Biometrika (2011), **98**, 1, *pp.* 1–15

Printed in Great Britain

doi: 10.1093/biomet/asq060
Advance Access publication 9 February 2011

Joint estimation of multiple graphical models

By JIAN GUO, ELIZAVETA LEVINA, GEORGE MICHAILIDIS and JI ZHU

Department of Statistics, University of Michigan, 1085 South University, Ann Arbor, Michigan 48109-1107, U.S.A.

guojian@umich.edu elevina@umich.edu gmichail@umich.edu jizhu@umich.edu

Summary

Gaussian graphical models explore dependence relationships between random variables, through the estimation of the corresponding inverse covariance matrices. In this paper we develop an estimator for such models appropriate for data from several graphical models that share the same variables and some of the dependence structure. In this setting, estimating a single graphical model would mask the underlying heterogeneity, while estimating separate models for each category does not take advantage of the common structure. We propose a method that jointly estimates the graphical models corresponding to the different categories present in the data, aiming 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 inverse covariance matrices across categories. We establish the asymptotic consistency and sparsity of the proposed estimator in the high-dimensional case, and illustrate its performance on a number of simulated networks. An application to learning semantic connections between terms from webpages collected from computer science departments is included.

Some key words: Covariance matrix; Graphical model; Hierarchical penalty; High-dimensional data; Network.

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 much attention both in the statistical and computer science literature; see, for example, the books by Lauritzen (1996) and Pearl (2009). These models 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, which correspond to a directed acyclic graph, and Gaussian models, which assume the joint distribution to be Gaussian. In the latter case, because the distribution is 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 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, if and only if the (j, j')th element in the inverse covariance matrix is zero; thus the problem of estimating a Gaussian graphical model is equivalent to estimating an inverse covariance matrix.

The literature on estimating an inverse covariance matrix goes back to Dempster (1972), who advocated the estimation of a sparse dependence structure, i.e., setting some elements of the inverse covariance 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 by removing the least significant edges from a fully connected graph, and continues removing edges until all remaining edges 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 & Perlman (2004).

More recently, the focus has shifted to using regularization for sparse estimation of the inverse covariance matrix and the corresponding graphical model. For example, Meinshausen & Bühlmann (2006) proposed to select edges for each node in the graph by regressing the variable on all other variables using ℓ_1-penalized regression. 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 & Lin (2007), Banerjee et al. (2008), 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 inverse covariance matrix, which can then be used to infer a graph, and has been referred to as the graphical lasso (Friedman et al., 2008) or sparse permutation invariant covariance estimator (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 $O_{\mathrm{p}}[\{q(\log p)/n\}^{1/2}]$, where q is the total number of nonzero elements in the precision matrix. Fan et al. (2009) and Lam & Fan (2009) extended this penalized maximum likelihood approach to general nonconvex penalties, such as the smoothly clipped absolute deviation penalty (Fan & Li, 2001), while Lam & Fan (2009) also established a so-called 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 pseudolikelihood instead of the likelihood have been recently proposed by G. V. Rocha, P. Zhao, and B. Yu, in a 2008 unpublished preprint, arXiv:0811.1239, 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. By heterogeneous data we mean data from several categories that share the same variables but differ in their dependence structure, with some edges common across all categories and other edges unique to each category. 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 §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 compared to 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 M. Kolar and E. P. Xing in a 2008 unpublished preprint, arXiv:0811.1239, Hoefling & Tibshirani (2009) and Ravikumar et al. (2009).

2. Methodology

2·1. *Problem set-up*

Suppose we have a heterogeneous dataset with p variables and K categories. The kth category contains n_k observations $(x_1^{(k)}, \ldots, x_{n_k}^{(k)})^{\mathrm{T}}$, where each $x_i^{(k)} = (x_{i,1}^{(k)}, \ldots, x_{i,p}^{(k)})$ is a p-dimensional row vector. Without loss of generality, we assume that the observations in the same category are centred along each variable, i.e., $\sum_{i=1}^{n_k} x_{i,j}^{(k)} = 0$ for all $j = 1, \ldots, p$ and $k = 1, \ldots, K$. We further assume that $x_1^{(k)}, \ldots, x_{n_k}^{(k)}$ are an independent and identically distributed sample from a p-variate Gaussian distribution with mean zero, without loss of generality since the data are centred, and covariance matrix $\Sigma^{(k)}$. Let $\Omega^{(k)} = (\Sigma^{(k)})^{-1} = (\omega_{j,j'}^{(k)})_{p\times p}$. The loglikelihood of the observations in the kth category is

$$l(\Omega^{(k)}) = -\frac{n_k}{2}\log(2\pi) + \frac{n_k}{2}[\log\{\det(\Omega^{(k)})\} - \mathrm{tr}(\hat{\Sigma}^{(k)}\Omega^{(k)})],$$

where $\hat{\Sigma}^{(k)}$ is the sample covariance matrix for the kth category, and $\det(\cdot)$ and $\mathrm{tr}(\cdot)$ are the determinant and the trace of a matrix, respectively.

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

$$\min_{\Omega^{(k)}} \mathrm{tr}(\hat{\Sigma}^{(k)}\Omega^{(k)}) - \log\{\det(\Omega^{(k)})\} + \lambda_k \sum_{j\neq j'} |\omega_{j,j'}^{(k)}|, \tag{1}$$

where the minimum is taken over symmetric positive definite matrices. The ℓ_1 penalty shrinks some of the off-diagonal elements in $\Omega^{(k)}$ to zero and the tuning parameter λ_k controls the degree of the sparsity in the estimated inverse covariance matrix. Problem (1) can be efficiently solved by existing algorithms such as the graphical lasso (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.

2·2. *The joint estimation method*

To improve estimation when graphical models for different categories may share some common structure, we propose a joint estimation method. First, we reparameterize each off-diagonal element $\omega_{j,j'}^{(k)}$ as $\omega_{j,j'}^{(k)} = \theta_{j,j'}\gamma_{j,j'}^{(k)}$ $(1 \leqslant j \neq j' \leqslant p;\ k = 1, \ldots, K)$. An analogous parameterization in a dimension reduction setting was used in Michailidis & de Leeuw (2001). To avoid sign ambiguity between θ and γ, we restrict $\theta_{j,j'} \geqslant 0$, $1 \leqslant j \neq j' \leqslant p$. To preserve symmetry, we require that $\theta_{j,j'} = \theta_{j',j}$ and $\gamma_{j,j'}^{(k)} = \gamma_{j',j}^{(k)}$ $(1 \leqslant j \neq j' \leqslant p;\ k = 1, \ldots, K)$. For all diagonal elements, we also require $\theta_{j,j} = 1$ and $\gamma_{j,j}^{(k)} = \omega_{j,j}^{(k)}$ $(j = 1, \ldots, p; k = 1, \ldots, K)$. 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 $\Theta = (\theta_{j,j'})_{p\times p}$ and $\Gamma^{(k)} = (\gamma_{j,j'}^{(k)})_{p\times p}$. To estimate this model, we propose the following penalized criterion subject to all constraints mentioned above:

$$\min_{\Theta,(\Gamma^{(k)})_{k=1}^K} \sum_{k=1}^{K}[\mathrm{tr}(\hat{\Sigma}^{(k)}\Omega^{(k)}) - \log\{\det(\Omega^{(k)})\}] + \eta_1 \sum_{j\neq j'} \theta_{j,j'} + \eta_2 \sum_{j\neq j'}\sum_{k=1}^{K} |\gamma_{j,j'}^{(k)}|, \tag{2}$$

where η_1 and η_2 are two tuning parameters. The first, η_1, controls the sparsity of the common factors $\theta_{j,j'}$ and can effectively identify the common zero elements across $\Omega^{(1)}, \ldots, \Omega^{(K)}$; 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 $\gamma_{j,j'}^{(k)}$, and hence some of the $\omega_{j,j'}^{(k)}$, can still be set to zero by the second penalty. This allows graphs belonging to different categories to have different structures. This decomposition has also been used by N. Zhou and J. Zhu in a 2007 unpublished preprint, arXiv:1006.2871, for group variable selection in regression problems.

Criterion (2) involves two tuning parameters η_1 and η_2; it turns out that this could be reduced to an equivalent problem with a single tuning parameter. Specifically, consider

$$\min_{\Theta,(\Gamma^{(k)})_{k=1}^K} \sum_{k=1}^K [\mathrm{tr}(\hat{\Sigma}^{(k)}\Omega^{(k)}) - \log\{\det(\Omega^{(k)})\}] + \sum_{j\neq j'} \theta_{j,j'} + \eta \sum_{j\neq j'} \sum_{k=1}^K |\gamma_{j,j'}^{(k)}|, \tag{3}$$

where $\eta = \eta_1\eta_2$. For two matrices A and B of the same size, we denote their Schur–Hadamard product by $A \cdot B$. Criteria (2) and (3) are equivalent in the following sense.

LEMMA 1. *Let* $\{\hat{\Theta}^*, (\hat{\Gamma}^{(k)*})_{k=1}^K\}$ *be a local minimizer of criterion* (3). *Then, there exists a local minimizer of criterion* (2), *denoted as* $\{\hat{\Theta}^{**}, (\hat{\Gamma}^{(k)**})_{k=1}^K\}$, *such that* $\hat{\Theta}^{**} \cdot \hat{\Gamma}^{(k)**} = \hat{\Theta}^* \cdot \hat{\Gamma}^{(k)*}$ *for all* $k = 1, \ldots, K$. *Similarly, if* $\{\hat{\Theta}^{**}, (\hat{\Gamma}^{(k)**})_{k=1}^K\}$ *is a local minimizer of criterion* (2), *then there exists a local minimizer of criterion* (3), *denoted as* $\{\hat{\Theta}^*, (\hat{\Gamma}^{(k)*})_{k=1}^K\}$, *such that* $\hat{\Theta}^{**} \cdot \hat{\Gamma}^{(k)**} = \hat{\Theta}^* \cdot \hat{\Gamma}^{(k)*}$ *for all* $k = 1, \ldots, K$.

The proof follows closely the proof of the lemma in Zhou and Zhu's unpublished 2007 preprint, 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.

2·3. *The algorithm*

First we reformulate the problem (3) in a more convenient form for computational purposes.

LEMMA 2. *Let* $(\hat{\Omega}^{(k)})_{k=1}^K$ *be a local minimizer of*

$$\min_{(\Omega^{(k)})_{k=1}^K} \sum_{k=1}^K [\mathrm{tr}(\hat{\Sigma}^{(k)}\Omega^{(k)}) - \log\{\det(\Omega^{(k)})\}] + \lambda \sum_{j\neq j'} \left(\sum_{k=1}^K |\omega_{j,j'}^{(k)}| \right)^{1/2}, \tag{4}$$

where $\lambda = 2\eta^{1/2}$. *Then, there exists a local minimizer of* (3), $\{\hat{\Theta}, (\hat{\Gamma}^{(k)})_{k=1}^K\}$, *such that* $\hat{\Omega}^{(k)} = \hat{\Theta} \cdot \hat{\Gamma}^{(k)}$, *for all* $k = 1, \ldots, K$. *On the other hand, if* $\{\hat{\Theta}, (\hat{\Gamma}^{(k)})_{k=1}^K\}$ *is a local minimizer of* (3), *then there also exists a local minimizer of* (4), $(\hat{\Omega}^{(k)})_{k=1}^K$, *such that* $\hat{\Omega}^{(k)} = \hat{\Theta} \cdot \hat{\Gamma}^{(k)}$, *for all* $k = 1, \ldots, K$.

The proof follows closely the proof of the lemma in Zhou and Zhu's unpublished 2007 preprint, and is omitted. To optimize (4) we use an iterative approach based on local linear approximation (Zou & Li, 2008). Specifically, letting $(\omega_{j,j'}^{(k)})^{(t)}$ denote the estimates from the previous iteration t, we approximate $(\sum_{k=1}^K |\omega_{j,j'}^{(k)}|)^{1/2} \sim \sum_{k=1}^K |\omega_{j,j'}^{(k)}| / \{\sum_{k=1}^K |(\omega_{j,j'}^{(k)})^{(t)}|\}^{1/2}$. Thus, at the $(t+1)$th iteration, problem (4) is decomposed into K individual optimization problems:

$$(\Omega^{(k)})^{(t+1)} = \arg\min_{\Omega^{(k)}} [\mathrm{tr}(\hat{\Sigma}^{(k)}\Omega^{(k)}) - \log\{\det(\Omega^{(k)})\}] + \lambda \sum_{j\neq j'} \tau_{j,j'}^{(k)} |\omega_{j,j'}^{(k)}|, \tag{5}$$

where $\tau_{j,j'}^{(k)} = \{\sum_{k=1}^{K} |(\omega_{j,j'}^{(k)})^{(t)}|\}^{-1/2}$. Criterion (5) is exactly the sparse inverse covariance matrix estimation problem with weighted ℓ_1 penalty; the solution can be efficiently computed using the graphical lasso algorithm of Friedman et al. (2008). For numerical stability, we threshold $\{\sum_{k=1}^{K} |(\omega_{j,j'}^{(k)})^{(t)}|\}^{1/2}$ at 10^{-10}. In summary, the proposed algorithm for solving (4) is:

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

Step 1. Update $\hat{\Omega}^{(k)}$ by (5) for all $k = 1, \ldots, K$ using graphical lasso.

Step 2. Repeat Step 1 until convergence is achieved.

2·4. *Model selection*

The tuning parameter λ in (4) controls the sparsity of the resulting estimator. It can be selected either by some type of Bayesian information criterion or through crossvalidation. The former balances the goodness of fit of the model and its complexity, while the latter seeks to optimize its predictive power. Specifically, for the proposed joint estimation method we define

$$\text{BIC}(\lambda) = \sum_{k=1}^{K} [\text{tr}(\hat{\Sigma}^{(k)} \hat{\Omega}_{\lambda}^{(k)}) - \log\{\det(\hat{\Omega}_{\lambda}^{(k)})\} + \text{df}_k \log(n_k)],$$

where $\hat{\Omega}_{\lambda}^{(1)}, \ldots, \hat{\Omega}_{\lambda}^{(K)}$ are the estimates from (4) with tuning parameter λ and the degrees of freedom are defined as $\text{df}_k = \#\{(j, j') : j < j', \hat{\omega}_{j,j'}^{(k)} \neq 0\}$. An analogous definition of the degrees of freedom for the lasso has been proposed by Zou et al. (2007).

The crossvalidation method randomly splits the dataset into D segments of equal size. For the kth category, we denote the sample covariance matrix using the data in the dth segment $(d = 1, \ldots, D)$ by $\hat{\Sigma}^{(k,d)}$ and the inverse covariance matrix estimated using all the data excluding those in the dth segment and the tuning parameter λ by $\hat{\Omega}_{\lambda}^{(k,-d)}$. Then we choose λ that minimizes the average predictive negative loglikelihood as follows:

$$\text{CV}(\lambda) = \sum_{d=1}^{D} \sum_{k=1}^{K} [\text{tr}(\hat{\Sigma}^{(k,d)} \hat{\Omega}_{\lambda}^{(k,-d)}) - \log\{\det(\hat{\Omega}_{\lambda}^{(k,-d)})\}].$$

Crossvalidation can in general be expected to be more accurate than the heuristic BIC; it is much more computationally intensive however, which is why we consider both options. We provide some comparisons between the tuning parameter selection methods in §4.

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 zero at a certain rate. First, we introduce the necessary notation and state certain regularity conditions on the true precision matrices $(\Omega_0^{(1)}, \ldots, \Omega_0^{(K)})$, where $\Omega_0^{(k)} = (\omega_{0,j,j'}^{(k)})_{p \times p}$ $(k = 1, \ldots, K)$.

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 $\Omega^{(k)}$, 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. In general, T_k and q_k depend on p. In addition, let $\|\cdot\|_F$ and $\|\cdot\|$ be the Frobenius norm and the 2-norm of matrices, respectively. We assume that the following regularity conditions hold.

Condition 1. There exist constants τ_1, τ_2 such that for all $p \geqslant 1$ and $k = 1, \ldots, K$, $0 < \tau_1 < \phi_{\min}(\Omega_0^{(k)}) \leqslant \phi_{\max}(\Omega_0^{(k)}) < \tau_2 < \infty$, where $\phi_{\min}$ and $\phi_{\max}$ indicate the minimal and maximal eigenvalues.

Condition 2. There exists a constant $\tau_3 > 0$ such that $\min_{k=1,\ldots,K} \min_{(j,j') \in T_k} |\omega_{0,j,j'}^{(k)}| \geqslant \tau_3$.

Condition 1 is standard, and is also used in Bickel & Levina (2008) and Rothman et al. (2008), which guarantees that the inverse exists and is well conditioned. Condition 2 ensures that nonzero elements are bounded away from zero.

THEOREM 1 (CONSISTENCY). *Suppose Conditions* 1 *and* 2 *hold,* $(p+q)(\log p)/n = o(1)$ *and* $\Lambda_1\{(\log p)/n\}^{1/2} \leqslant \lambda \leqslant \Lambda_2\{(1 + p/q)(\log p)/n\}^{1/2}$ *for some positive constants* Λ_1 *and* Λ_2. *Then there exists a local minimizer* $(\hat{\Omega}^{(k)})_{k=1}^K$ *of* (4), *such that*

$$\sum_{k=1}^{K} \|\hat{\Omega}^{(k)} - \Omega_0^{(k)}\|_F = O_{\mathrm{p}}\left[\left\{\frac{(p+q)\log p}{n}\right\}^{1/2}\right].$$

THEOREM 2 (SPARSISTENCY). *Suppose all conditions in Theorem* 1 *hold. We further assume* $\sum_{k=1}^K \|\hat{\Omega}^{(k)} - \Omega_0^{(k)}\|^2 = O_{\mathrm{p}}(\eta_n)$, *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* $(\hat{\Omega}^{(k)})_{k=1}^K$ *in Theorem* 1 *satisfies* $\hat{\omega}_{j,j'}^{(k)} = 0$ *for all* $(j, j') \in T_k^c$, $k = 1, \ldots, K$.

This theorem is analogous to Theorem 2 of Lam & Fan (2009). The consistency requires both an upper and a lower bound on λ, whereas sparsistency requires consistency and an additional lower bound on λ. To make the bounds compatible, we require $\{(\log p)/n\}^{1/2} + \eta_n^{1/2} = O[\{(1 + p/q)(\log p)/n\}^{1/2}]$. Since η_n is the rate of convergence in the operator norm, we can bound it using the fact that $\|M\|_F^2/p \leqslant \|M\|^2 \leqslant \|M\|_F^2$. This leads to two extreme cases. In the worst-case scenario, $\sum_k \|\hat{\Omega}^{(k)} - \Omega_0^{(k)}\|$ has the same rate as $\sum_k \|\hat{\Omega}^{(k)} - \Omega_0^{(k)}\|_F$ and thus $\eta_n = O\{(p+q)(\log p)/n\}$. The two bounds are compatible only when $q = O(1)$. In the best-case scenario, $\sum_k \|\hat{\Omega}^{(k)} - \Omega_0^{(k)}\|$ has the same rate as $\sum_k \|\hat{\Omega}^{(k)} - \Omega_0^{(k)}\|_F/p^{1/2}$. 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

4·1. *Simulation settings*

In this section, we assess the performance of the joint estimation method on three types of simulated networks: a chain, a nearest-neighbour and a scale-free network. In all cases, we set $p = 100$ and $K = 3$. For each $k = 1, \ldots, K$, we generate $n_k = 100$ independently and identically distributed observations from a multivariate normal distribution $N\{0, (\Omega^{(k)})^{-1}\}$, where $\Omega^{(k)}$ is the inverse covariance matrix of the kth category. The details of the three simulated examples are as follows.

In the first example, we follow the simulation set-up in Fan et al. (2009) to generate a chain network, which corresponds to a tridiagonal inverse covariance matrix. The covariance matrices $\Sigma^{(k)}$ 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 $s_j - s_{j-1} \sim \mathrm{Un}(0{\cdot}5, 1)$ $(j = 2, \ldots, p)$.

Further, let $\Omega^{(k)} = (\Sigma^{(k)})^{-1}$. The K precision matrices generated by this procedure share the same pattern of zeros, i.e. the common structure, but the values of their nonzero off-diagonal elements may be different. The left panel of Fig. 1 shows the common link structure across

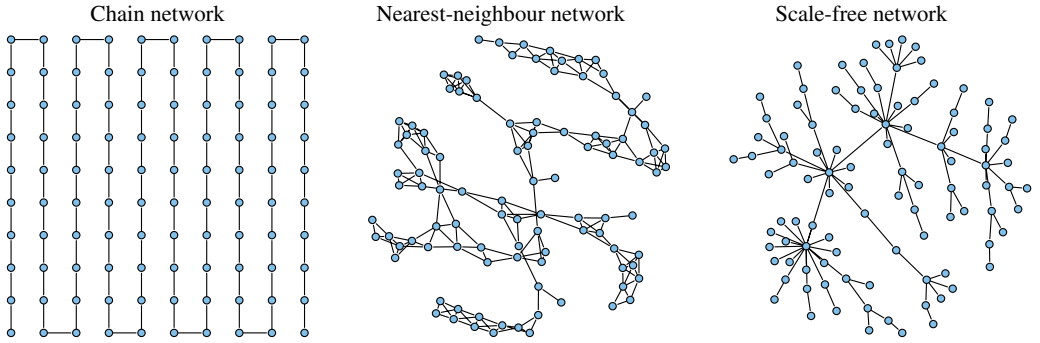

Fig. 1. The common links present in all categories in the three simulated networks.

the K categories. Further, we add heterogeneity to the common structure by creating additional individual links as follows: for each $\Omega^{(k)}$ $(k = 1, \ldots, K)$, we randomly pick a pair of symmetric zero elements and replace them with a value uniformly generated from the $[-1, -0{\cdot}5] \cup [0{\cdot}5, 1]$ interval. This procedure is repeated ρM times, where M is the number of off-diagonal nonzero elements in the lower triangular part of $\Omega^{(k)}$ and ρ is the ratio of the number of individual links to the number of common links. In the simulations, we considered values of $\rho = 0$, 1/4, 1 and 4, thus gradually increasing the proportion of individual links.

In the second example, the nearest-neighbour networks are generated by modifying the data generating mechanism described in Li & Gui (2006). Specifically, we generate p points randomly on a unit square, calculate all $p(p-1)/2$ pairwise distances, and find m nearest neighbours of each point in terms of this distance. The nearest neighbour network is obtained by linking any two points that are m-nearest neighbours of each other. The integer m controls the degree of sparsity of the network and the value $m = 5$ was chosen in our study. The middle panel of Fig. 1 illustrates a realization of the common structure of a nearest-neighbour network. Subsequently, K individual graphs were generated, by adding some individual links to the common graph with $\rho = 0$, 1/4, 1, 4 by the same method as described in Example 1, with values for the individual links $\omega^{(k)}_{j,j'}$ generated from a uniform distribution on $[-1, -0{\cdot}5] \cup [0{\cdot}5, 1]$.

In the last example, we generate the common structure of a scale-free network using the Barabasi–Albert algorithm (Barabasi & Albert, 1999); a realization is depicted in the right panel of Fig. 1. The individual links in the kth network $(k = 1, \ldots, K)$, are randomly added as before, with $\rho = 0$, 1/4, 1, 4 and the associated elements in $\Omega^{(k)}$ are generated uniformly on $[-1, -0{\cdot}5] \cup [0{\cdot}5, 1]$.

We compare the joint estimation method to the method that estimates each category separately via (1). A number of metrics are used to assess performance, including receiver operating characteristic curves, average entropy loss, average Frobenius loss, average false positive and average false negative rates, and the average rate of misidentified common zeros among the categories. For the receiver operating characteristic 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 and average Frobenius loss are defined as

$$\begin{aligned} \text{EL} &= \frac{1}{K}\sum_{k=1}^{K} \text{tr}\{(\Omega^{(k)})^{-1}\hat{\Omega}^{(k)}\} - \log[\det\{(\Omega^{(k)})^{-1}\hat{\Omega}^{(k)}\}] - p, \\ \text{FL} &= \frac{1}{K}\sum_{k=1}^{K} \|\Omega^{(k)} - \hat{\Omega}^{(k)}\|_F^2 / \|\Omega^{(k)}\|_F^2. \end{aligned} \tag{6}$$

The average false positive rate gives the proportion of false discoveries, that is, true zeros estimated as nonzero; 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 $\Omega^{(1)}, \ldots, \Omega^{(K)}$ estimated as nonzero. The respective formal definitions are

$$\begin{aligned}
\text{FP} &= \frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{1\leqslant j<j'\leqslant p} \mathrm{I}(\omega_{j,j'}^{(k)}=0, \hat{\omega}_{j,j'}^{(k)}\neq 0)}{\sum_{1\leqslant j<j'\leqslant p} \mathrm{I}(\omega_{j,j'}^{(k)}=0)},\\
\text{FN} &= \frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{1\leqslant j<j'\leqslant p} \mathrm{I}(\omega_{j,j'}^{(k)}\neq 0, \hat{\omega}_{j,j'}^{(k)}=0)}{\sum_{1\leqslant j<j'\leqslant p} \mathrm{I}(\omega_{j,j'}^{(k)}\neq 0)},\\
\text{CZ} &= \frac{\sum_{1\leqslant j<j'\leqslant p} \mathrm{I}(\sum_{k=1}^{K}|\omega_{j,j'}^{(k)}|=0, \sum_{k=1}^{K}|\hat{\omega}_{j,j'}^{(k)}|\neq 0)}{\sum_{1\leqslant j<j'\leqslant p} \mathrm{I}(\sum_{k=1}^{K}|\omega_{j,j'}^{(k)}|=0)}.
\end{aligned} \tag{7}$$

4·2. *Simulation results*

Figure 2 shows the estimated ROC, receiver operating characteristic, 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. As ρ increases, the structures become more and more different, and the joint and separate methods move closer together, with the separate method eventually slightly outperforming the joint method at $\rho = 4$, although the results are still fairly similar. This is precisely as it should be, since the joint estimation method has the biggest advantage with the most overlap in structure. In order to assess the variability of the two methods, we drew the boxplots of the sensitivity of the two models with the false positive rate controlled at 5%; the results indicate that as long as there is a substantial common structure, the joint method is superior to the separate method and the difference is statistically significant.

Table 1 summarizes the results based on 50 replications with the tuning parameter selected by BIC(λ) and crossvalidation as described in §2·4. In general, the joint estimation method produces lower entropy and Frobenius norm losses for both model selection criteria, with the difference most pronounced at low values of ρ. For the joint method, the two model selection criteria exhibit closer agreement in false positive and false negative rates and the proportion of misidentified common zeros. For the separate method, however, crossvalidation tends to select more false positive links, which result in more misidentified common zeros.

5. UNIVERSITY WEBPAGES EXAMPLE

The dataset was collected in 1997 and includes webpages from computer science departments at Cornell, the University of Texas, University of Washington and University of Wisconsin. The original data have been pre-processed using standard text processing procedures, such as removing stopwords and stemming the words. The pre-processed dataset can be downloaded from http://web.ist.utl.pt/∼acardoso/datasets/. The webpages were manually classified into seven categories, from which we selected the four largest for our analysis: student, faculty, course and project, with 544, 374, 310 and 168 webpages, respectively. The log-entropy weighting method (Dumais, 1991) was used to calculate the term-document matrix $X = (x_{i,j})_{n\times p}$, with n and p denoting the number of webpages and distinct terms, respectively. Let $f_{i,j}$ $(i = 1, \ldots, n;\ \ j = 1, \ldots, p)$ be the number of times the jth term appears in

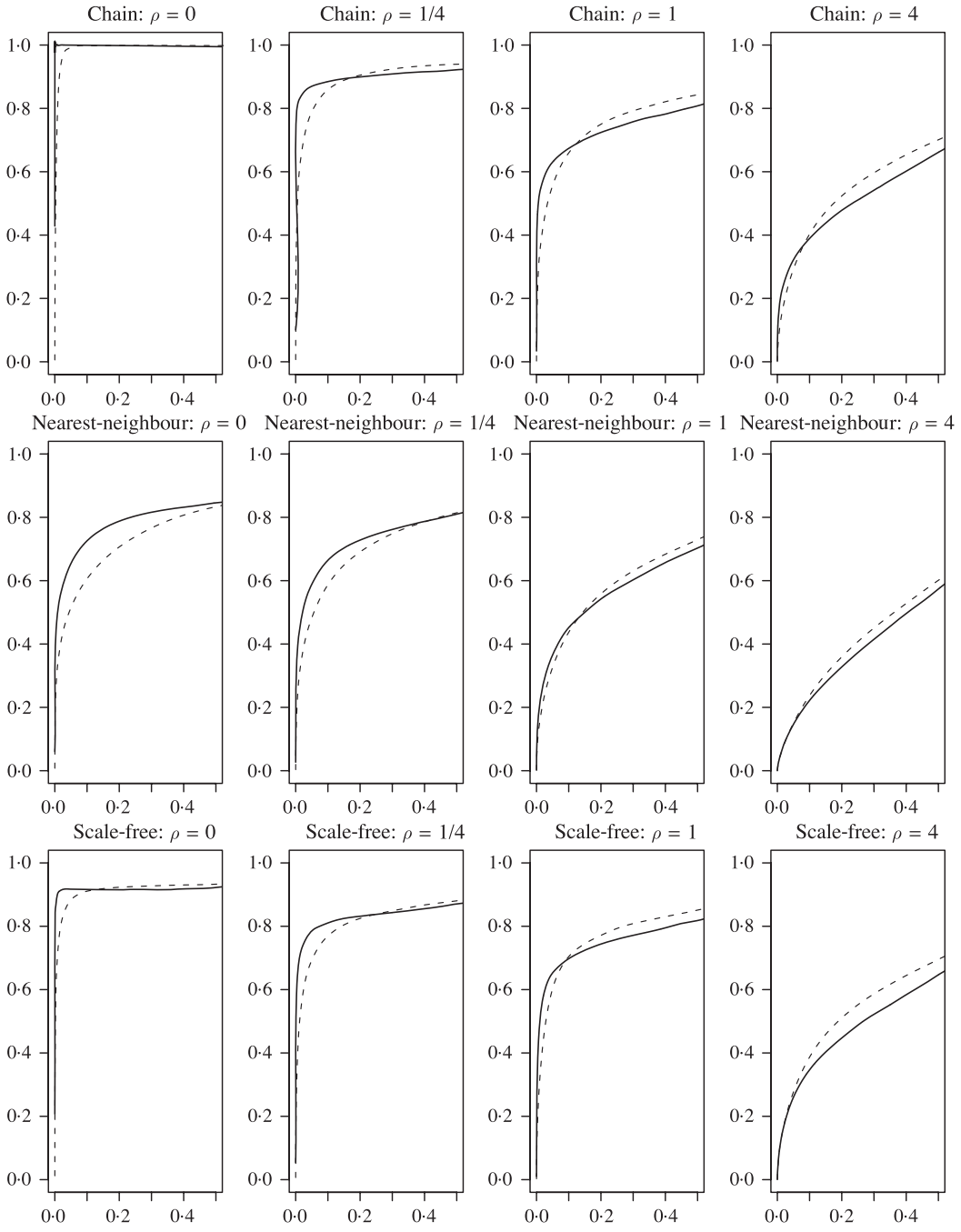

Fig. 2. Receiver operating characteristic curves. The horizontal and vertical axes in each panel are false positive rate and sensitivity, respectively. The solid line corresponds to the joint estimation method, and the dashed line corresponds to the separate estimation method. ρ is the ratio of the number of individual links to the number of common links.

the ith webpage and let $p_{i,j} = f_{i,j} / \sum_{i=1}^{n} f_{i,j}$. Then, the log-entropy weight of the jth 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 X is defined as $x_{i,j} = e_j \log(1 + f_{i,j})$ $(i = 1, \ldots, n;\ \ j = 1, \ldots, p)$ and it is normalized along each column. We applied the proposed joint estimation method to $n = 1396$ documents in the four largest

Table 1. *Results from the three simulated examples. In each cell, the numbers before and after the slash correspond to the results from selected by* BIC *and crossvalidation, respectively*

Example	ρ	Method	EL	FL	FN (%)	FP (%)	CZ (%)
Chain	0	S	20·7 / 21·9	0·5 / 0·5	0·8 / 0·1	5·7 / 21·8	14·5 / 51·0
		J	12·8 / 6·6	0·3 / 0·3	0·0 / 0·0	4·3 / 0·5	7·0 / 1·2
	1/4	S	21·3 / 16·6	0·5 / 0·5	41·3 / 9·0	1·3 / 18·7	3·8 / 46·0
		J	9·5 / 8·7	0·3 / 0·3	15·6 / 17·6	1·7 / 0·7	3·2 / 1·4
	1	S	23·0 / 17·1	0·5 / 0·5	73·7 / 24·4	0·7 / 18·8	1·9 / 46·4
		J	12·5 / 12·4	0·4 / 0·4	44·2 / 45·8	1·6 / 1·1	3·0 / 2·0
	4	S	29·8 / 20·2	0·6 / 0·5	97·3 / 47·5	0·1 / 19·5	0·3 / 47·8
		J	20·0 / 20·7	0·5 / 0·5	75·5 / 76·2	1·9 / 1·8	3·2 / 3·0
Nearest-neighbour	0	S	11·9 / 15·9	0·4 / 0·5	40·1 / 33·5	2·2 / 16·1	6·1 / 40·5
		J	6·1 / 11·3	0·3 / 0·4	18·5 / 52·7	1·6 / 0·6	3·2 / 1·3
	1/4	S	13·9 / 17·1	0·4 / 0·5	44·0 / 32·5	2·4 / 17·6	6·9 / 43·9
		J	8·1 / 14·5	0·3 / 0·4	27·4 / 57·5	1·7 / 1·0	2·9 / 1·7
	1	S	18·5 / 18·0	0·5 / 0·5	48·5 / 45·3	4·0 / 17·8	11·2 / 44·3
		J	13·0 / 19·0	0·4 / 0·5	40·0 / 77·3	2·8 / 1·2	3·8 / 2·0
	4	S	24·8 / 20·1	0·5 / 0·5	98·7 / 65·5	0·1 / 18·1	0·3 / 44·9
		J	19·3 / 23·8	0·7 / 0·5	80·8 / 95·0	3·2 / 1·0	4·8 / 1·6
Scale-free	0	S	16·9 / 15·5	0·5 / 0·5	20·7 / 6·4	1·9 / 17·1	5·3 / 42·1
		J	8·1 / 7·0	0·3 / 0·3	9·4 / 11·2	1·5 / 0·5	2·8 / 1·0
	1/4	S	17·1 / 14·5	0·5 / 0·4	49·6 / 17·5	1·2 / 16·6	3·7 / 41·8
		J	9·4 / 9·1	0·3 / 0·3	29·3 / 32·2	1·3 / 0·8	2·4 / 1·4
	1	S	22·3 / 18·1	0·5 / 0·5	51·8 / 22·5	2·8 / 19·3	8·2 / 47·4
		J	15·2 / 15·3	0·4 / 0·4	42·5 / 43·1	2·2 / 2·0	3·2 / 2·9
	4	S	27·9 / 20·0	0·6 / 0·5	99·6 / 49·6	0·0 / 19·1	0·0 / 47·0
		J	23·0 / 23·8	0·5 / 0·5	82·5 / 84·1	2·1 / 1·8	3·2 / 2·7

S, the separate method; J, the joint method; EL, FL, FN, FP and CZ are defined in equations (6) and (7); ρ the ratio of the number of individual links to the number of common links.

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 Fig. 3(a). 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 Fig. 3(b)–(d); each graph clearly has its own semantic meaning, which we loosely label as webpage generic, research area/lab and parallel programming.

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 Fig. 4. 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, faculti-student, assist-professor and associ-professor. Similarly, we illustrate the differences between the course and project categories in Fig. 5. Some teaching-related terms are linked only in the course category, such as office-hour, office-instructor and

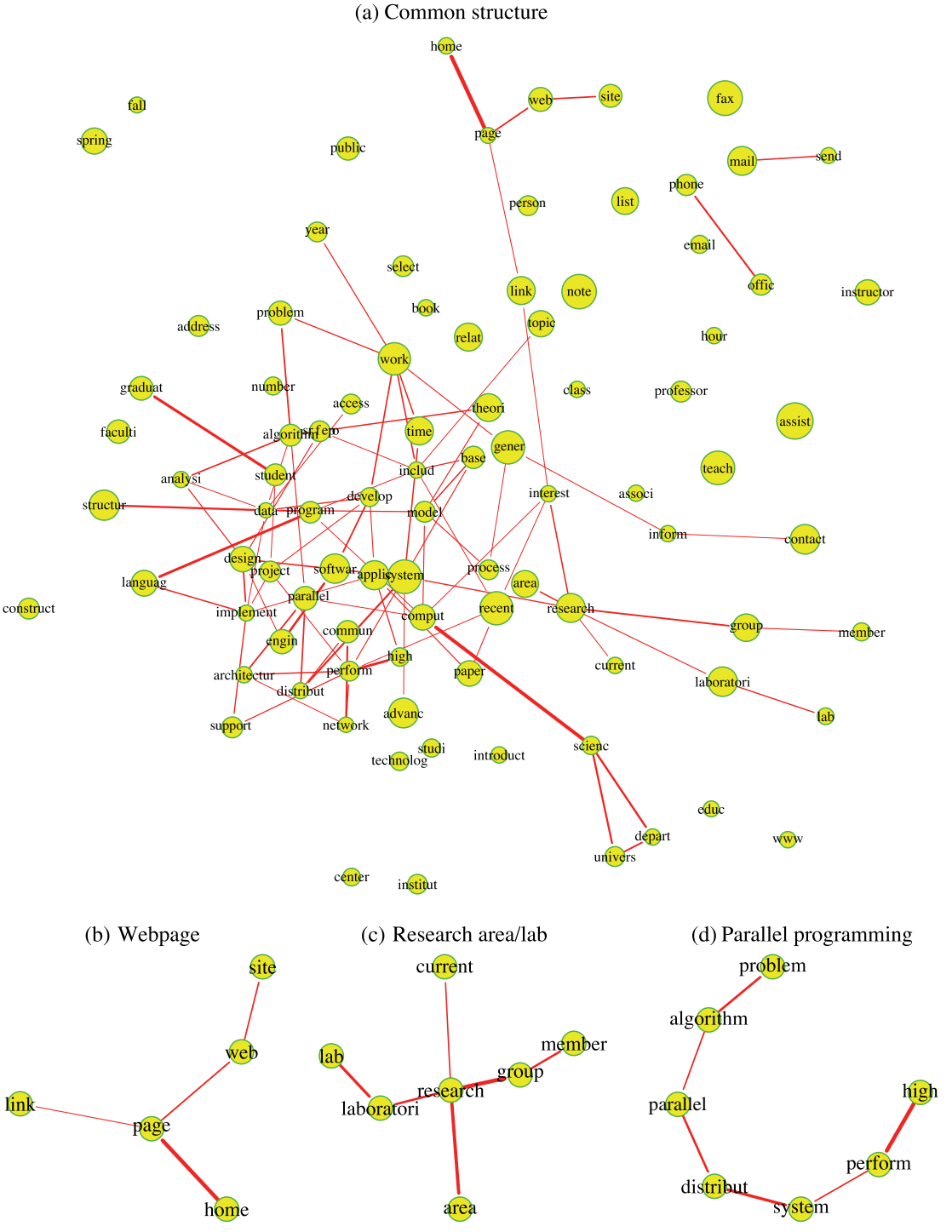

Fig. 3. Common structure in the webpages data. Panel (a) shows the estimated common structure for the four categories. The nodes represent 100 terms with the highest log-entropy weights. The area of the circle representing a node is proportional to its log-entropy weight. The width of an edge is proportional to the magnitude of the associated partial correlation. Panels (b)–(d) show subgraphs extracted from the graph in panel (a).

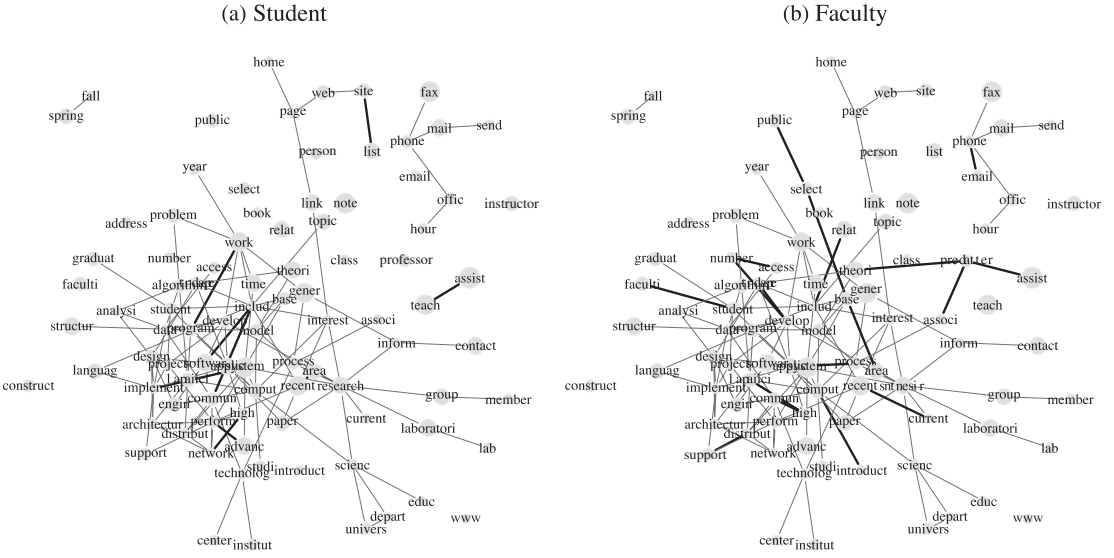

Fig. 4. 'Student' and 'Faculty' graphs. The thin light lines are the links appearing in both categories, and the thick dark lines are the links only appearing in one category.

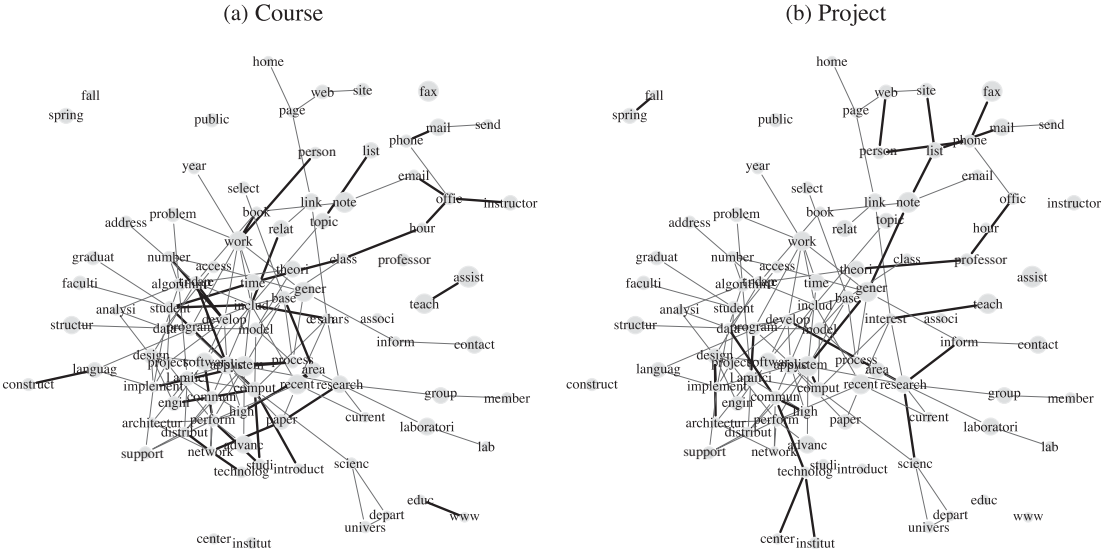

Fig. 5. 'Course' and 'Project' graphs. The thin light lines are the links appearing in both categories, and the thick dark lines are the links only appearing in one category.

teach-assist, while pairs in the project category are connected to research, such as 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. When each category is estimated separately, individual links dominate, and the results are not as easy to interpret. The graphical models obtained by separate estimation are not shown for lack of space.

ACKNOWLEDGEMENT

The authors thank the editor, the associate editor, two reviewers and Sijian Wang from the University of Wisconsin for helpful suggestions. E.L. and J.Z. are partially supported by National Science Foundation grants, and G.M. is partially supported by grants from the National Institutes of Health and the Michigan Economic Development Corporation.

APPENDIX

In the beginning, we state some results used in the proof of Theorem 1 that were established in Rothman et al. (2008, Theorem 1). We use the following notation: for a matrix $M = (m_{j,j'})_{p\times p}$, $|M|_1 = \sum_{j,j'} |m_{j,j'}|$, M^+ is a diagonal matrix with the same diagonal as M, $M^- = M - M^+$ and M_S is M with all elements outside an index set S replaced by zeros. We also write $\widetilde{M}$ for the vectorized $p^2 \times 1$ form of M, and $\otimes$ for the Kronecker product of two matrices. In addition, we denote $\Sigma_0^{(k)} = (\Omega_0^{(k)})^{-1}$ as the true covariance matrix of the kth category ($k = 1, \ldots, K$).

LEMMA A1. *Let* $l(\Omega^{(k)}) = \mathrm{tr}(\hat{\Sigma}^{(k)}\Omega^{(k)}) - \log\{\det(\Omega^{(k)})\}$. *Then for any* $k = 1, \ldots, K$, *the following decomposition holds:*

$$l(\Omega_0^{(k)} + \Delta^{(k)}) - l(\Omega_0^{(k)}) = \mathrm{tr}\{(\hat{\Sigma}^{(k)} - \Sigma_0^{(k)})\Delta^{(k)}\} + (\tilde{\Delta}^{(k)})^{\mathrm{T}} \left\{ \int_0^1 (1-v)(\Omega_0^{(k)} + v\Delta^{(k)})^{-1} \otimes (\Omega_0^{(k)} + v\Delta^{(k)})^{-1} dv \right\} \tilde{\Delta}^{(k)}. \tag{A1}$$

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

$$|\mathrm{tr}\{(\hat{\Sigma}^{(k)} - \Sigma_0^{(k)})\Delta^{(k)}\}| \leqslant C_1 \left(\frac{\log p}{n}\right)^{1/2} |\Delta^{(k)-}|_1 + C_2 \left(\frac{p\log p}{n}\right)^{1/2} \|\Delta^{(k)+}\|_F, \tag{A2}$$

$$(\tilde{\Delta}^{(k)})^{\mathrm{T}} \left\{ \int_0^1 (1-v)(\Omega_0^{(k)} + v\Delta^{(k)})^{-1} \otimes (\Omega_0^{(k)} + v\Delta^{(k)})^{-1} dv \right\} \tilde{\Delta}^{(k)} \geqslant \frac{1}{4\tau_2^2} \|\Delta^{(k)}\|_F^2. \tag{A3}$$

Proof of Theorem 1. In a slight abuse of notation, we will write $\Omega = (\Omega^{(k)})_{k=1}^K$, $\Omega_0 = (\Omega_0^{(k)})_{k=1}^K$, and $\Delta = (\Delta^{(k)})_{k=1}^K$, where $\Delta^{(k)} = (\delta_{j,j'}^{(k)})_{p\times p}$ is defined as $\Delta^{(k)} = \Omega^{(k)} - \Omega_0^{(k)}$ ($k = 1, \ldots, K$). Let $Q(\Omega)$ be the objective function of (4), and let $G(\Delta) = Q(\Omega_0 + \Delta) - Q(\Omega_0)$. 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 $G(0) = 0$. Specifically, we define $\mathcal{A} = \{\Delta : (\sum_{k=1}^K \|\Delta^{(k)}\|_F) \leqslant Mr_n\}$, with boundary $\partial\mathcal{A} = \{\Delta : (\sum_{k=1}^K \|\Delta^{(k)}\|_F) = Mr_n\}$, where M is a positive constant and $r_n = \{(p+q)(\log p)/n\}^{1/2}$.

By the decomposition (A1) in Lemma A1, we can write $G(\Delta) = I_1 + I_2 + I_3 + I_4$, where

$$I_1 = \sum_{k=1}^K \mathrm{tr}\{(\hat{\Sigma}^{(k)} - \Sigma_0^{(k)})\Delta^{(k)}\},$$

$$I_2 = \sum_{k=1}^K (\tilde{\Delta}^{(k)})^{\mathrm{T}} \left\{ \int_0^1 (1-v)(\Omega_0^{(k)} + v\Delta^{(k)})^{-1} \otimes (\Omega_0^{(k)} + v\Delta^{(k)})^{-1} dv \right\} \tilde{\Delta}^{(k)},$$

$$I_3 = \lambda \sum_{(j,j')\in T^c} \left(\sum_{k=1}^K |\delta_{j,j'}^{(k)}| \right)^{1/2},$$

$$I_4 = \lambda \sum_{j\neq j':(j,j')\in T} \left\{ \left(\sum_{k=1}^K |\omega_{j,j'}^{(k)}| \right)^{1/2} - \left(\sum_{k=1}^K |\omega_{0,j,j'}^{(k)}| \right)^{1/2} \right\}.$$

We first consider I_1. By applying inequality (A2) in Lemma A1, we have $|I_1| \leqslant I_{1,1} + I_{1,2}$, where $I_{1,1} = C_1\{(\log p)/n\}^{1/2} \sum_{k=1}^K |\Delta_T^{(k)-}|_1 + C_2\{(p \log p)/n\}^{1/2} \sum_{k=1}^K \|\Delta^{(k)+}\|_F$ and $I_{1,2} = C_1\{(\log p)/n\}^{1/2} \sum_{k=1}^K |\Delta_{T^c}^{(k)-}|_1$. By applying the bound $|\Delta_T^{(k)-}|_1 \leqslant q_k^{1/2} \|\Delta_T^{(k)-}\|_F$, we have

$$I_{1,1} \leqslant C_1 \left(\frac{q \log p}{n}\right)^{1/2} \sum_{k=1}^K \|\Delta_T^{(k)-}\|_F + C_2 \left(\frac{p \log p}{n}\right)^{1/2} \sum_{k=1}^K \|\Delta^{(k)+}\|_F$$

$$\leqslant (C_1 + C_2) \left\{\frac{(p+q)\log p}{n}\right\}^{1/2} \sum_{k=1}^K \|\Delta^{(k)}\|_F \leqslant M(C_1 + C_2) \frac{(p+q)\log p}{n}$$

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

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

$$I_3 - I_{1,2} \geqslant \lambda \sum_{k=1}^K |\Delta_{T^c}^{(k)-}|_1 - C_1 \left(\frac{\log p}{n}\right)^{1/2} \sum_{k=1}^K |\Delta_{T^c}^{(k)-}|_1$$

$$\geqslant (\Lambda_1 - C_1) \left(\frac{\log p}{n}\right)^{1/2} \sum_{k=1}^K |\Delta_{T^c}^{(k)-}|_1.$$

The last inequality uses the condition $\lambda \geqslant \Lambda_1\{(\log p)/n\}^{1/2}$. Therefore, $I_3 - I_{1,2} \geqslant 0$ when Λ_1 is large enough. Next we consider I_2. By applying inequality (A3) in Lemma A1, we have $I_2 \geqslant (1/4\tau_2^2) \sum_{k=1}^K \|\Delta^{(k)}\|_F^2 \geqslant \{M^2/(8\tau_2^2)\}\{(p+q)(\log p)/n\}$. Finally consider the remaining term I_4. Using Condition 2, we have

$$|I_4| \leqslant \lambda \sum_{j \neq j':(j,j') \in T} \frac{\sum_{k=1}^K \|\omega_{j,j'}^{(k)}| - |\omega_{0,j,j'}^{(k)}\|}{\left(\sum_{k=1}^K |\omega_{j,j'}^{(k)}|\right)^{1/2} + \left(\sum_{k=1}^K |\omega_{0,j,j'}^{(k)}|\right)^{1/2}}$$

$$\leqslant \frac{\lambda}{\tau_3^{1/2}} \sum_{k=1}^K \sum_{j \neq j':(j,j') \in T} |\omega_{j,j'}^{(k)} - \omega_{0,j,j'}^{(k)}| \leqslant \frac{\lambda}{\tau_3^{1/2}} q^{1/2} \sum_{k=1}^K \|\Delta^{(k)}\|_F \leqslant \frac{M\Lambda_2}{\tau_3^{1/2}} \frac{(p+q)(\log p)}{n}.$$

The last inequality uses the condition $\lambda \leqslant \Lambda_2\{(1 + p/q)(\log p)/n\}^{1/2}$. Putting everything together and using $I_2 > 0$ and $I_3 - I_{1,2} > 0$, we have

$$G(\Delta) \geqslant I_2 - I_{1,1} - |I_4| \geqslant M^2 \frac{(p+q)\log p}{n} \left(\frac{1}{8\tau_2^2} - \frac{C_1 + C_2 + \Lambda_2/\tau_3^{1/2}}{M}\right).$$

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

Proof of Theorem 2. It suffices to show that for all $(j, j') \in T_k^c$ $(k = 1, \ldots, K)$, the derivative $\partial Q/\partial\omega_{j,j'}^{(k)}$ at $\hat{\omega}_{j,j'}^{(k)}$ has the same sign as $\hat{\omega}_{j,j'}^{(k)}$ with probability tending to 1. To see that, suppose that for some $(j, j') \in T_k^c$, the estimate $\hat{\omega}_{j,j'}^{(k)} \neq 0$. Without loss of generality, suppose $\hat{\omega}_{j,j'}^{(k)} > 0$. Then there exists $\xi > 0$ such that $\hat{\omega}_{j,j'}^{(k)} - \xi > 0$. Since $\hat{\Omega}$ is a local minimizer of $Q(\Omega)$, we have $\partial Q/\partial\omega_{j,j'}^{(k)} < 0$ at $\hat{\omega}_{j,j'}^{(k)} - \xi$ for ξ small, contradicting the claim that $\partial Q/\partial\omega_{j,j'}^{(k)}$ at $\hat{\omega}_{j,j'}^{(k)}$ has the same sign as $\hat{\omega}_{j,j'}^{(k)}$.

The derivative of the objective function can be written as

$$\frac{\partial Q}{\partial\omega_{j,j'}^{(k)}} = 2\{\alpha_{j,j'}^{(k)} + \beta_{j,j'} \operatorname{sgn}(\omega_{j,j'}^{(k)})\}, \tag{A4}$$

where $\alpha_{j,j'}^{(k)} = \hat{\sigma}_{j,j'}^{(k)} - \sigma_{j,j'}^{(k)}$ and $\beta_{j,j'} = \lambda/(\sum_{k=1}^K |\omega_{j,j'}^{(k)}|)^{1/2}$. Arguing as in Lam & Fan (2009, Theorem 2), one can show that $\max_{k=1,\ldots,K} \max_{j,j'} |\alpha_{j,j'}^{(k)}| = O_p[\{(\log p)/n\}^{1/2} + \eta_n^{1/2}]$. On the other hand, by

Theorem 1, we have $\sum_{k=1}^{K} |\omega_{j,j'}^{(k)} - \omega_{0,j,j'}^{(k)}| \leqslant \sum_{k=1}^{K} \|\Omega^{(k)} - \Omega_0^{(k)}\|_F = O_p(\eta_n) = o(1)$. Then for any $\epsilon > 0$ and large enough n we have $\sum_{k=1}^{K} |\omega_{j,j'}^{(k)}| \leqslant \sum_{k=1}^{K} |\omega_{0,j,j'}^{(k)}| + \epsilon$. Then we have $|\beta_{j,j'}| \geqslant \lambda/(1 + \sum_{k=1}^{K} |\omega_{0,j,j'}^{(k)}|)^{-1/2}$. By assumption, $\{(\log p)/n\}^{1/2} + \eta_n^{1/2} = O(\lambda)$, and thus the term $\beta_{j,j'}$ dominates $\alpha_{j,j'}^{(k)}$ in (A4) for any $(j, j') \in T_k^c$ $(k = 1, \ldots, K)$. Therefore, $\operatorname{sgn}\{(\partial Q/\partial \omega_{j,j'}^{(k)})|_{\omega_{j,j'}^{(k)} = \hat{\omega}_{j,j'}^{(k)}}\} = \operatorname{sgn}(\hat{\omega}_{j,j'}^{(k)})$. □

References

Banerjee, O., El Ghaoui, L. & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation. *J. Mach. Learn. Res.* **9**, 485–516.

Barabasi, A.-L. & Albert, R. (1999). Emergence of scaling in random networks. *Science* **286**, 509–12.

Bickel, P. J. & Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199–227.

d'Aspremont, A., Banerjee, O. & El Ghaoui, L. (2008). First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. Appl.* **30**, 56–66.

Dempster, A. P. (1972). Covariance selection. *Biometrics* **28**, 157–75.

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

Dumais, S. T. (1991). Improving the retrieval of information from external source. *Behav. Res. Meth. Instr. Comp.* **23**, 229–36.

Edwards, D. (2000). *Introduction to Graphical Modelling*. New York: Springer.

Fan, J., Feng, Y. & Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *Ann. Appl. Statist.* **3**, 521–41.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

Friedman, J., Hastie, T. & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–41.

Hoefling, H. & Tibshirani, R. (2009). Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *J. Mach. Learn. Res.* **10**, 883–906.

Lam, C. & Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrices estimation. *Ann. Statist.* **37**, 4254–78.

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

Li, H. & Gui, J. (2006). Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics* **7**, 302–17.

Meinshausen, N. & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 1436–62.

Michailidis, G. & de Leeuw, J. (2001). Multilevel homogeneity analysis with differential weighting. *Comp. Statist. Data Anal.* **32**, 411–42.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Oxford: Cambridge University Press.

Peng, J., Wang, P., Zhou, N. & Zhu, J. (2009). Partial correlation estimation by joint sparse regression model. *J. Am. Statist. Assoc.* **104**, 735–46.

Ravikumar, P., Wainwright, M. J. & Lafferty, J. D. (2009). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38**, 1287–319.

Rothman, A. J., Bickel, P. J., Levina, E. & Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Statist.* **2**, 494–515.

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

Zou, H., Hastie, T. & Tibshirani, R. (2007). On the degrees of freedom of the LASSO. *Ann. Statist.* **35**, 2173–92.

Zou, H. & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36**, 1108–26.

[*Received December* 2009. *Revised June* 2010]