individual links to the number of common links ("I/C ratio"). In the simulations, we considered values of ρ=0, 1/4, 1 and 4, thus gradually increasing the proportion of individual links.

Example 2: Nearest Neighbor Networks

To generate these networks, we modify the data generating mechanism described in Li and Gui (2006). Specifically, we generate p points randomly on a unit square, calculate all $p(p-1)/2$ pairwise distances, and find m nearest neighbors of each point in terms of this distance. The nearest neighbor network is obtained by linking any two points that are m-nearest neighbors of each other. The integer m controls the degree of sparsity of the network and the value $m = 5$ was chosen in our study. Figure 1 (middle panel) illustrates a realization of the common structure of a nearest-neighbor network. Subsequently, K

individual graphs were generated, by adding some individual links to the common graph with I/C ratio $\rho = 0, 1/4, 1, 4$ by the same method as described in Example 1, with values for the individual links $\omega_{j,j'}^{(k)}$ generated from a uniform distribution on $[-1, -0.5] \cup [0.5, 1]$.

Example 3: Scale-free Networks

To generate the common structure of a power-law network, the Barabasi-Albert algorithm (Barabasi and Albert, 1999) was employed; a realization is depicted in the right panel of Figure 1. The individual links in the k-th network, $1 \leq k \leq K$, are randomly added as before, with I/C ratio $\rho = 0, 1/4, 1, 4$ and the associated elements in $\mathbf{\Omega}^{(k)}$ generated uniformly on $[-1, -0.5] \cup [0.5, 1]$.

Results

Figure 2 shows the estimated ROC curves averaged over 50 replications for all three simulated examples, obtained by varying the tuning parameter. It can be seen that the curves estimated by the joint estimation method dominate those of the separate estimation method when the proportion of individual links is low. When the proportion of individual links is high, the methods perform very similarly. This is as expected, since in the presence of a common structure, the joint estimation method outperforms the separate estimation method, whereas in the absence of a common structure, their performance is fairly similar. Table 1 summarizes results based on 50 replications with the tuning parameter selected by BIC as described in Section 2.3. In general, the joint estimation method produces lower entropy and Frobenius losses, as well as lower false negative rates. When the proportion of common links is high enough, it also produces lower false positive rates and performs better at identifying common zeros than the separate estimation method.

Table 1: Results from the three simulated examples. "S" stands for the separate estimation method and "J" stands for the joint estimation method. EL is the average entropy loss, FL is the average Frobenius loss, FN and FP are the average false negative and false positive rates, and CZ is the average proportion of mis-identified common zeros.

Example	ρ	Method	EL	FL	FN (%)	FP (%)	CZ (%)
1	0	S	20.7(0.8)	0.54(0.01)	0.81(0.6)	5.70(0.6)	14.50(1.6)
		J	12.8(2.4)	0.32(0.02)	0.03(0.1)	4.33(1.0)	6.99(1.6)
	1/4	S	21.3(0.9)	0.52(0.01)	41.32(4.0)	1.32(0.4)	3.83(1.1)
		J	9.5(0.4)	0.32(0.01)	15.59(1.4)	1.65(0.5)	3.22(0.9)
	1	S	23.0(1.6)	0.53(0.01)	73.65(8.1)	0.65(0.4)	1.91(1.2)
		J	12.5(0.6)	0.37(0.01)	44.22(2.7)	1.62(0.4)	2.97(0.7)
	4	S	29.8(0.6)	0.56(0.00)	97.27(1.9)	0.10(0.1)	0.27(0.3)
		J	20.0(0.5)	0.46(0.01)	75.45(2.4)	1.89(0.5)	3.18(0.7)
2	0	S	11.9(0.4)	0.44(0.01)	40.07(2.0)	2.20(0.3)	6.05(0.7)
		J	6.1(0.4)	0.29(0.02)	18.54(4.0)	1.61(0.8)	3.18(1.4)
	1/4	S	13.9(0.6)	0.44(0.01)	44.01(2.8)	2.42(0.4)	6.88(1.0)
		J	8.1(0.3)	0.31(0.01)	27.38(2.5)	1.71(0.3)	2.93(0.4)
	1	S	18.5(1.1)	0.48(0.01)	48.50(3.3)	3.95(0.7)	11.22(2.0)
		J	13.0(0.3)	0.37(0.01)	39.97(1.9)	2.82(0.4)	3.82(0.5)
	4	S	24.8(0.3)	0.54(0.00)	98.72(0.9)	0.09(0.1)	0.25(0.9)
		J	19.3(0.2)	0.47(0.00)	80.80(0.6)	3.24(0.2)	4.75(0.4)
3	0	S	16.9(0.6)	0.47(0.01)	20.71(2.4)	1.89(0.3)	5.30(0.7)
		J	8.1(0.8)	0.29(0.01)	9.41(1.4)	1.45(1.0)	2.78(1.6)
	1/4	S	17.1(0.9)	0.48(0.01)	49.63(4.6)	1.24(0.3)	3.65(1.1)
		J	9.4(0.4)	0.33(0.01)	29.25(2.8)	1.31(0.4)	2.43(0.7)
	1	S	22.3(1.3)	0.51(0.01)	51.84(4.2)	2.80(0.4)	8.21(1.1)
		J	15.2(0.7)	0.40(0.01)	42.53(2.5)	2.16(0.4)	3.19(0.7)
	4	S	27.9(0.2)	0.55(0.01)	99.63(0.5)	0.01(0.0)	0.02(0.0)
		J	23.0(0.5)	0.50(0.01)	82.47(2.3)	2.14(0.6)	3.21(1.0)

5 Data Example

The data set comes from the "World Wide Knowledge Base" project at Carnegie Mellon University. It was collected in 1997 and includes webpages from websites at computer science departments in the following four universities: Cornell, Texas, Washington, and Wisconsin. The webpages were manually classified into seven categories: student, faculty, staff, department, course, project, and other. The full data set can be downloaded from the machine learning repository at the University of California, Irvine (Asuncion and Newman, 2007).

For our analysis, only 1396 webpages corresponding to the four largest categories were selected: student (544 webpages), faculty (374 webpages), course (310 webpages) and project (168 webpages). The original data set was preprocessed by Cardoso-Cachopo (2009) following the following steps: (1) Substituting space for tab, newline, and return characters; (2) Keeping only letters (that is, turning punctuation, numbers, etc. into spaces) and turning all letters to lowercase; (3) Removing words less than 3 characters long and removing the 524 "smart" stopwords; (4) Substituting a single space for multiple spaces; (5) Stemming the documents by applying a stemmer algorithm (Porter, 1980) to the remaining text.

The log-entropy weighting method (Dumais, 1991) was used to calculate the term-document matrix $\boldsymbol{X} = (x_{i,j})_{n\times p}$, with n and p denoting the number of webpages and distinct terms (words), respectively. Let $f_{i,j}$, $1 \leq i \leq n, 1 \leq j \leq p$ be the number of times the j-th term appears in the i-th webpage and let $p_{i,j} = f_{i,j}/\sum_{i=1}^{n} f_{i,j}$. Then, the log-entropy weight of the j-th term is defined as

$$e_j = 1 + \sum_{i=1}^{n} p_{i,j} \log(p_{i,j})/\log(n) \ .$$

Finally, the term-document matrix $\boldsymbol{X}$ is defined as

$$x_{i,j} = e_j \log(1 + f_{i,j}) \ , 1 \leq i \leq n \ , \ 1 \leq j \leq p \ ,$$

and it is normalized along each column. We applied the proposed joint estimation method to $n = 1396$ documents in the four largest categories and $p = 100$ terms with the highest log-entropy weights out of a total of 4800 terms. The resulting common network structure is shown in Figure 3. The area of the circle representing a node is proportional to its log-entropy weight, while the thickness of an edge is proportional to the magnitude of the associated partial correlation. The plot reveals the existence of some high degree nodes, such as "research", "data", "system", "perform", that are part of the computer science vocabulary. Further, some standard phrases in computer science, such as "home–page", "comput–scienc",

"program–languag", "data–structur", "distribut–system" and "high–perform", have high partial correlations among their constituent words in all four categories. A few subgraphs extracted from the common network are shown in Figure 4; each graph clearly has its own semantic meaning, which we loosely label "webpage generic", "research area/lab", "parallel programming" and "program development".

The model also allows us to explore the heterogeneity between different categories. As an example, we show the graphs for the "student" and "faculty" categories in Figure 5. It can be seen that terms "teach" and "assist" are only linked in the "student" category, since many graduate students are employed as teaching assistants. On the other hand, some term pairs only have links in the "faculty" category, such as "select–public" (for selected publications), "faculti–student", "assist–professor" and "associ–professor". Similarly, we illustrate the differences between the "course" and "project" categories (the former emphasizes teaching/learning and the latter research) in Figure 6. Some course-related terms are linked only in the "course" category, such as "office–hour", "office–instructor" and "teach–assist", while prominent pairs in the "project" category include "technolog–center", "technolog–institut", "research–scienc" and "research–inform". Overall, the model captures the basic common semantic structure of the websites, but also identifies meaningful differences across the various categories.

Acknowledgments

The authors thank Sijian Wang for his helpful suggestions. E. Levina's research is partially supported by NSF grant DMS-0805798. G. Michailidis's research is partially supported by NIH grant 1RC1CA145444-0110 and MEDC grant GR-687. J. Zhu's research is partially supported by NSF grants DMS-0705532 and DMS-0748389.

References

Asuncion, A. and Newman, D. (2007), "UCI machine learning repository," .

Barabasi, A.-L. and Albert, R. (1999), "Emergence of scaling in random networks," *Science*, 286, 509–512.

Bickel, P. and Levina, E. (2008), "Regularized estimation of large covariance matrices," *Annals of Statistics*, 36, 199–227.

Cardoso-Cachopo, A. (2009), `http://web.ist.utl.pt/~acardoso/datasets/`.

d'Aspremont, A., Banerjee, O., and Ghaoui, L. (2008), "First-order methods for sparse covariance selection," *SIAM Journal on matrix Analysis and its Applications*, 30, 56–66.

Dempster, A. (1972), "Covariance selection," *Biometrics*, 28, 157–175.

Drton, M. and Perlman, M. (2004), "Model selection for Gaussian concentration graphs," *Biometrika*, 91, 591–602.

Dumais, S. (1991), "Improving the retrieval of information from external source," *Behavior Research Methods, Instruments and Computers*, 23, 229–236.

Edwards, D. (2000), *Introduction to graphical modelling*, Springer, New York.

Fan, J., Feng, Y., and Wu, Y. (2009), "Network exploration via the adaptive LASSO and SCAD penalties," *Annals of Applied Statistics*, To appear.

Fan, J. and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Asscociation*, 96, 1348–1360.

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, 9, 432–441.

Kolar, M. and Xing, E. (2008), "Improved estimation of high-dimensional Ising models," in *Eprint arXiv:0811.1239*.

Lam, C. and Fan, J. (2009), "Sparsistency and rates of convergence in large covariance matrices estimation," *Annals of Statistics*, To appear.

Lauritzen, S. (1996), *Graphical Models*, Oxford University Press, Oxford.

Li, H. and Gui, J. (2006), "Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks," *Biostatistics*, 7, 302–317.

Meinshausen, N. and Buhlmann, P. (2006), "High-dimensional graphs with the lasso," *Annals of Statistics*, 34, 1436–1462.

Michailidis, G. and de Leeuw, J. (2001), "Multilevel homogeneity analysis with differential weighting," *Computational Statistics and Data Analysis*, 32, 411–442.

Pearl, J. (2000), *Causality: models, reasoning, and inference*, Cambridge University Press, Oxford.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009), "Partial Correlation Estimation by Joint Sparse Regression Model," *Journal of the American Statistical Asscociation*, 104, 735–746.

Porter, M. (1980), "An algorithm for suffix stripping," *Program*, 14, 130–137.

Ravikumar, P., Wainwright, M., and Lafferty, J. (2009), "High-dimensional Ising model selection using ℓ_1-regularized logistic regression," *Annals of Statistics*, To appear.

Rocha, G., Zhao, P., and Yu, B. (2008), "A path following algorithm for Sparse Pseudo-Likelihood Inverse Covariance Estimation (SPLICE)," Tech. rep., Department of Statistics, University of California, Berkeley.

Rothman, A., Bickel, P., Levina, E., and Zhu, J. (2008), "Sparse permutation invariant covariance estimation," *Electronic Journal of Statistics*, 2, 494–515.

Yuan, M. and Lin, Y. (2007), "Model selection and estimation in the Gaussian graphical model," *Biometrika*, 94, 19–35.

Zhou, N. and Zhu, J. (2007), "Group variable selection via a hierarchical lasso and its oracle property," Tech. rep., Department of Statistics, University of Michigan.

Zou, H. and Li, R. (2008), "One-step sparse estimates in nonconcave penalized likelihood models," *Annals of Statistics*, 36, 1108–1126.

Appendix

Proof of Lemma 2. Denote $Q_\lambda(\{\mathbf{\Omega}^{(k)}\}_{k=1}^K)$ as the criterion in (6). Suppose $\{\widehat{\mathbf{\Omega}}^{(k)}\}_{k=1}^K$ is a local minimizer of $Q_\lambda(\{\mathbf{\Omega}^{(k)}\}_{k=1}^K)$. Define $\widehat{\mathbf{\Theta}}$ and $\{\widehat{\mathbf{\Gamma}}^{(k)}\}_{k=1}^K$ as $\widehat{\theta}_{j,j'} = \sqrt{\eta \sum_{k=1}^K |\widehat{\omega}_{j,j'}^{(k)}|}$ and $\widehat{\gamma}_{j,j'}^{(k)} = \widehat{\omega}_{j,j'}^{(k)} / \sqrt{\eta \sum_{k=1}^K |\widehat{\omega}_{j,j'}^{(k)}|}$, $1 \le j < j' \le p; 1 \le k \le K$, respectively. Then we need to show that $(\widehat{\mathbf{\Theta}}, \{\widehat{\mathbf{\Gamma}}^{(k)}\}_{k=1}^K)$ is a local minimizer of (5). By the definition of a local minimizer, there exists a constant $\epsilon > 0$ such that for any $\{\mathbf{\Omega}^{(k)}\}_{k=1}^K$ satisfying $\sum_{1\le j,j'\le p} \sum_{k=1}^K |\omega_{j,j'}^{(k)} - \widehat{\omega}_{j,j'}^{(k)}| \le \epsilon$, we have $Q_\lambda(\{\mathbf{\Omega}^{(k)}\}_{k=1}^K) \ge Q_\lambda(\{\widehat{\mathbf{\Omega}}^{(k)}\}_{k=1}^K)$. Let $\mathcal{F} = \{(\mathbf{\Theta}, \{\mathbf{\Gamma}^{(k)}\}_{k=1}^K) : \|\Delta\mathbf{\Theta}\|_1 + \sum_{k=1}^K \|\Delta\mathbf{\Gamma}^{(k)}\|_1 \le \epsilon'\}$, where $\Delta\theta_{j,j'} = \theta_{j,j'} - \widehat{\theta}_{j,j'}$, $\Delta\gamma_{j,j'}^{(k)} = \gamma_{j,j'}^{(k)} - \widehat{\gamma}_{j,j'}^{(k)}$, and ϵ' is some constant satisfying $0 \le \epsilon' \le (-U + \sqrt{U^2 + 4\epsilon})/2$, where

$$U = \max\{1, 4/\lambda^2\} \max_{j\neq j'} \lambda \sqrt{\sum_{k=1}^K |\widehat{\omega}_{j,j'}^{(k)}|} + \max_{1\le j\le p} (1 + \sum_{k=1}^K |\widehat{\omega}_{j,j}^{(k)}|). \tag{11}$$

Then, for any $(\mathbf{\Theta}, \{\mathbf{\Gamma}^{(k)}\}_{k=1}^K) \in \mathcal{F}$, we have

$$\begin{aligned}
& \sum_{k=1}^K |\omega_{j,j'}^{(k)} - \widehat{\omega}_{j,j'}^{(k)}| \\
= & \sum_{k=1}^K |\theta_{j,j'}\gamma_{j,j'}^{(k)} - \widehat{\theta}_{j,j'}\widehat{\gamma}_{j,j'}^{(k)}| \\
= & \sum_{k=1}^K |(\widehat{\theta}_{j,j'} + \Delta\theta_{j,j'})(\widehat{\gamma}_{j,j'}^{(k)} + \Delta\gamma_{j,j'}^{(k)}) - \widehat{\theta}_{j,j'}\widehat{\gamma}_{j,j'}^{(k)}| \\
\leq & |\Delta\theta_{j,j'}| \sum_{k=1}^K |\widehat{\gamma}_{j,j'}^{(k)}| + |\widehat{\theta}_{j,j'}| \sum_{k=1}^K |\Delta\gamma_{j,j'}^{(k)}| + |\Delta\theta_{j,j'}| \sum_{k=1}^K |\Delta\gamma_{j,j'}^{(k)}| \qquad (12)
\end{aligned}$$

By the fact that $\|\Delta\mathbf{\Theta}\|_1 = \sum_{j\neq j'} |\Delta\theta_{j,j'}|$ and $\|\Delta\mathbf{\Gamma}^{(k)}\|_1 = \sum_{j=1}^p |\omega_{j,j}^{(k)} - \widehat{\omega}_{j,j}^{(k)}| + \sum_{j\neq j'} |\Delta\gamma_{j,j'}^{(k)}|$, we have

$$\begin{aligned}
& \sum_{1\leq j,j'\leq p} \sum_{k=1}^K |\omega_{j,j'}^{(k)} - \widehat{\omega}_{j,j'}^{(k)}| \\
\leq & (\sum_{1\leq j,j'\leq p} |\Delta\theta_{j,j'}|)(\max_{1\leq j,j'\leq p} \sum_{k=1}^K |\widehat{\gamma}_{j,j'}^{(k)}|) + (\max_{1\leq j,j'\leq p} |\widehat{\theta}_{j,j'}|)(\sum_{1\leq j,j'\leq p} \sum_{k=1}^K |\Delta\gamma_{j,j'}^{(k)}|) \\
& +(\sum_{1\leq j,j'\leq p} |\Delta\theta_{j,j'}|)(\sum_{1\leq j,j'\leq p} \sum_{k=1}^K |\Delta\gamma_{j,j'}|) \\
\leq & \max_{1\leq j,j'\leq p} (|\widehat{\theta}_{j,j'}| + \sum_{k=1}^K |\widehat{\gamma}_{j,j'}^{(k)}|)\{\|\Delta\mathbf{\Theta}\|_1 + \sum_{k=1}^K \|\Delta\mathbf{\Gamma}^{(k)}\|_1\} \\
& +\{\|\Delta\mathbf{\Theta}\|_1 + \sum_{k=1}^K \|\Delta\mathbf{\Gamma}^{(k)}\|_1\}^2
\end{aligned} \qquad (13)$$

By the construction of $\widehat{\boldsymbol{\Theta}}$ and $\widehat{\boldsymbol{\Gamma}}^{(k)}$, we know $|\widehat{\theta}_{j,j'}| + \eta \sum_{k=1}^{K} |\widehat{\gamma}_{j,j'}^{(k)}| = 2\sqrt{\eta \sum_{k=1}^{K} |\widehat{\omega}_{j,j'}^{(k)}|} = \lambda \sqrt{\sum_{k=1}^{K} |\widehat{\omega}_{j,j'}^{(k)}|}$. Then

$$\begin{aligned}
& \max_{1\le j,j'\le p} (|\widehat{\theta}_{j,j'}| + \sum_{k=1}^{K} |\widehat{\gamma}_{j,j'}^{(k)}|) \\
\le\ & \max_{j\neq j'} (|\widehat{\theta}_{j,j'}| + \sum_{k=1}^{K} |\widehat{\gamma}_{j,j'}^{(k)}|) + \max_{1\le j\le p} (1 + \sum_{k=1}^{K} |\widehat{\omega}_{j,j}^{(k)}|) \\
\le\ & \max\{1, 1/\eta\} \max_{j\neq j'} (|\widehat{\theta}_{j,j'}| + \eta \sum_{k=1}^{K} |\widehat{\gamma}_{j,j'}^{(k)}|) + \max_{1\le j\le p} (1 + \sum_{k=1}^{K} |\widehat{\omega}_{j,j}^{(k)}|) \\
=\ & \max\{1, 4/\lambda^2\} \max_{j\neq j'} \lambda \sqrt{\sum_{k=1}^{K} |\widehat{\omega}_{j,j'}^{(k)}|} + \max_{1\le j\le p} (1 + \sum_{k=1}^{K} |\widehat{\omega}_{j,j}^{(k)}|) \\
=\ & U. \qquad (14)
\end{aligned}$$

Thus, we have $\sum_{1\le j,j'\le p} \sum_{k=1}^{K} |\omega_{j,j'}^{(k)} - \widehat{\omega}_{j,j'}^{(k)}| \le U\epsilon' + (\epsilon')^2 \le \epsilon$. Consequently,

$$\begin{aligned}
Q_\eta^*(\boldsymbol{\Theta}, \{\boldsymbol{\Gamma}^{(k)}\}_{k=1}^K) &= Q_\lambda(\{\boldsymbol{\Omega}^{(k)}\}_{k=1}^K) \\
&\ge Q_\lambda(\{\widehat{\boldsymbol{\Omega}}^{(k)}\}_{k=1}^K) \\
&= Q_\eta^*(\widehat{\boldsymbol{\Theta}}, \{\widehat{\boldsymbol{\Gamma}}^{(k)}\}_{k=1}^K) \qquad (15)
\end{aligned}$$

Thus, $(\widehat{\boldsymbol{\Theta}}, \{\widehat{\boldsymbol{\Gamma}}^{(k)}\}_{k=1}^K)$ is a local minimizer of (5).

On the other hand, we can follow the proof in Theorem 1 of Zhou and Zhu (2007) to show that: if $(\widehat{\boldsymbol{\Theta}}, \{\widehat{\boldsymbol{\Gamma}}^{(k)}\}_{k=1}^K)$ is a local minimizer of (5), then there also exists a local minimizer of (6), $\{\widehat{\boldsymbol{\Omega}}^{(k)}\}_{k=1}^K$, such that $\widehat{\boldsymbol{\Omega}}^{(k)} = \widehat{\boldsymbol{\Theta}} \cdot \widehat{\boldsymbol{\Gamma}}^{(k)}$, for all $1 \le k \le K$.

□

Next, we state some results used in the proof of Theorem 1 that were established in Theorem 1 of Rothman et al. (2008). We will use the following notation: for a matrix $\boldsymbol{M} = [m_{j,j'}]_{p\times p}$, $|\boldsymbol{M}|_1 = \sum_{j,j'} |m_{j,j'}|$, $\boldsymbol{M}^+$ is a diagonal matrix with the same diagonal as $\boldsymbol{M}$, $\boldsymbol{M}^- = \boldsymbol{M} - \boldsymbol{M}^+$, and $\boldsymbol{M}_S$ is $\boldsymbol{M}$ with all elements outside an index set S replaced by zeros. We also write $\widetilde{\boldsymbol{M}}$ for the vectorized $p^2 \times 1$ form of $\boldsymbol{M}$, and $\otimes$ for the Kronecker product of two matrices.

Lemma 3 *Let* $l(\mathbf{\Omega}^{(k)}) = \text{trace}(\widehat{\mathbf{\Sigma}}^{(k)}\mathbf{\Omega}^{(k)}) - \log\{\det(\mathbf{\Omega}^{(k)})\}$. *Then for any* $1 \le k \le K$, *the following decomposition holds:*

$$l(\mathbf{\Omega}_0^{(k)}+\mathbf{\Delta}^{(k)}) - l(\mathbf{\Omega}_0^{(k)}) = \text{trace}[(\widehat{\mathbf{\Sigma}}^{(k)} - \mathbf{\Sigma}_0^{(k)})\mathbf{\Delta}^{(k)}] + (\widetilde{\mathbf{\Delta}}^{(k)})^{\intercal}[\int_0^1 (1-v)(\mathbf{\Omega}_0^{(k)} + v\mathbf{\Delta}^{(k)})^{-1} \otimes (\mathbf{\Omega}_0^{(k)} + v\mathbf{\Delta}^{(k)})^{-1} dv]\widetilde{\mathbf{\Delta}}^{(k)} \ . \quad (16)$$

Further, there exist positive constants C_1 *and* C_2 *such that with probability tending to 1*

$$\left|\text{trace}[(\widehat{\mathbf{\Sigma}}^{(k)} - \mathbf{\Sigma}_0^{(k)})\mathbf{\Delta}^{(k)}]\right| \ \le \ C_1\sqrt{\frac{\log p}{n}}|\mathbf{\Delta}^{(k)-}|_1 + C_2\sqrt{\frac{p\log p}{n}}\|\mathbf{\Delta}^{(k)+}\|_F \ , \quad (17)$$

$$(\widetilde{\mathbf{\Delta}}^{(k)})^{\intercal}\big[\int_0^1 (1-v)(\mathbf{\Omega}_0^{(k)} + v\mathbf{\Delta}^{(k)})^{-1} \otimes (\mathbf{\Omega}_0^{(k)} + v\mathbf{\Delta}^{(k)})^{-1} dv\big]\widetilde{\mathbf{\Delta}}^{(k)} \ \ge \ \frac{1}{4\tau_2^2}\|\mathbf{\Delta}^{(k)}\|_F^2 \ . \quad (18)$$

Proof of Theorem 1

In a slight abuse of notation, we will write $\mathbf{\Omega} = \{\mathbf{\Omega}^{(k)}\}_{k=1}^K$, $\mathbf{\Omega}_0 = \{\mathbf{\Omega}_0^{(k)}\}_{k=1}^K$, and $\mathbf{\Delta} = \{\mathbf{\Delta}^{(k)}\}_{k=1}^K$, where $\mathbf{\Delta}^{(k)} = (\delta_{j,j'}^{(k)})_{p\times p}$ is defined as $\mathbf{\Delta}^{(k)} = \mathbf{\Omega}^{(k)} - \mathbf{\Omega}_0^{(k)}$, $1 \le k \le K$. Let $Q(\mathbf{\Omega})$ be the objective function of (6), and let $G(\mathbf{\Delta}) = Q(\mathbf{\Omega}_0 + \mathbf{\Delta}) - Q(\mathbf{\Omega}_0)$. If we take a closed bounded convex set $\mathcal{A}$ which contains 0, and show that G is strictly positive everywhere on the boundary $\partial\mathcal{A}$, then it implies that G has a local minimum inside $\mathcal{A}$, since G is continuous and $G(\mathbf{0}) = 0$. Specifically, we define $\mathcal{A} = \{\mathbf{\Delta} : \sum_{k=1}^K \|\mathbf{\Delta}^{(k)}\|_F \le Mr_n\}$, with boundary $\partial\mathcal{A} = \{\mathbf{\Delta} : \sum_{k=1}^K \|\mathbf{\Delta}^{(k)}\|_F = Mr_n\}$, where M is a positive constant and $r_n = \sqrt{(p+q)(\log p)/n}$.

By Lemma 3, we can write $G(\mathbf{\Delta}) = I_1 + I_2 + I_3 + I_4$, where

$$
\begin{aligned}
I_1 &= \sum_{k=1}^{K} \mathrm{trace}[(\widehat{\mathbf{\Sigma}}^{(k)} - \mathbf{\Sigma}_0^{(k)})\mathbf{\Delta}^{(k)}] && (19)\\
I_2 &= \sum_{k=1}^{K} \widetilde{\mathbf{\Delta}}^{(k)\mathrm{T}} [\int_0^1 (1-v)(\Omega_0^{(k)} + v\mathbf{\Delta}^{(k)})^{-1} \otimes (\Omega_0^{(k)} + v\mathbf{\Delta}^{(k)})^{-1} dv] \widetilde{\mathbf{\Delta}}^{(k)} && (20)\\
I_3 &= \lambda \sum_{(j,j')\in T^c} \sqrt{\sum_{k=1}^{K} |\delta_{j,j'}^{(k)}|} && (21)\\
I_4 &= \lambda \sum_{j\neq j':(j,j')\in T} \Big(\sqrt{\sum_{k=1}^{K} |\omega_{j,j'}^{(k)}|} - \sqrt{\sum_{k=1}^{K} |\omega_{0,j,j'}^{(k)}|} \Big) && (22)
\end{aligned}
$$

We first consider I_1. By applying inequality (17) in Lemma 3, we have

$$
I_1 \leq C_1 \sqrt{\frac{\log p}{n}} \sum_{k=1}^{K} |\mathbf{\Delta}^{(k)-}|_1 + C_2 \sqrt{\frac{p \log p}{n}} \sum_{k=1}^{K} \|\mathbf{\Delta}^{(k)+}\|_F = I_{1,1} + I_{1,2} \ , \tag{23}
$$

where

$$
\begin{aligned}
I_{1,1} &= C_1 \sqrt{\frac{\log p}{n}} \sum_{k=1}^{K} |\mathbf{\Delta}_T^{(k)-}|_1 + C_2 \sqrt{\frac{p \log p}{n}} \sum_{k=1}^{K} \|\mathbf{\Delta}^{(k)+}\|_F \ , \\
I_{1,2} &= C_1 \sqrt{\frac{\log p}{n}} \sum_{k=1}^{K} |\mathbf{\Delta}_{T^c}^{(k)-}|_1 \ . && (24)
\end{aligned}
$$

By applying the bound $|\mathbf{\Delta}_T^{(k)-}|_1 \leq \sqrt{q_k} \|\mathbf{\Delta}_T^{(k)-}\|_F$, we have

$$
\begin{aligned}
|I_{1,1}| &\leq C_1 \sqrt{\frac{q \log p}{n}} \sum_{k=1}^{K} \|\mathbf{\Delta}_T^{(k)-}\|_F + C_2 \sqrt{\frac{p \log p}{n}} \sum_{k=1}^{K} \|\mathbf{\Delta}^{(k)+}\|_F \\
&\leq (C_1 + C_2) \sqrt{\frac{(p+q)\log p}{n}} \sum_{k=1}^{K} \|\mathbf{\Delta}^{(k)}\|_F = M(C_1 + C_2) \frac{(p+q)\log p}{n} && (25)
\end{aligned}
$$

on the boundary $\partial\mathcal{A}$.

Next, since for r_n small enough we have $I_3 \geq \lambda \sum_{k=1}^{K} |\mathbf{\Delta}_{T^c}^{(k)-}|_1$, the term $I_{1,2}$ is dominated by the positive term I_3:

$$
\begin{aligned}
I_3 + I_{1,2} &\geq \lambda \sum_{k=1}^{K} |\mathbf{\Delta}_{T^c}^{(k)-}|_1 - C_1 \sqrt{\frac{\log p}{n}} \sum_{k=1}^{K} |\mathbf{\Delta}_{T^c}^{(k)-}|_1 \\
&\geq (\Lambda_1 - C_1) \sqrt{\frac{\log p}{n}} \sum_{k=1}^{K} |\mathbf{\Delta}_{T^c}^{(k)-}|_1 && (26)
\end{aligned}
$$

The last inequality uses the condition $\lambda \geq \Lambda_1\sqrt{(\log p)/n}$. Therefore, $I_3 + I_{1,2} \geq 0$ when Λ_1 is large enough. Next we consider I_2. By Lemma 3 (iii), we have

$$I_2 \geq \frac{1}{4\tau_2^2}\sum_{k=1}^{K}\|\mathbf{\Delta}^{(k)}\|_F^2 \geq \frac{M^2}{8\tau_2^2}\,\frac{(p+q)\log p}{n} \tag{27}$$

Finally consider the remaining term I_4. Using condition (B), we have

$$\begin{aligned}
|I_4| &\leq \lambda \sum_{j\neq j':(j,j')\in T}\Big|\Big(\sum_{k=1}^{K}|\omega_{j,j'}^{(k)}|\Big)^{1/2} - \Big(\sum_{k=1}^{K}|\omega_{0,j,j'}^{(k)}|\Big)^{1/2}\Big| \\
&\leq \lambda \sum_{j\neq j':(j,j')\in T}\frac{\sum_{k=1}^{K}\Big||\omega_{j,j'}^{(k)}| - |\omega_{0,j,j'}^{(k)}|\Big|}{\big(\sum_{k=1}^{K}|\omega_{j,j'}^{(k)}|\big)^{1/2} + \big(\sum_{k=1}^{K}|\omega_{0,j,j'}^{(k)}|\big)^{1/2}} \\
&\leq \frac{\lambda}{\sqrt{\tau_3}}\sum_{k=1}^{K}\sum_{j\neq j':(j,j')\in T}|\omega_{j,j'}^{(k)} - \omega_{0,j,j'}^{(k)}| \\
&\leq \frac{\lambda}{\sqrt{\tau_3}}\sqrt{q}\sum_{k=1}^{K}\|\mathbf{\Delta}^{(k)}\|_F \\
&\leq \frac{M\Lambda_2}{\sqrt{\tau_3}}\frac{(p+q)(\log p)}{n}\,.
\end{aligned} \tag{28}$$

The last inequality uses the condition $\lambda \leq \Lambda_2\sqrt{[(p+q)(\log p)]/(nq)}$. Putting everything together and using $I_2 > 0$ and $I_3 + I_{1,2} > 0$, we have

$$\begin{aligned}
G(\mathbf{\Delta}) &\geq I_2 - I_{1,1} - I_4 \\
&\geq \frac{M^2}{8\tau_2^2}\frac{(p+q)\log p}{n} - M(C_1+C_2)\frac{(p+q)\log p}{n} - \frac{M\Lambda_2}{\sqrt{\tau_3}}\frac{(p+q)\log p}{n} \\
&= M^2\frac{(p+q)\log p}{n}\Big(\frac{1}{8\tau_2^2} - \frac{C_1+C_2+\Lambda_2/\sqrt{\tau_3}}{M}\Big)\,.
\end{aligned} \tag{29}$$

Thus for M sufficiently large, we have $G(\mathbf{\Delta}) > 0$ for any $\mathbf{\Delta} \in \partial\mathcal{A}$.

□

Proof of Theorem 2

It suffices to show that for all $(j,j') \in T_k^c$, $1 \leq k \leq K$, the derivative $\partial Q/\partial\omega_{j,j'}^{(k)}$ at $\widehat{\omega}_{j,j'}^{(k)}$ has the same sign as $\widehat{\omega}_{j,j'}^{(k)}$ with probability tending to 1. To see that, suppose that for some

$(j, j') \in T_k^c$, the estimate $\widehat{\omega}_{j,j'}^{(k)} \neq 0$. Without loss of generality, suppose $\widehat{\omega}_{j,j'}^{(k)} > 0$. Then there exists $\xi > 0$ such that $\widehat{\omega}_{j,j'}^{(k)} - \xi > 0$. Since $\widehat{\mathbf{\Omega}}$ is a local minimizer of $Q(\mathbf{\Omega})$, we have $\partial Q / \partial \omega_{j,j'}^{(k)} < 0$ at $\widehat{\omega}_{j,j'}^{(k)} - \xi$ for ξ small, contradicting the claim $\partial Q / \partial \omega_{j,j'}^{(k)}$ has the same sign as $\omega_{j,j'}^{(k)}$.

The derivative of the objective function can be written as

$$\frac{\partial Q}{\partial \omega_{j,j'}} = 2(W_1 + W_2 \mathrm{sgn}(\omega_{j,j'}^{(k)})) \ , \tag{30}$$

where $W_1 = \widehat{\sigma}_{j,j'}^{(k)} - \sigma_{j,j'}^{(k)}$ and $W_2 = \lambda / \sqrt{\sum_{k=1}^K |\omega_{j,j'}^{(k)}|}$. Arguing as in Theorem 2 of Lam and Fan (2009), one can show that $W_1 = O([(\log p)/n]^{1/2} + \eta_n^{1/2})$. On the other hand, by Theorem 1, we have $\sum_{k=1}^K |\omega_{j,j'}^{(k)} - \omega_{0,j,j'}^{(k)}| \leq \sum_{k=1}^K \|\mathbf{\Omega}^{(k)} - \mathbf{\Omega}_0^{(k)}\|_F = O(\eta_n) = o(1)$. Then for any $\epsilon > 0$ and large enough n we have $\sum_{k=1}^K |\omega_{j,j'}^{(k)}| \leq \sum_{k=1}^K |\omega_{0,j,j'}^{(k)}| + \epsilon$. Then we have $|W_2| \geq \lambda / (1 + \sum_{k=1}^K |\omega_{0,j,j'}^{(k)}|)$. By assumption, $[(\log p)/n]^{1/2} + \eta_n^{1/2} = O(\lambda)$, and thus the term W_2 dominates W_1 in (30). Therefore,

$$\mathrm{sgn}\left(\frac{\partial Q}{\partial \omega_{j,j'}^{(k)}}\Big|_{\omega_{j,j'}^{(k)} = \widehat{\omega}_{j,j'}^{(k)}}\right) = \mathrm{sgn}(\widehat{\omega}_{j,j'}^{(k)}) \ .$$

□

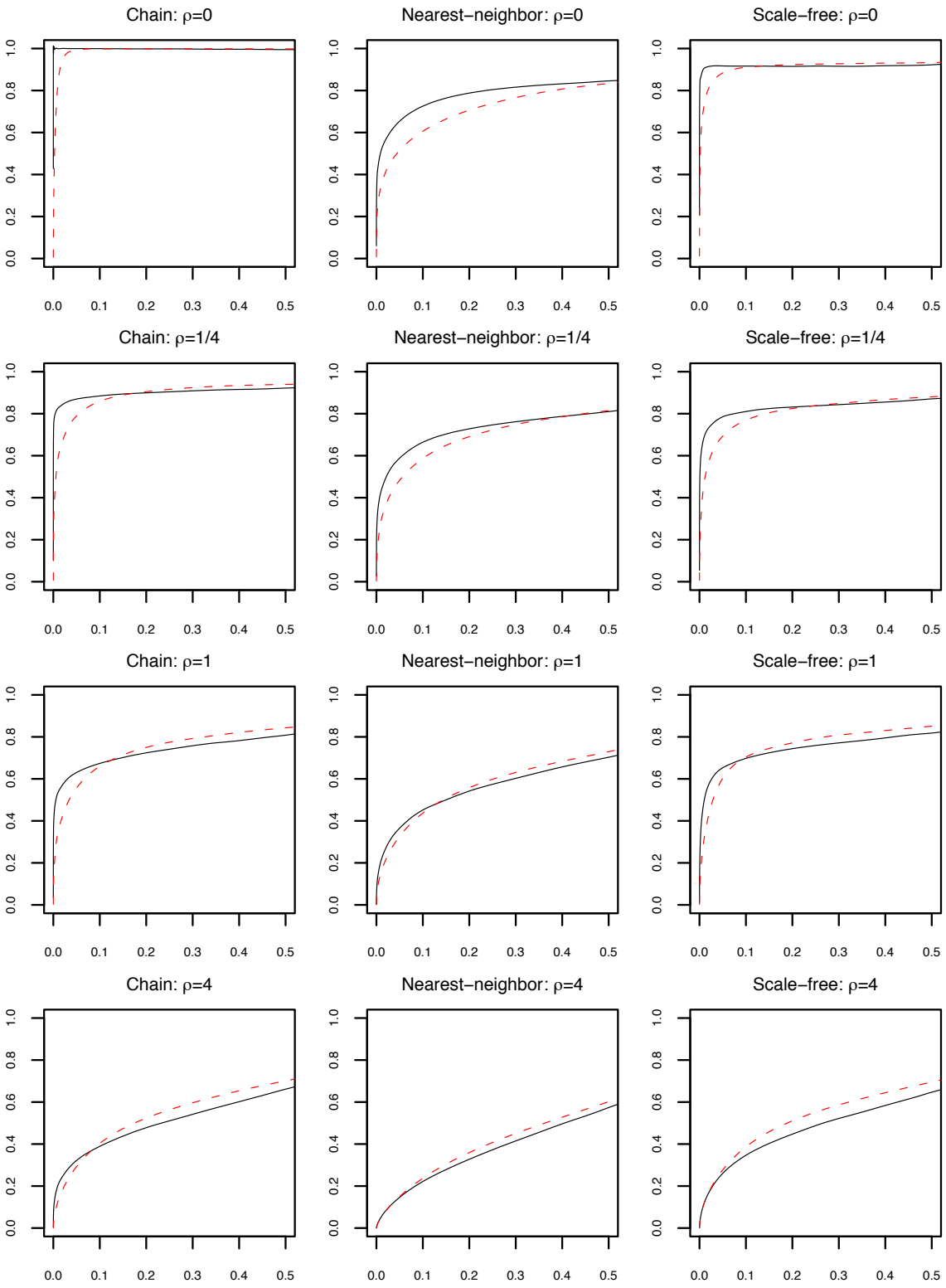


Figure 2: ROC curves. The solid line corresponds to the joint estimation method, and the dashed line to the separate estimation method.

Common Structure

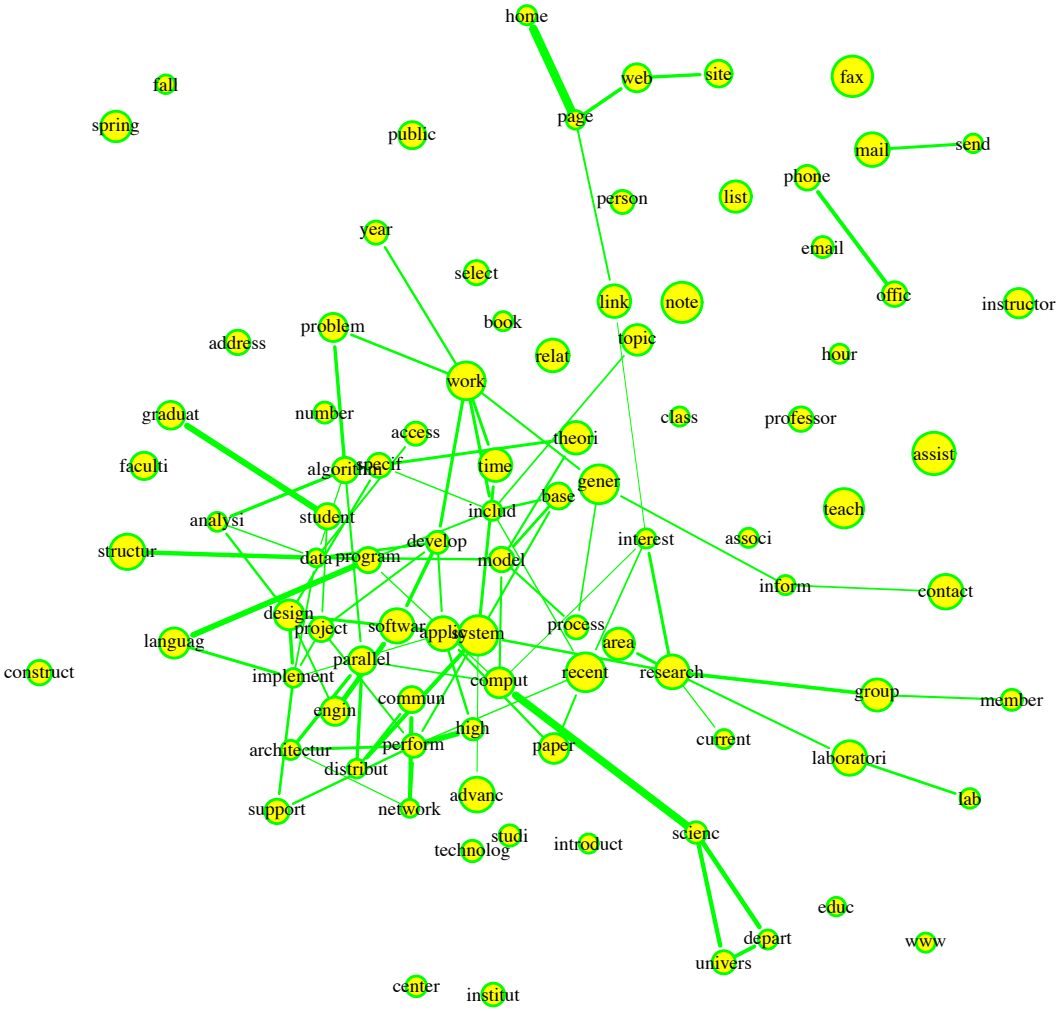


Figure 3: The common structure in the estimated graph. The nodes represent the 100 terms with highest log-entropy weights. The areas of the yellow circles are proportional to the degrees of the associated nodes in the common structure. The width of each edge is proportional to the magnitude of the associated partial correlation.

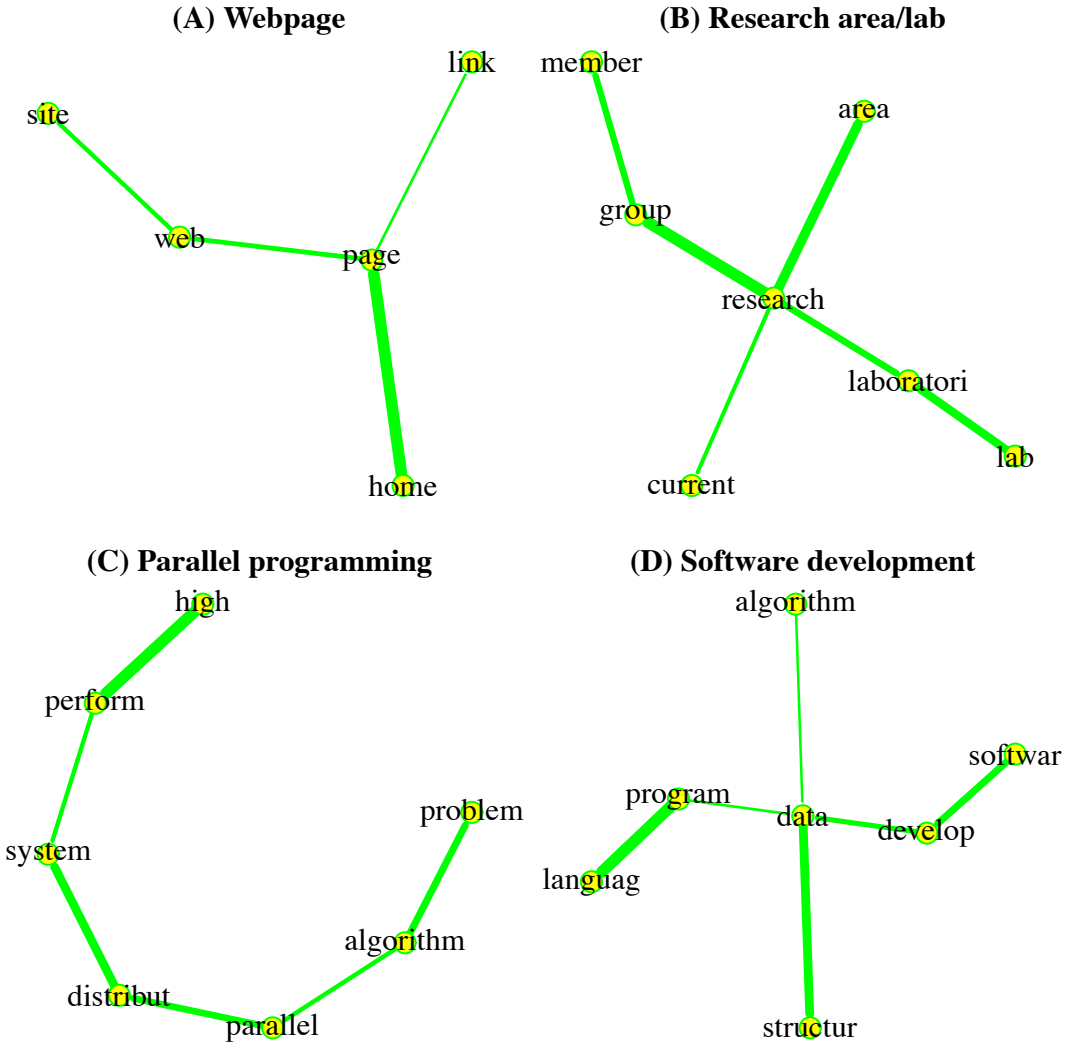
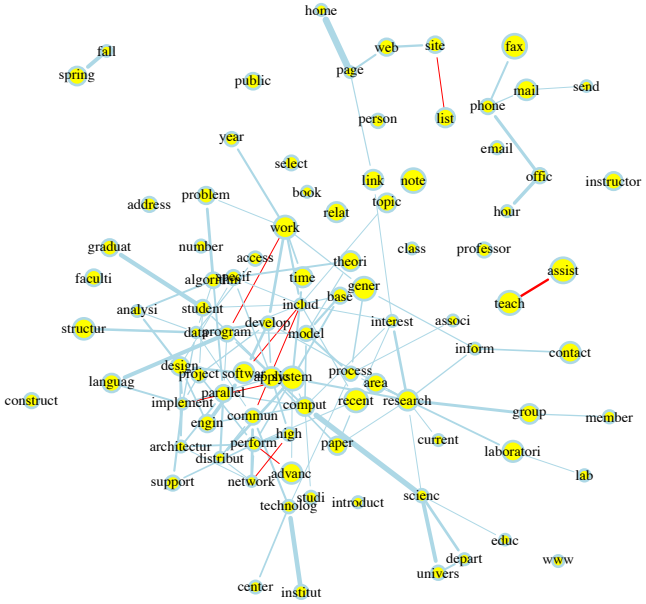


Figure 4: Subgraphs extracted from the common structure in Figure 3.

(A) Student



(B) Faculty

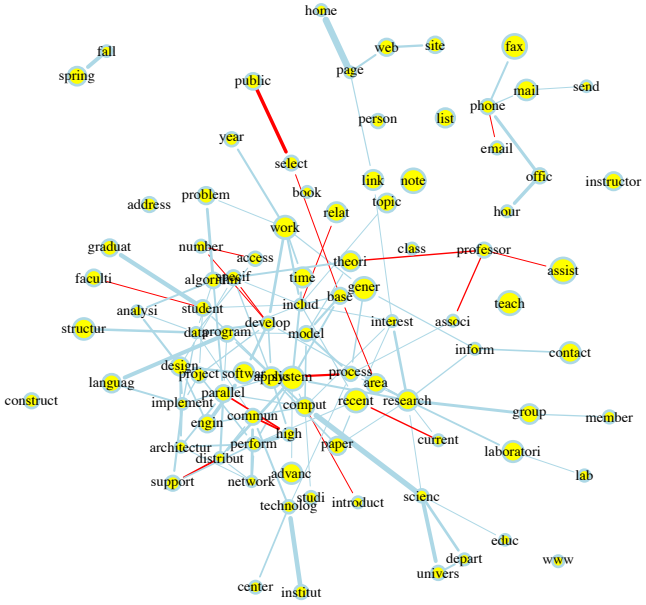
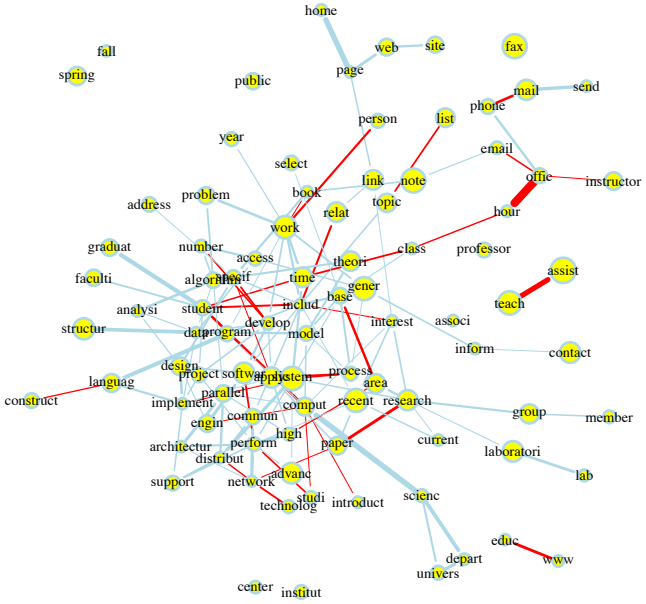


Figure 5: "Student" and "Faculty" graphs. The light blue lines are the links appearing in both categories, and the red lines are the links only appearing in one category.

(A) Course



(B) Project

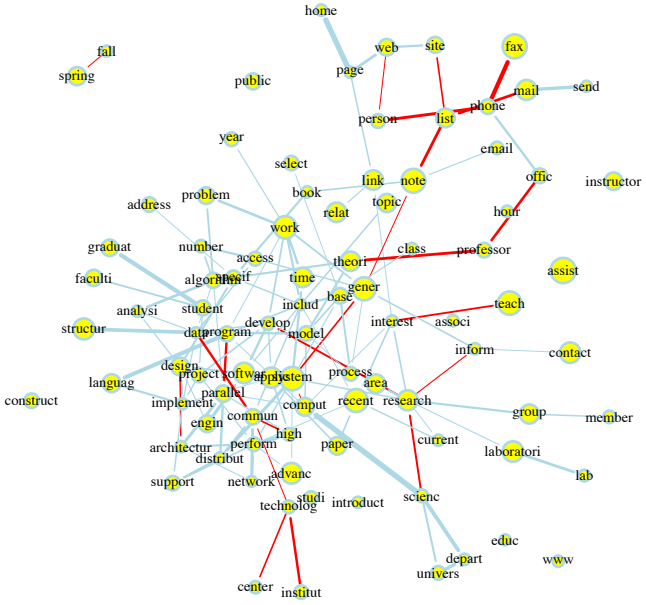


Figure 6: "Course" and "Project" graphs. The light blue lines are the links appearing in both categories, and the red lines are the links only appearing in one category.