Biometrika (2016), **103**, 3, *pp.* 493–511
Printed in Great Britain
doi: 10.1093/biomet/asw035

Joint estimation of multiple dependent Gaussian graphical models with applications to mouse genomics

BY YUYING XIE

Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, Michigan 48824, U.S.A.
xyy@stt.msu.edu

YUFENG LIU

Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A.
yfliu@email.unc.edu

AND WILLIAM VALDAR

Department of Genetics, University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A.
william.valdar@unc.edu

SUMMARY

Gaussian graphical models are widely used to represent conditional dependencies among random variables. In this paper, we propose a novel estimator for data arising from a group of Gaussian graphical models that are themselves dependent. A motivating example is that of modelling gene expression collected on multiple tissues from the same individual: here the multivariate outcome is affected by dependencies acting not only at the level of the specific tissues, but also at the level of the whole body; existing methods that assume independence among graphs are not applicable in this case. To estimate multiple dependent graphs, we decompose the problem into two graphical layers: the systemic layer, which affects all outcomes and thereby induces cross-graph dependence, and the category-specific layer, which represents graph-specific variation. We propose a graphical EM technique that estimates both layers jointly, establish estimation consistency and selection sparsistency of the proposed estimator, and confirm by simulation that the EM method is superior to a simpler one-step method. We apply our technique to mouse genomics data and obtain biologically plausible results.

Some key words: EM algorithm; Gaussian graphical model; Mouse genomics; Shrinkage; Sparsity; Variable selection.

1. INTRODUCTION

Gaussian graphical models are widely used to represent conditional dependencies among sets of normally distributed outcome variables. For example, observed, and potentially dense, correlations between measurements of expression for multiple genes, stock market prices of different asset classes, or blood flow for multiple voxels in functional magnetic resonance imaging, i.e., fMRI-measured brain activity, can often be more parsimoniously explained by an underlying

graph whose structure may be relatively sparse. As methods for estimating these underlying graphs have matured, a number of elaborations to basic Gaussian graphical models have been proposed, including those that seek either to model the sampling distribution of the data more closely, or to model prior expectations of the analyst about structural similarities among graphs representing related datasets (Guo et al., 2011; Danaher et al., 2014; Lee & Liu, 2015). In this paper, we propose an elaboration that seeks to model an additional feature of the sampling distribution increasingly encountered in biomedical data, whereby correlations among the outcome variables are considered to be the by-product of underlying conditional dependencies acting at different levels. For illustration, consider gene expression data obtained from multiple tissues, such as liver, kidney, and brain, collected on each individual. In this setting, observed correlations between expressed genes may be caused by dependence structures not only within a specific tissue but also across tissues at the level of the whole body. We describe these distinct graphical strata respectively as the category-specific and systemic layers, and model their respective outcomes as latent variables.

The conditional dependence relationships among p outcome variables, $Y = (Y_1, \ldots, Y_p)$, can be represented by a graph $\mathcal{G} = (\Gamma, E)$, where each variable is a node in the set Γ and conditional dependencies are represented by the edges in the set E. If the joint distribution of the outcome variables is multivariate Gaussian, $Y \sim N(0, \Sigma)$, then conditional dependencies are reflected in the nonzero entries of the precision matrix $\Omega = \Sigma^{-1}$. Specifically, two variables Y_i and Y_j are conditionally independent given the other variables if and only if the (i, j)th entry of Ω is zero. Inferring the dependence structure of such a Gaussian graphical model is thus the same as estimating which elements of its precision matrix are nonzero.

When the underlying graph is sparse, as is often assumed, the maximum likelihood estimator is dominated in terms of the false positive rate by shrinkage estimators. The maximum likelihood estimate of Ω typically implies a graph that is fully connected, which is unhelpful for estimating graph topology. To impose sparsity, and thereby provide a more informative inference about network structure, a number of methods have been introduced that estimate Ω under ℓ_1 regularization. Meinshausen & Bühlmann (2006) proposed to iteratively determine the edges of each node in $\mathcal{G}$ by fitting an ℓ_1-penalized regression model to the corresponding variable Y_j using the remaining variables Y_{-j} as predictors, an approach which can be viewed as optimizing a pseudolikelihood (Ambroise et al., 2009; Peng et al., 2009). More recently, numerous papers have proposed estimation using sparse penalized maximum likelihood (Yuan & Lin, 2007; Banerjee et al., 2008; d'Aspremont et al., 2008; Rothman et al., 2008; Ravikumar et al., 2011). Efficient implementations include the graphical lasso algorithm (Friedman et al., 2008) and the quadratic inverse covariance algorithm (Hsieh et al., 2014). The convergence rate and selection consistency of such penalized estimation schemes have also been investigated in theoretical studies (Rothman et al., 2008; Lam & Fan, 2009).

Although a single graph provides a useful representation of an underlying dependence structure, several extensions have been proposed. In the context where the precision matrix, and hence the graph, is dynamic over time, Zhou et al. (2010) proposed a weighted method to estimate the graph's temporal evolution. Another extension is to simultaneously estimate multiple graphs that may share some common structure. For example, when inferring how brain regions interact using fMRI data, each subject's brain corresponds to a different graph, but we would nonetheless expect some common interaction patterns across subjects, as well as patterns specific to an individual. In such cases, joint estimation of multiple related graphs can be more efficient than estimating the graphs separately. For joint estimation of Gaussian graphs, Varoquaux et al. (2010) and Honorio & Samaras (2010) proposed methods using group lasso and multi-task lasso, respectively. Both assume that all the precision matrices have the same pattern of zeros. To provide

greater flexibility, Guo et al. (2011) proposed a joint penalized method using a hierarchical penalty, and derived the convergence rate and sparsistency properties of the resulting estimators. In the same setting, Danaher et al. (2014) extended the graphical lasso (Friedman et al., 2008) to estimate multiple graphs from independent datasets using penalties based on the generalized fused lasso or, alternatively, the sparse group lasso.

The above methods for estimating multiple Gaussian graphs focus on the settings in which data collected from different categories are stochastically independent. In some applications, however, data from different categories are more naturally considered as dependent. In each of two studies considered here, gene expression data have been collected on multiple tissues in multiple mice. For each mouse we have expression measurements for p genes in each of K different tissues, that is, K different categories, represented by the p-dimensional vectors Y_k ($k = 1, \ldots, K$). In this setting, the gene expression profiles of different mice may have arisen from the same network structure, but they are otherwise stochastically independent; in contrast, the gene expression profiles of different tissues within the same mouse are stochastically dependent. For such data, increasingly common in biomedical research, the above methods are not applicable.

To explore the gene network structure across different tissues, and to characterize the dependence among tissues, we consider a decomposition of the observed gene expression Y_k into two latent vectors. In our model, we define

$$Y_k = Z + X_k, \tag{1}$$

where $Z, X_1, \ldots, X_K$ are mutually independent. Because $\mathrm{cov}(Y_k, Y_l) = \mathrm{var}(Z)$ for any $k \neq l$, Z represents dependence across different tissues. Letting Ω_j denote the precision matrix of X_j for tissue j, and defining $\mathrm{var}(Z) = \Omega_0^{-1}$, we aim to estimate Ω_k ($k = 0, \ldots, K$) from the observed outcome data on $\{Y_1, \ldots, Y_K\}$. To accomplish this joint estimation of multiple dependent networks, we propose a one-step method and an EM method.

In the above decomposition, Z can be viewed as representing systemic variation in gene expression, that is, variation manifesting simultaneously in all measured tissues of the same mouse, whereas X_k represents category-specific variation, that is, variation unique to tissue k. An important property of this two-layer model is that sparsity in the systemic and category-specific networks can produce networks for the outcome variable Y that are highly connected. Conversely, highly connected graphs for the outcome Y can easily arise from relatively sparse underlying dependencies acting at two levels. This phenomenon is illustrated in Fig. 1, which depicts category-specific networks Ω_1 and Ω_2 for two categories $C1$ and $C2$, which might correspond to, for example, liver and brain tissue-types, and a systemic network Ω_0, which reflects relationships affecting all tissues at once, for example, gene interactions that are responsive to hormone levels or other globally acting processes. Although all three underlying networks, Ω_0, Ω_1 and Ω_2, are sparse, the precision matrix of observed variables within each tissue, that is, the aggregate network $\Omega_{Y_k} = (\Omega_0^{-1} + \Omega_k^{-1})^{-1}$ following (1), is highly connected. Existing methods aiming to estimate a single sparse network layer are therefore ill-suited to this problem because they impose sparsity on the aggregate network rather than on the two simpler layers that generate it.

2. Methodology

2·1. *Problem formulation*

The following notation is used throughout the paper. We denote the true precision and covariance matrices by Ω^* and Σ^*. For any matrix $W = (\omega_{ij})$, we denote the determinant by $\det(W)$, the

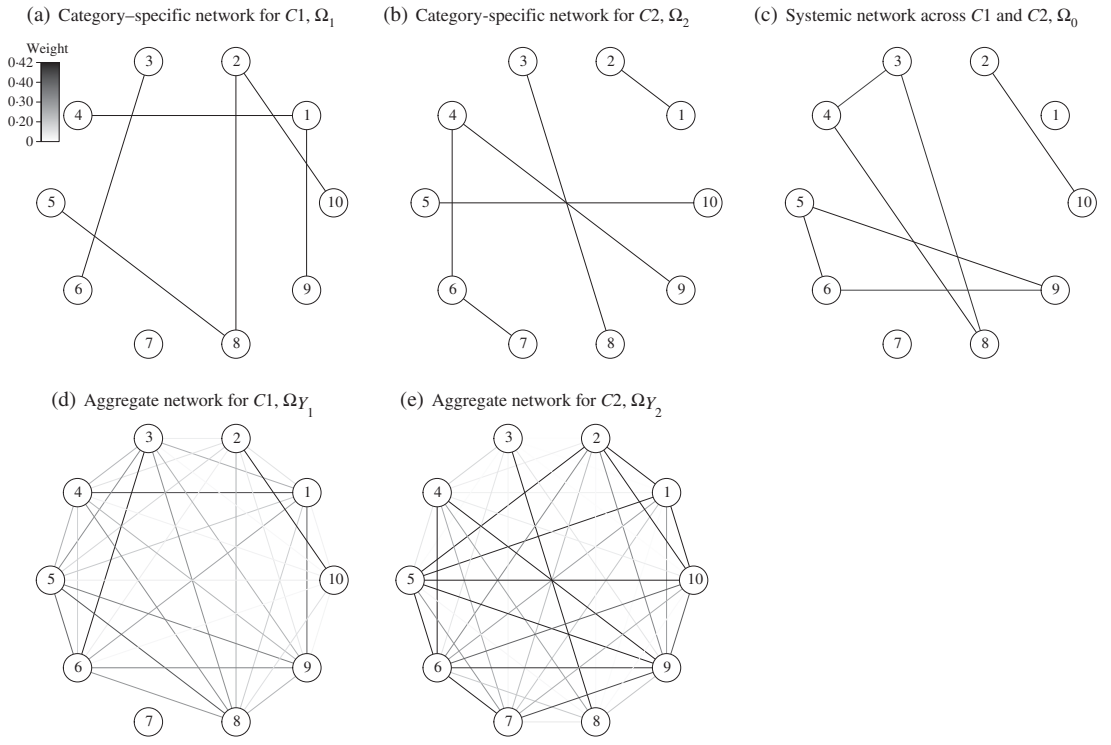

Fig. 1. Illustration of systemic and category-specific networks using a toy example with two categories, $C1$ and $C2$, and $p = 10$ variables. (a) Category-specific network for $C1$. (b) Category-specific network for $C2$. (c) Systemic network affecting variables in both $C1$ and $C2$. (d) Aggregate network $\Omega_{Y_1} = (\Omega_1^{-1} + \Omega_0^{-1})^{-1}$ for category $C1$. (e) Aggregate network $\Omega_{Y_2} = (\Omega_2^{-1} + \Omega_0^{-1})^{-1}$ for $C2$.

trace by $\mathrm{tr}(W)$ and the off-diagonal entries of W by W^-. We further denote the jth eigenvalue of W by $\phi_j(W)$ and the minimum and maximum eigenvalues of W by $\phi_{\min}(W)$ and $\phi_{\max}(W)$. The Frobenius norm, $\|W\|_{\mathrm{F}}$, is defined as $\sum_{i,j} \omega_{ij}^2$; the operator/spectral norm, $\|W\|^2$, is defined as $\phi_{\max}(WW^{\mathrm{T}})$; the infinity norm, $\|W\|_\infty$, is defined as $\max|w_{ij}|$; and the elementwise L_1-norm, $|W|_1$, is defined as $\sum_{i,j} |\omega_{ij}|$.

In the problem we address, measurements are available on the same p outcome variables in each of K distinct categories on each of n individuals. Some dependence is anticipated among outcomes both at the level of the category and at the level of the individual: dependence at the level of the category is described as category-specific, and dependence at the level of the individual is described as systemic, that is, modelled as if affecting outcomes in all categories of the same individual simultaneously. Our primary example is the measurement of gene expression giving rise to transcript abundance readings on p genes in K tissues, such as liver, kidney and brain, in n laboratory mice.

Letting $Y_{k,i}$ be the ith data vector for the kth category, we model

$$Y_{k,i} = X_{k,i} + Z_i \quad (i = 1, \ldots, n; k = 1, \ldots, K), \tag{2}$$

where Z_i is the random vector corresponding to the shared systemic random effect, and $X_{k,i}$ is the random vector corresponding to the kth category. We assume that $X_{k,i}$ and Z_i $(i = 1, \ldots, n; k = 1, \ldots, K)$ are independent and identically distributed p-dimensional random

vectors with mean zero and covariance matrices Σ_k and Σ_0 respectively. We further assume that $X_{k,i}$ and Z_i are independent of each other and each follows a multivariate Gaussian distribution.

For the ith sample in the kth category, we observe the p-dimensional realization of $Y_{k,i}$, vector $y_{k,i} = (y_{k,i,1}, \ldots, y_{k,i,p})^{\mathrm{T}}$. Without loss of generality, we assume that these observations are centred, i.e., $\sum_{i=1}^{n} y_{k,i,j} = 0$ $(j = 1, \ldots, p;\ k = 1, \ldots, K)$. Let $y_{\cdot,i}$ be the combined data vector with $y_{\cdot,i} = (y_{1,i}^{\mathrm{T}}, \ldots, y_{K,i}^{\mathrm{T}})^{\mathrm{T}}$, such that $y_{\cdot,i}$ follows a Gaussian distribution with covariance $\Sigma_Y = \{_d\Sigma_k\} + J \otimes \Sigma_0 = \{\Sigma_{Y(l,m)}\}_{1 \leqslant l,m \leqslant K}$, where $\{_d\cdot\}$ is a block-diagonal matrix, J is a square matrix with all 1s as the entries, $\otimes$ is the Kronecker product, and $\Sigma_{Y(l,m)}$ is the covariance matrix between Y_l and Y_m. We denote the $n \times Kp$ dimensional data matrix by $y = (y_{\cdot,1}, \ldots, y_{\cdot,n})^{\mathrm{T}}$, and let $\Omega_k = (\Sigma_k)^{-1} = (\omega_{k(i,j)})$ and $\Omega_Y = (\Sigma_Y)^{-1}$. Our goal is to estimate Ω_k $(k = 0, \ldots, K)$. Although X_k and Z are latent variables, we can show that Ω_k is identifiable under the model set-up in (2) with $K \geqslant 2$. More details can be found in the Supplementary Material. For simplicity, we write Ω and Σ for $\{\Omega_k\}_{k=0}^{K}$ and $\{\Sigma_k\}_{k=0}^{K}$ respectively in the following derivation.

The loglikelihood of the data can be written as

$$\mathcal{L}(\Omega; y) = -\frac{npK}{2}\log(2\pi) + \frac{n}{2}\big\{\log\det(\Omega_Y) - \mathrm{tr}(\hat{\Sigma}_Y\Omega_Y)\big\}, \tag{3}$$

where

$$\hat{\Sigma}_Y = n^{-1}\sum_{i=1}^{n} y_{\cdot,i}y_{\cdot,i}^{\mathrm{T}} = \{\hat{\Sigma}_{Y(l,m)}\}_{1 \leqslant l,m \leqslant K}$$

is the $Kp \times Kp$ sample covariance matrix. In our setting,

$$\begin{aligned}\mathcal{L}(\Omega; y) \propto \sum_{k=1}^{K}\big\{\log\det(\Omega_k) - \mathrm{tr}(\hat{\Sigma}_{Y(k,k)}\Omega_k)\big\} + \log\det(\Omega_0)\\ - \log\det(A) + \sum_{l,m=1}^{K}\mathrm{tr}\big(\Omega_l\hat{\Sigma}_{Y(l,m)}\Omega_m A^{-1}\big),\end{aligned}$$

where $A = \sum_{k=0}^{K}\Omega_k$; see the Supplementary Material for details.

A natural way to obtain a sparse estimate of Ω is to maximize the penalized loglikelihood

$$\hat{\Omega} = \underset{\Omega \succ 0}{\arg\max}\,\mathcal{P}(\Omega; y) = \underset{\Omega \succ 0}{\arg\max}\,\mathcal{L}(\Omega; y) - \lambda_1\sum_{k=1}^{K}|\Omega_k^-|_1 - \lambda_2|\Omega_0^-|_1. \tag{4}$$

Because the likelihood is complicated, direct estimation of the precision matrices in (4) is difficult. Estimation can proceed directly, however, given the values z of the latent outcome vector Z. Therefore, we first estimate Σ_0 and then the other parameters. In §§ 2·2 and 2·3, we consider estimation of these multiple dependent graphs using a one-step procedure and a method based on the EM algorithm.

2·2. *One-step method*

The idea behind our one-step method is to generate a good initial estimate for Σ and then obtain estimates for Ω by one-step optimization. Because $\mathrm{var}(Z) = \mathrm{cov}(Y_l, Y_m)$ for any $m \neq l$, it

is natural to use the covariance matrix $\Sigma_{Y(l,m)}$ between all pairs of Y_l and Y_m to estimate Σ_0 by

$$\hat{\Sigma}_0 = \frac{1}{K(K-1)} \sum_{m \neq l} \hat{\Sigma}_{Y(m,l)} = \frac{1}{K(K-1)n} \sum_{m \neq l} \sum_{i=1}^{n} \left(y_{m,i} y_{l,i}^{\mathrm{T}} \right). \tag{5}$$

Using the fact that $\mathrm{var}(X_k) = \mathrm{var}(Y_k) - \mathrm{var}(Z)$, we can then obtain an estimate for Σ_k as

$$\hat{\Sigma}_k = \hat{\Sigma}_{Y(k,k)} - \hat{\Sigma}_0 = \frac{1}{n} \sum_{i=1}^{n} \left(y_{k,i} y_{k,i}^{\mathrm{T}} \right) - \hat{\Sigma}_0. \tag{6}$$

Although $\hat{\Sigma}_k$ is symmetric, it may not be positive semidefinite, but this can be ensured using projection (Xu & Shao, 2012). For any symmetric matrix $\hat{\Sigma}_k$ $(k = 0, \ldots, K)$, the positive-semidefinite projection is

$$\hat{\Sigma}_k' = \underset{\Sigma \succeq 0}{\arg\min} \, \| \Sigma - \hat{\Sigma}_k \|_\infty. \tag{7}$$

Lastly, we estimate Ω by minimizing $K + 1$ separate functions,

$$\mathcal{W}_k(\Omega_k) = \mathrm{tr}(\hat{\Sigma}_k' \Omega_k) - \log \det(\Omega_k) + \lambda |\Omega_k^-|_1 \quad (k = 0, \ldots, K), \tag{8}$$

where $\lambda = \lambda_2$ when $k = 0$ and $\lambda = \lambda_1$ otherwise. The minimization of (8) can be solved efficiently by algorithms such as the graphical lasso (Friedman et al., 2008) or by the quadratic inverse covariance algorithm (Hsieh et al., 2014). We refer to this approach as the one-step method and later compare its performance with the EM method defined next.

2·3. *Graphical EM method*

The one-step method provides an estimate of Ω. In the spirit of the classic EM algorithm (Dempster et al., 1977), this estimate of Ω can be used to obtain a better estimate of Σ, which in turn can be used to obtain a better estimate of Ω. This procedure is iterated until the estimates of Ω converge, leading to a graphical EM algorithm, described below.

First, we rewrite the sampling model as

$$\begin{pmatrix} Z \\ Y_1 - Z \\ \vdots \\ Y_K - Z \end{pmatrix} \sim N \left\{ \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_0 & 0 & \ldots & 0 \\ 0 & \Sigma_1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \Sigma_K \end{pmatrix} \right\}$$

and the loglikelihood given $Y = y$ and $Z = z = (z_1, \ldots, z_n)^{\mathrm{T}}$ as

$$\begin{aligned} \mathcal{L}(\Omega; y, z) \propto \log \det(\Omega_0) - \mathrm{tr} \left(\Omega_0 \sum_{i=1}^{n} z_i z_i^{\mathrm{T}} / n \right) \\ + \sum_{k=1}^{K} \left[\log \det(\Omega_k) - \mathrm{tr} \left\{ \Omega_k \sum_{i=1}^{n} (y_{k,i} - z_i)(y_{k,i} - z_i)^{\mathrm{T}} / n \right\} \right]. \end{aligned} \tag{9}$$

Expression (9) cannot be calculated directly because z_i and $z_i z_i^{\mathrm{T}}$ are unobserved. However, we can replace them with their expected values conditional on Ω and y, and develop the EM algorithm with the following steps:

Step 1 (E-step). Update the expectation of the loglikelihood conditional on Ω using

$$\begin{aligned}\mathcal{Q}(\Omega;\Omega^{(t)}) &= E_{Z|\Omega^{(t)}}\{\mathcal{L}(\Omega;y,z)\}\\ &\propto \log\det(\Omega_0) - \operatorname{tr}\left\{\Omega_0\, E_{Z|\Omega^{(t)}}\left(\sum_{i=1}^{n} z_i z_i^{\mathrm{T}}/n\right)\right\} + \sum_{k=1}^{K}\log\det(\Omega_k)\\ &\quad - \sum_{k=1}^{K}\operatorname{tr}\left[\Omega_k E_{Z|\Omega^{(t)}}\left\{\sum_{i=1}^{n}(y_{k,i}-z_i)(y_{k,i}-z_i)^{\mathrm{T}}/n\right\}\right]\\ &= \sum_{k=0}^{K}\left\{\log\det(\Omega_k) - \operatorname{tr}\big(\Omega_k\dot{\Sigma}_k^{(t)}\big)\right\}.\end{aligned}$$

Step 2 (M-step). Update Ω that maximizes

$$\Omega^{(t+1)} = \arg\min_{\Omega\succ 0} -\mathcal{Q}(\Omega;\Omega^{(t)}) + \lambda_1\sum_{k=1}^{K}|\Omega_k^{-}|_1 + \lambda_2|\Omega_0^{-}|_1, \tag{10}$$

where $\Omega^{(t)}$ denotes the estimates from the tth iteration, $E_{Z|\Omega^{(t)}}(\cdot)$ denotes the conditional expectation with respect to Z given $\Omega^{(t)}$, and

$$\begin{aligned}\dot{\Sigma}_k^{(t)} &= E_{Z|\Omega^{(t)}}\left\{\sum_{i=1}^{n}(y_{k,i}-z_i)(y_{k,i}-z_i)^{\mathrm{T}}/n\right\}\\ &= \ddot{\Sigma}_{Y(k,k)} - \sum_{l=1}^{K}\left(\ddot{\Sigma}_{Y(k,l)}\Omega_l^{(t)}\right)(A^{(t)})^{-1} - (A^{(t)})^{-1}\sum_{l=1}^{K}\left(\Omega_l^{(t)}\ddot{\Sigma}_{Y(l,k)}\right)\\ &\quad + (A^{(t)})^{-1}\sum_{l,k=1}^{K}\left(\Omega_l^{(t)}\ddot{\Sigma}_{Y(l,k)}\Omega_k^{(t)}\right)(A^{(t)})^{-1} + (A^{(t)})^{-1}\quad (k=1,\dots,K),\end{aligned} \tag{11}$$

$$\dot{\Sigma}_0^{(t)} = \sum_{i=1}^{n} E_{Z|\Omega^{(t)}}\big(z_i z_i^{\mathrm{T}}/n\big) = (A^{(t)})^{-1} + (A^{(t)})^{-1}\sum_{l,k=1}^{K}\big(\Omega_l^{(t)}\ddot{\Sigma}_{Y(l,k)}\Omega_k^{(t)}\big)(A^{(t)})^{-1}, \tag{12}$$

where $\ddot{\Sigma}_Y = \hat{\Sigma}_Y$ is an estimator for Σ_Y^*, the true covariance matrix of Y. Therefore, problem (10) is decomposed into $K+1$ separate optimization problems:

$$\Omega_k^{(t+1)} = \arg\min_{\Omega_k\succ 0}\left\{\operatorname{tr}\big(\Omega_k\dot{\Sigma}_k^{(t)}\big) - \log\det(\Omega_k) + \lambda|\Omega_k^{-}|_1\right\}\quad (k=0,\dots,K), \tag{13}$$

where $\lambda=\lambda_2$ when $k=0$ and $\lambda=\lambda_1$ otherwise. We can then use the graphical lasso (Friedman et al., 2008) to solve (13).

We summarize the proposed EM method in the following algorithm.

Algorithm 1. The graphical EM algorithm.

(Initial value). Initialize $\hat{\Sigma}_0'$ and $\hat{\Sigma}_k'$ $(k = 1, \ldots, K)$ using (3), (5)–(7).
(Updating rule: the M-step). Update Ω_k $(k = 0, \ldots, K)$ by (13) using the graphical lasso.
(Updating rule: the E-step). Update $\dot{\Sigma}_k$ using (11) and (12).
Iterate the M- and E-steps until convergence.
Output $\hat{\Omega}_k$ $(k = 0, \ldots, K)$.

The next proposition demonstrates convergence of our graphical EM algorithm.

PROPOSITION 1. *With any given n, p,* $\lambda_1 > 0$, *and* $\lambda_2 > 0$, *the graphical EM algorithm solving* (4) *has the following properties:*

Property 1. The penalized loglikelihood in (4) is bounded above.

Property 2. For each iteration, the penalized loglikelihood is nondecreasing.

Property 3. For a prespecified threshold δ, after a finite number of steps, the objective function in (4) converges in the sense that $\left|\mathcal{P}(\Omega^{(t+1)}; y) - \mathcal{P}(\Omega^{(t)}; y)\right| < \delta$.

2·4. *Model selection*

We consider two options for selecting the tuning parameter $\lambda = (\lambda_1, \lambda_2)$, minimization of the extended Bayesian information criterion (Chen & Chen, 2008) and crossvalidation. The extended Bayesian information criterion is quick to compute and takes into account both goodness of fit and model complexity. Crossvalidation, by contrast, is more computationally demanding and focuses on the predictive power of the model.

In our model, we define the extended Bayesian information criterion

$$\text{BIC}_\gamma(\lambda) = -2\mathcal{L}(\{\hat{\Omega}_k\}_{k=0}^K; y) + \nu(\lambda)\log n + 2\gamma \log \tau\{\nu(\lambda)\},$$

where $\{\hat{\Omega}_k\}_{k=0}^K$ are the estimates with the tuning parameter set at λ, $\mathcal{L}(\cdot)$ is the loglikelihood function, the degrees of freedom $\nu(\lambda)$ is the sum of the number of nonzero off-diagonal elements in $\{\hat{\Omega}_k\}_{k=0}^K$, and $\tau\{\nu(\lambda)\}$ is the number of models with size $\nu(\lambda)$, which equals $a!/\{b!(a-b)!\}$ where $a = Kp(p-1)/2$ and $b = \nu(\lambda)$. This criterion is indexed by a parameter $\gamma \in [0, 1]$. The tuning parameter λ is selected as $\hat{\lambda} = \arg\min\{\text{BIC}_\gamma(\lambda) : \lambda_1, \lambda_2 \in (0, \infty)\}$.

In describing the crossvalidation procedure, we define the predictive negative loglikelihood function as $\mathcal{F}(\Sigma, \Omega) = \text{tr}(\Sigma\Omega) - \log\det(\Omega)$. To select λ using crossvalidation, we randomly split the dataset equally into J groups, and denote the sample covariance matrix from the jth group by $\hat{\Sigma}_{Y(j,\lambda)}$ and the precision matrix estimated from the remaining groups by $\hat{\Omega}_{Y(-j,\lambda)}$. Then we choose

$$\hat{\lambda} = \arg\min_{\lambda} \left\{ \sum_{j=1}^{J} \mathcal{F}(\Sigma_{Y(j,\lambda)}, \hat{\Omega}_{Y(-j,\lambda)}) : \lambda_1, \lambda_2 \in (0, \infty) \right\}.$$

The performance of these two selection methods is reported in § 4.

3. Asymptotic properties

We introduce some notation and the regularity conditions. Let $\{\Omega_k^*\}_{k=0}^K$ be the true precision matrices with $\Omega_k^* = (\omega_{k(i,j)}^*)$, $T_k = \{(i,j) : i \neq j, \omega_{k(i,j)}^* \neq 0\}$ the index set corresponding to the nonzero off-diagonal entries in Ω_k^*, $q_k = |T_k|$ the cardinality of T_k, and $q = \sum_{k=0}^K q_k$. Let $\{\Sigma_k^*\}_{k=0}^K$ be the true covariance matrices for Z and $\{X_k\}_{k=1}^K$, and $\Sigma_Y^* = \{\Sigma_{Y(l,m)}^*\}_{1 \leqslant l,m \leqslant K}$ be the true covariance matrices for Y. We assume that the following regularity conditions hold.

Condition 1. There exist constants τ_1 and τ_2 such that $0 < \tau_1 < \phi_{\min}(\Omega_k^*) \leqslant \phi_{\max}(\Omega_k^*) < \tau_2 < \infty$ $(k = 0, \ldots, K)$.

Condition 2. There exist constants a and b such that $a\{(\log p)/n\}^{1/2} \leqslant \lambda_j \leqslant b\{(1 + p/q)(\log p)/n\}^{1/2}$ $(j = 1, 2)$.

Condition 1 bounds the eigenvalues of Ω_k^* and guarantees the existence of its inverse. Condition 2 is needed to facilitate the proof of consistency. The following theorems discuss estimation consistency and selection sparsistency of our methods.

Theorem 1 (Consistency of the one-step method). *Under Conditions* 1 *and* 2, *if* $(p + q)(\log p)/n = o(1)$, *then the solution* $\{\hat{\Omega}_k^{\text{one}}\}_{k=0}^K$ *of the one-step method satisfies*

$$\sum_{k=0}^{K} \|\hat{\Omega}_k^{\text{one}} - \Omega_k^*\|_{\text{F}} = O_{\text{p}}\left[\left\{\frac{(p+q)\log p}{n}\right\}^{1/2}\right].$$

We next present a corollary of Theorem 1 which gives a good estimator of Σ_Y^*.

Corollary 1. *Under the assumptions of Theorem* 1 *and with* $\hat{\Omega}_k^{\text{one}} (k = 0, \ldots, K)$ *being the one-step solution,* $\check{\Sigma}_k = (\hat{\Omega}_k^{\text{one}})^{-1}$ *satisfies*

$$\|\check{\Sigma}_k - \Sigma_k^*\|_{\text{F}} = O_{\text{p}}\left[\left\{\frac{(p+q)\log p}{n}\right\}^{1/2}\right].$$

To study our EM estimator, we need an estimator for Σ_Y^* that satisfies the following condition.

Condition 3. We assume there exists an estimator $\tilde{\Sigma}_Y$ such that

$$\|\tilde{\Sigma}_Y - \Sigma_Y^*\|_{\text{F}} = O_{\text{p}}\left[\left\{\frac{(p+q)\log p}{n}\right\}^{1/2}\right].$$

The rate in Condition 3 is required to control the convergence rate of the E-step estimating Σ_k^* and thus the consistency of the estimate from the EM method. Under the conditions in Theorem 1, we can use the one-step estimator $\hat{\Omega}_k^{\text{one}}$ $(k = 0, \ldots, K)$ to obtain $\tilde{\Sigma}_Y = J \otimes \hat{\Omega}_0^{-1} + \{_d\hat{\Omega}_k^{-1}\}$, where $\{_d\cdot\}$ is a block-diagonal matrix. The resulting $\tilde{\Sigma}_Y$ satisfies Condition 3 by Corollary 1.

Theorem 2 (Consistency of the EM method). *If Conditions* 1–3 *hold and* $(p + q)(\log p)/n = o(1)$, *then after a finite number of iterations, the solution* $\{\hat{\Omega}_k^{\text{EM}}\}_{k=0}^K$ *of the EM method satisfies*

$$\sum_{k=0}^{K} \|\hat{\Omega}_k^{\text{EM}} - \Omega_k^*\|_{\text{F}} = O_{\text{p}}\left[\left\{\frac{(p+q)\log p}{n}\right\}^{1/2}\right].$$

Theorem 3 (Sparsistency of the one-step method). *Under the assumptions of Theorem* 1, *if we further assume that the one-step solution* $\{\hat{\Omega}_k^{\text{one}}\}_{k=0}^K$ *satisfies* $\sum_{k=0}^K \|\hat{\Omega}_k^{\text{one}} - \Omega_k^*\| = O_{\text{p}}(\eta_n)$

for a sequence $\eta_n \to 0$, *and if* $\{(\log p)/n\}^{1/2} + \eta_n = O(\lambda_1) = O(\lambda_2)$, *then with probability tending to* 1, $\hat{\omega}^{\text{one}}_{k(i,j)} = 0$ *for all* $(i, j) \in T^{\text{c}}_k$ $(k = 0, \ldots, K)$.

For sparsistency we require a lower bound on the rates of λ_1 and λ_2, but for consistency we need an upper bound for λ_1 and λ_2 to control the biases. In order to have consistency and sparsistency simultaneously, we need the bounds to be compatible, that is, we need $\{(\log p)/n\}^{1/2} + \eta_n = O(\lambda_1, \lambda_2) = \{(1 + p/q) \log p/n\}^{1/2}$. From the inequalities $\|W\|_F^2/p \leqslant \|W\|^2 \leqslant \|W\|_F^2$, there are two extreme scenarios describing the rate of η_n, as discussed in Lam & Fan (2009). In the worst case, where $\sum_{k=0}^K \|\hat{\Omega}_k - \Omega_k^*\|$ has the same rate as $\sum_{k=0}^K \|\hat{\Omega}_k - \Omega_k^*\|_{\text{F}}$, we achieve both consistency and sparsistency only when $q = O(1)$. In the most optimistic case, where $\sum_{k=0}^K \|\hat{\Omega}_k - \Omega_k^*\|^2 = \sum_{k=0}^K \|\hat{\Omega}_k - \Omega_k^*\|_{\text{F}}^2/p$, we have $\eta_n^2 = (1 + q/p) \log p/n$, and the compatibility of the bounds requires $q = O(p)$.

THEOREM 4 (Sparsistency of the EM method). *Under the assumptions of Theorem 2, if we further assume the EM solution* $\{\hat{\Omega}_k^{\text{EM}}\}_{k=0}^K$ *satisfies* $\sum_{k=0}^K \|\hat{\Omega}_k^{\text{EM}} - \Omega_k^*\| = O_{\text{p}}(\zeta_n)$ *for a sequence* $\zeta_n \to 0$, *and if* $\{(p + q)(\log p)/n\}^{1/2} + \zeta_n = O(\lambda_1) = O(\lambda_2)$, *then with probability tending to* 1, $\hat{\omega}^{\text{EM}}_{k(i,j)} = 0$ *for all* $(i, j) \in T_k^c$ $(k = 0, \ldots, K)$.

Similar to the discussion above for the EM algorithm, we have both consistency and sparsistency when $q = O(1)$. See the Supplementary Material.

4. SIMULATION

4·1. *Simulating category-specific and systemic networks*

We assessed the performance of the one-step and EM methods by applying them to simulated data generated by two types of synthetic networks: a chain network and a nearest-neighbour network as shown in Fig. 2. Twelve simulation settings were considered. These varied the base architecture of the category-specific network, the degree to which the actual structure could deviate from this base architecture, and the number of outcome variables.

Under each of the 12 simulation conditions, samples were independently and identically distributed, with systemic outcomes generated as $Z_i \sim N(0, \Omega_0^{-1})$, category-specific outcomes as $X_{k,i} \sim N(0, \Omega_k^{-1})$, and observed outcomes as $y_{k,i} = x_{k,i} + z_i$, for $K = 4$ and $n = 300$. The following architectures were considered for the five networks $\{\Omega_k\}_{k=0}^4$:

(I) the K category-specific networks are chain-networks and the systemic network is a nearest-neighbour network with the number of neighbours $m = 5$ and 25 for $p = 100$ and 1000;

(II) the K category-specific networks and the systemic network are all nearest-neighbour networks with $m = 5$ and 25 for $p = 100$ and 1000 respectively.

Chain and nearest-neighbour networks were generated using the algorithms in Fan et al. (2009) and Li & Guo (2006). The structures of network (I) are shown in Figs. 2(a) and (d). Simulated networks were allowed to deviate from their base architectures by a specified degree ρ, through a random addition of edges following the method of Guo et al. (2011). Specifically, for each Ω_k $(k = 0, 1, \ldots, K)$ generated above, a symmetric pair of zero elements is randomly selected and replaced with a value generated uniformly from $[-1, -0{\cdot}5] \cup [0{\cdot}5, 1]$. We repeat this procedure ρT times, with T being the number of links in the initial structure and $\rho \in \{0, 0{\cdot}2, 1\}$.

We compared the performance of the one-step and EM methods by examining the average false positive rate, average false negative rate, average Hamming distance, average entropy loss

$$\text{EL} = \frac{1}{K+1} \sum_{k=0}^{K} \left\{ \text{tr}\left(\Omega_k^{*-1} \hat{\Omega}_k\right) - \log \det \left(\Omega_k^{*-1} \hat{\Omega}_k\right) \right\} - p,$$

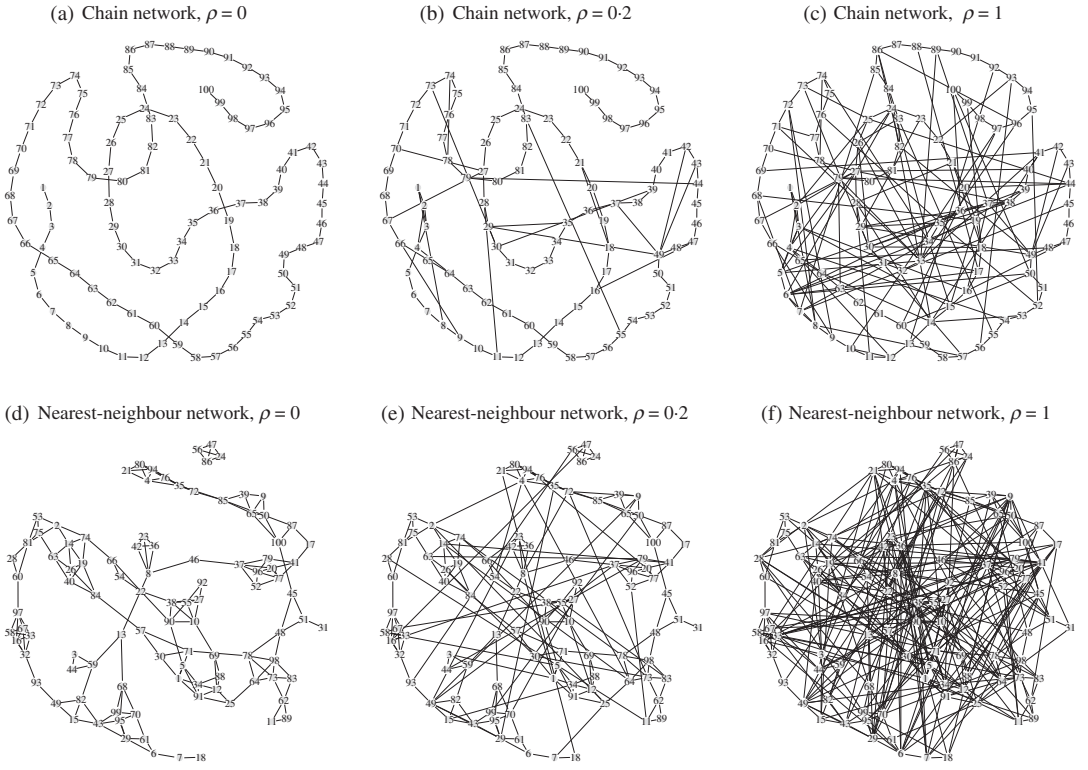

Fig. 2. Network topologies generated in the simulations. Panels (a)–(c) show chain networks with noise ratios $\rho = 0, 0{\cdot}2$, and 1. Panels (d)–(f) show nearest-neighbour networks with $\rho = 0, 0{\cdot}2$, and 1.

and average Frobenius loss

$$\text{FL} = \frac{1}{K+1} \sum_{k=0}^{K} \frac{\|\Omega_k^* - \hat{\Omega}_k\|_{\text{F}}^2}{\|\Omega_k^*\|_{\text{F}}^2}.$$

We also examined receiver operating curves for the two methods.

4·2. *Estimation of category-specific Ω_k and systemic networks Ω_0*

As shown in Fig. 1, existing methods are designed to estimate the aggregate networks Ω_{Y_k} instead of category-specific Ω_k and systemic Ω_0 networks. In this subsection, we focus only on our proposed one-step and EM methods.

Results of the simulations are reported in Table 1. Summary statistics are based on 50 replicate trials under each of the 12 conditions, and given for model fitting under both the extended Bayesian information criterion with $\gamma = 0{\cdot}1$ and crossvalidation. In general, the one-step method under either model selection criterion resulted in higher values of entropy loss, Frobenius loss, false positive rates and Hamming distance. For both methods, crossvalidation tended to choose models with more false positive links but fewer false negative links, leading to a denser graph. For model selection, a rule of thumb is to use crossvalidation when $p > 500$, and to use the extended Bayesian information criterion otherwise.

Receiver operating curves for the one-step and EM methods are plotted in Fig. 3; each is based on 100 replications with the constraint $\lambda_1 = \lambda_2$. Under all settings, the EM method outperforms the one-step method, yielding greater improvements as the structures become more complicated.

Table 1. *Summary statistics reporting performance of the EM and one-step methods inferring graph structure for different networks. The numbers before and after the slash are the results based on the extended Bayesian information criterion and crossvalidation, respectively*

p	Network architecture	ρ	Method	EL	FL	FP (%)	FN (%)	HD (%)
100	(I)	0	One-step	12·1/10·0	0·24/0·16	5·5/20·9	4·2/0·9	5·5/20·4
		0	EM	6·7/4·7	0·15/0·08	4·2/15·8	3·4/0·6	4·2/15·4
		0·2	One-step	10·6/8·6	0·22/0·15	5·4/19·4	3·7/0·9	5·3/18·8
		0·2	EM	6·4/4·8	0·15/0·09	4·9/14·3	3·5/0·6	4·8/ 14·0
		1	One-step	12·6/9·9	0·24/0·17	7·3/23·3	9·5/2·9	7·5/22·3
		1	EM	8·3/6·0	0·17/0·11	6·7/15·3	5·3/1·6	6·6/14·6
	(II)	0	One-step	12·1/9·6	0·27/0·19	3·4/19·6	22·0/7·6	4·1/19·1
		0	EM	7,9/6·0	0·20/0·14	3·8/13·5	12·4/4·2	4·1 13·4
		0·2	One-step	12·5/9·7	0·26/0·18	4·6/21·0	23·0/7·8	5·5/20·4
		0·2	EM	8·7/6·1	0·19/0·12	4·5/15·2	14·1/3·2	5·0/14·6
		1	One-step	16·3/12·6	0·27/0·17	8·7/30·4	24·0/8·8	9·9/28·7
		1	EM	11·3/7·6	0·20/0·11	8·1/22·9	13·7/2·7	8·6/21·4
1000	(I)	0	One-step	276·7/240·6	0·44/0·36	0·6/5·5	52·1/34·6	0·9/5·6
		0	EM	120·3/94·9	0·22/0·16	0·5/2·5	48·9/35·7	0·8/2·7
		0·2	One-step	201·5/162·3	0·35/0·27	0·2/5·0	64·3/37·9	0·6/5·3
		0·2	EM	117·7/88·5	0·19/0·13	0·2/2·2	57·8/39·8	0·6/ 2·5
		1	One-step	171·6/146·0	0·28/0·22	0·0/5·3	100/54·1	1·2/5·9
		1	EM	147·1/108·1	0·20/0·14	0·0/2·3	99·2/56·5	1·2/2·9
	(II)	0	One-step	301·0/234·4	0·43/0·33	0·1/6·7	83·5/53·7	2·0/7·7
		0	EM	206·7/160·9	0·29/0·23	0·2/2·6	73·8/56·4	1·9/3·8
		0·2	One-step	349·8/257·5	0·44/0·31	0·1/8·4	89·2/52·9	2·5/9·6
		0·2	EM	275·0/190·8	0·32/0·23	0·2/3·9	82·7/53·8	2·4/5·2
		1	One-step	325·4/268·8	0·41/0·29	0·0/10·1	99·9/64·3	4·4/12·5
		1	EM	301·6/232·6	0·31/0·23	0·0/4·8	99·8/68·2	4·4/ 7·6

EL, the average entropy loss; FL, the average Frobenius loss; FN, the average false negative rate; FP, the average false positive rate; HD, the average Hamming distance; ρ, the noise ratio.

4·3. *Estimation of aggregate networks* Ω_{Y_k}

Although our goal is to estimate the two network layers, we can also use our estimators of Ω_k $(k = 0, \ldots, K)$ to estimate the aggregate network $\Omega_{Y_k} = (\Omega_k^{-1} + \Omega_0^{-1})^{-1}$ as a derived statistic. Doing so allows us to compare our method with methods that aim to estimate the aggregate network Ω_{Y_k}, as these methods are otherwise incomparable.

We compared the performance of the EM method with two existing single-level methods for estimating multiple graphs: the hierarchical penalized likelihood method of Guo et al. (2011) and the joint graphical lasso of Danaher et al. (2014). As shown by simulation results reported in the Supplementary Material, these two single-level methods tended to give similar, sparse estimates that were very different from the true aggregate graph. The true aggregate graph tended to be highly connected, as illustrated in Fig. 1, and under most settings was much better estimated by the EM method. An exception was setting (II) with $\rho = 0$ and 0·2: here Ω_{Y_k} is relatively sparse, and the best performance came from the method of Guo et al. (2011). Sparsity in Ω_{Y_k} arises under this setting because when Ω_k and Ω_0 are chain networks Ω_{Y_k} has a strong banding structure, with large absolute values within the band and small absolute values outside.

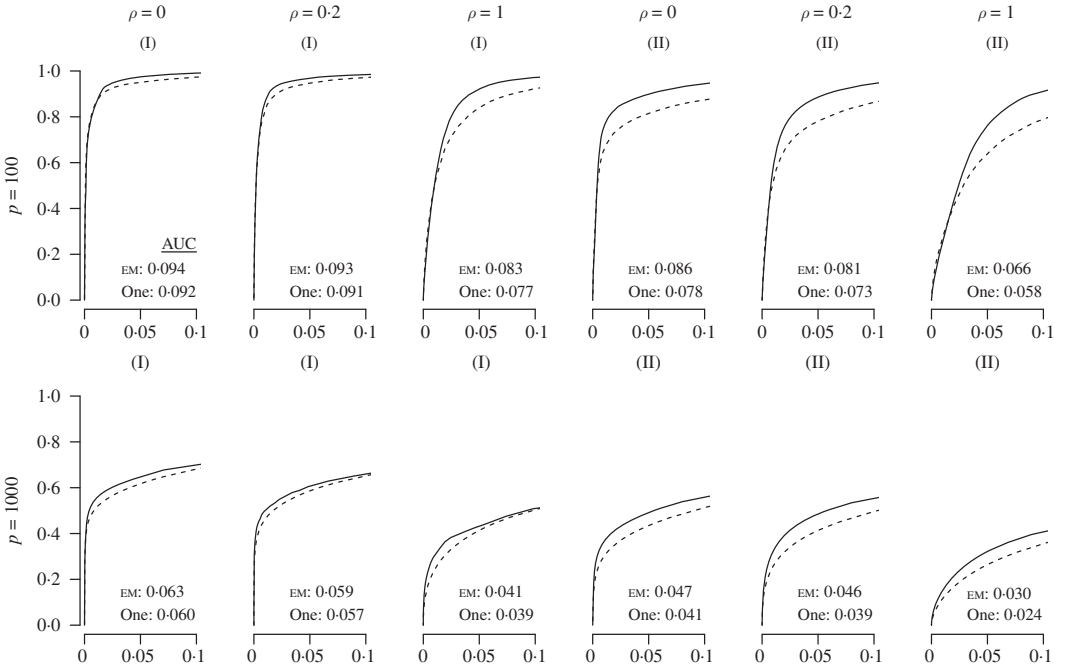

Fig. 3. Receiver operating characteristic curves assessing power and discrimination of graphical inference under different simulation settings. Each panel reports performance of the EM method (solid line) and the one-step method (dashed line), plotting true positive rates (y-axis) against false positive rates (x-axis) for a given noise ratio ρ, network base architecture I or II, sample size $n = 300$, and number of neighbours $m = 5$ and 25 for $p = 100$ and 1000 respectively. The numbers in each panel represent the areas under the curve for the two methods.

5. Application to gene expression data in mice

To demonstrate the potential utility of our approach, we apply the EM method to mouse genomics data from Dobrin et al. (2009) and Crowley et al. (2015). In each case, we aim to infer systemic and category-specific gene co-expression networks from transcript abundance as measured by microarrays. In describing our inference on these datasets we find it helpful to distinguish two interpretations of a network: the potential network is the network of biologically possible interactions in the type of system under study; the induced network is the subgraph of the potential network that could be inferred in the population sampled by the study. The induced network is therefore a statistical, not physical, phenomenon, and describes the dependence structure induced by the interventions, or perturbations, applied to the system.

A simple example is the relationship between caloric intake, sex, and body weight. Body weight is influenced by both the state of being male or female and the degree of caloric consumption; these relations constitute edges in the potential network. Yet in a population where caloric intake varies but where individuals are exclusively male, the effect of sex is undefined and the corresponding edges relating sex to body weight are undetectable; these edges are therefore absent in the induced network. More generally, the induced network for a system is defined by both the potential network and the intervention applied to it: two populations of mice could have the same potential network, but when subject to different interventions their induced networks could differ. Conversely, when estimating the dependence structure of variables arising from population data, the degree to which the induced network reflects the potential network is a function of the underlying conditions being varied and interventions at work.

The Dobrin et al. (2009) dataset comprises expression measurements for 23 698 transcripts on 301 male mice in adipose, liver, brain and muscle tissues. These mice arose from an F_2 cross between two contrasting inbred founder strains, one with normal body weight physiology and the other with a heritable tendency for rapid weight gain. In a cross of this type, the analysed offspring constitute an independent and identically distributed sample of individuals who are genetically distinct and have effectively been subject to a randomized allocation of normal and weight-inducing DNA variants, or alleles, at multiple locations along its genome. As a result of this allocation, gene expression networks inferred on such a population would be expected to emphasize more strongly those subgraphs of the underlying potential network that are related to body weight. Moreover, since the intervention alters a property affecting the entire individual, we might expect it to exert at least some of its effect systemically, that is, globally across all tissues in each individual.

Using a subset of the data, we inferred the dependence structure of gene co-expression among three groups of well-annotated genes in brain and liver: an obesity-related gene set, an imprinting-related gene set, and an extracellular matrix, i.e., the ECM-related gene set. These groups were chosen based on criteria independent of our analysis and represent three groups whose respective effects would be exaggerated under very different interventions. The tissue-specific and systemic networks inferred by our EM method are shown in Fig. 4. Each node represents a gene, and the darkness of an edge represents the magnitude of the associated partial correlation. The systemic network in Fig. 4(c) includes edges on the *Aif1* obesity-related pathway only, which is consistent with the F_2 exhibiting a dependence structure induced primarily by an obesity-related genetic intervention that acts systemically. The category-specific networks in Figs. 4(a) and (b) still include part of the *Aif1* pathway, suggesting that variation in this pathway tracks variation at both the systemic and the tissue-specific level; in other ways their dependence structures differ, with, for instance, *Aif1* and *Rgl2* being linked in the brain but not in the liver. The original analysis of Dobrin et al. (2009) used a correlation network approach, whereby unconditional correlations with statistical significance above a predefined threshold were declared as edges; that analysis also supported a role for *Aif1* in tissue-to-tissue co-expression.

The Crowley et al. (2015) data comprise expression measurements of 23 000 transcripts in brain, liver, lung and kidney tissues in 45 mice arising from three independent reciprocal F_1 crosses. A reciprocal F_1 cross between two inbred strains A and B generates two subpopulations: the progeny of strain-A females mated to strain-B males denoted by $A \times B$, and the progeny of strain-B females and strain-A males, denoted by $B \times A$. Across the two progeny groups, the set of alleles inherited is identical, with each mouse having inherited half of its alleles from A and the other half from B; but the route through which those alleles were inherited differs, with, for example, $A \times B$ offspring inheriting their A alleles only from their fathers and $B \times A$ inheriting them only from their mothers. The underlying intervention in a reciprocal cross is therefore not the varying of genetics as such but the varying of parent-of-origin, or epigenetics, and so we might expect some of this epigenetic effect to manifest across all tissues.

We applied our EM method to a normalized subset of the Crowley et al. (2015) data, restricting attention to brain and liver, and removing gross effects of genetic background. Our analysis identified three edges on the systemic network as shown in Fig. 5(c) that include the genes *Igf2*, *Tab1*, *Nrk* and *Pde4b*, all from the imprinting-related set implicated in mediating epigenetic effects. Thus, the inferred patterns of systemic-level gene relationships in the two studies coincide with the underlying interventions implied by the structure of those studies, with genes affecting body weight in the Dobrin et al. (2009) data and genes affected by parent-of-origin in the Crowley et al. (2015) data.

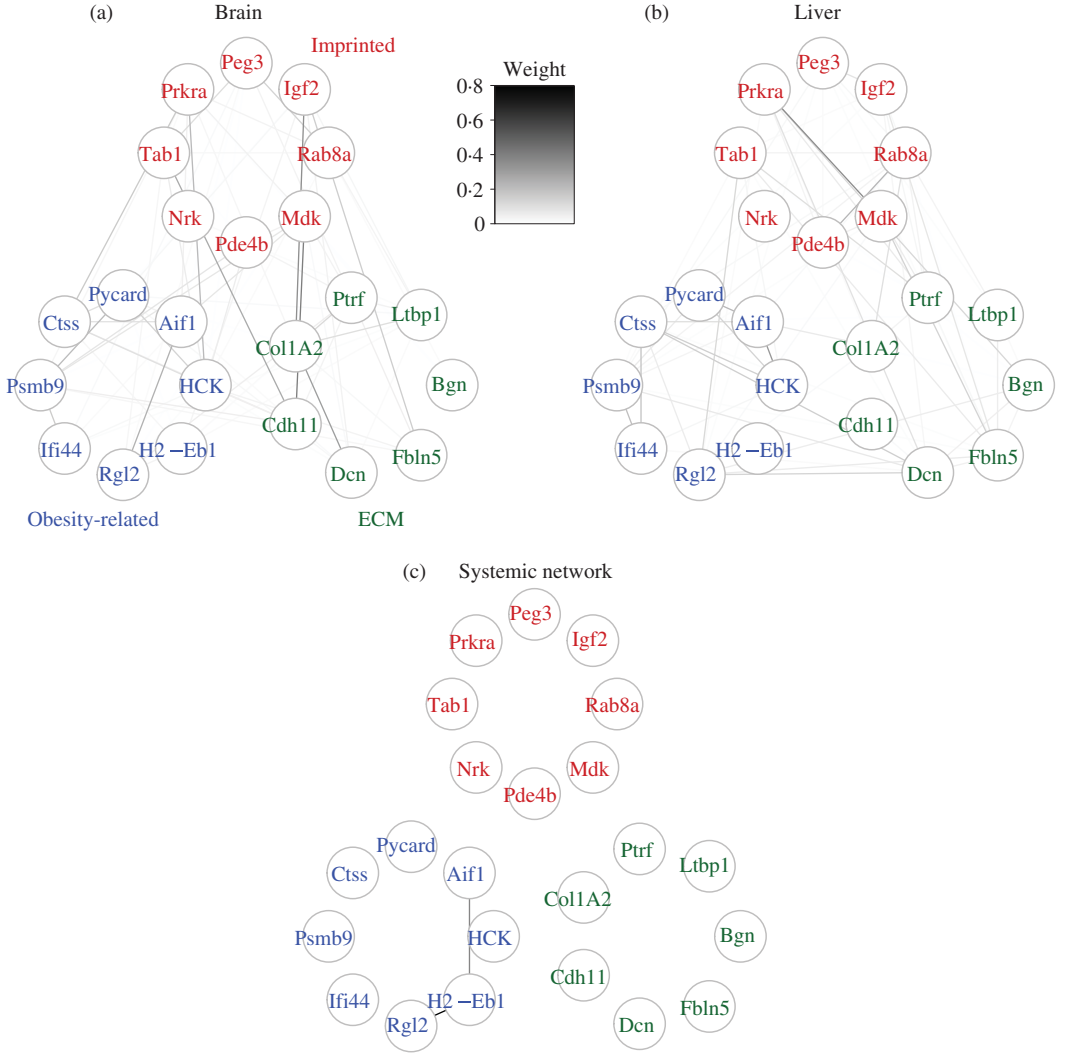

Fig. 4. Topology of gene co-expression networks inferred by the EM method for the data from a population of F_2 mice with randomly allocated high-fat versus normal gene variants. Panels (a) and (b) display the estimated brain-specific and liver-specific dependence structures. Panel (c) shows the estimated systemic structure describing whole-body interactions that simultaneously affect variables in both tissues.

To demonstrate the use of our method for higher-dimensional data, we examined a larger subset of genes from Dobrin et al. (2009). Selecting the $p = 1000$ genes that had the largest within-group variance among the four tissues in the F_2 population, we applied our graphical EM method, using the extended Bayesian information criterion to select the tuning parameters λ_1 and λ_2. The existence of a single, nonzero systemic layer for these data was strongly supported by significance testing, as described in the Supplementary Material. The topologies of the estimated tissue-specific and systemic networks are shown in Figs. 6 (a)–(d), with a zoomed-in view of the edges of the systemic network shown in Fig. 6(f). The systemic network is sparse, with 249 edges connecting 62 of the 1000 genes in Fig. 6(e); this sparsity may reflect there being few interactions simultaneously occurring across all tissues in this F_2 population, with one contributing reason being that some genes are expressed primarily in one tissue and not others. The systemic network

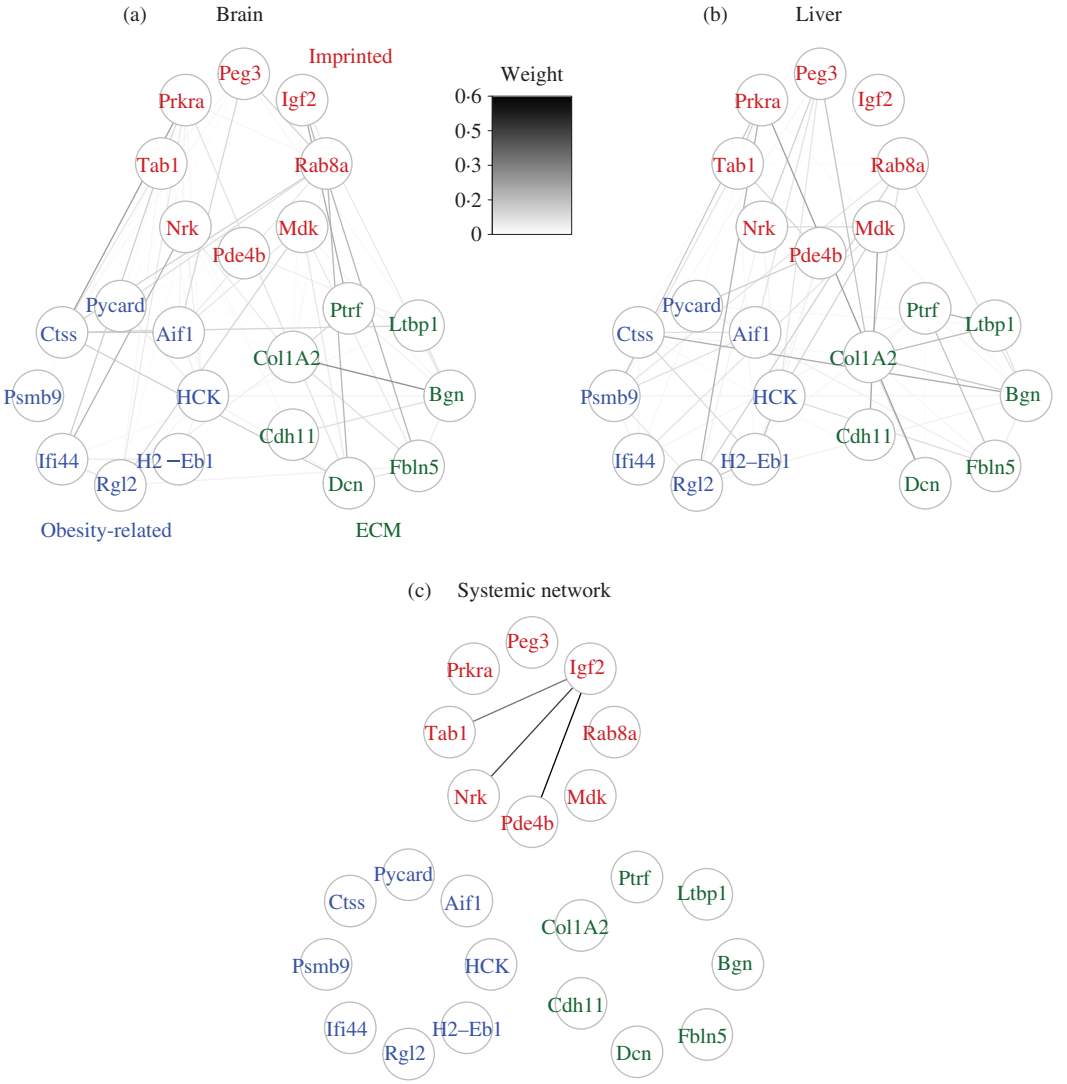

Fig. 5. Topology of gene co-expression networks inferred by the EM method for the data from a population of reciprocal F_1 mice. Panels (a) and (b) display the estimated brain-specific and liver-specific dependence structures. Panel (c) shows the estimated systemic structure describing whole-body interactions that simultaneously affect variables in both tissues.

also includes a connection between two genes, *Ifi44* and *H2-Eb1*, that are members of the *Aif1* network of Fig. 4. To characterize more broadly the genes identified in the systemic network, we conducted an analysis of gene ontology enrichment (Shamir et al., 2005), in which the distribution of gene ontology terms associated with connected genes in the systemic network was contrasted against the background distribution of gene ontology terms in the entire 1000-gene set; this showed that the systemic network is significantly enriched for genes associated with immune and metabolic processes, which accords with recent studies linking obesity to strong negative impacts on immune response to infection (Milner & Beck, 2012; Lumeng, 2013). The original study of Dobrin et al. (2009) also showed the enrichment of inflammatory response processes in co-expression from liver and adipose, again using unconditional correlations.

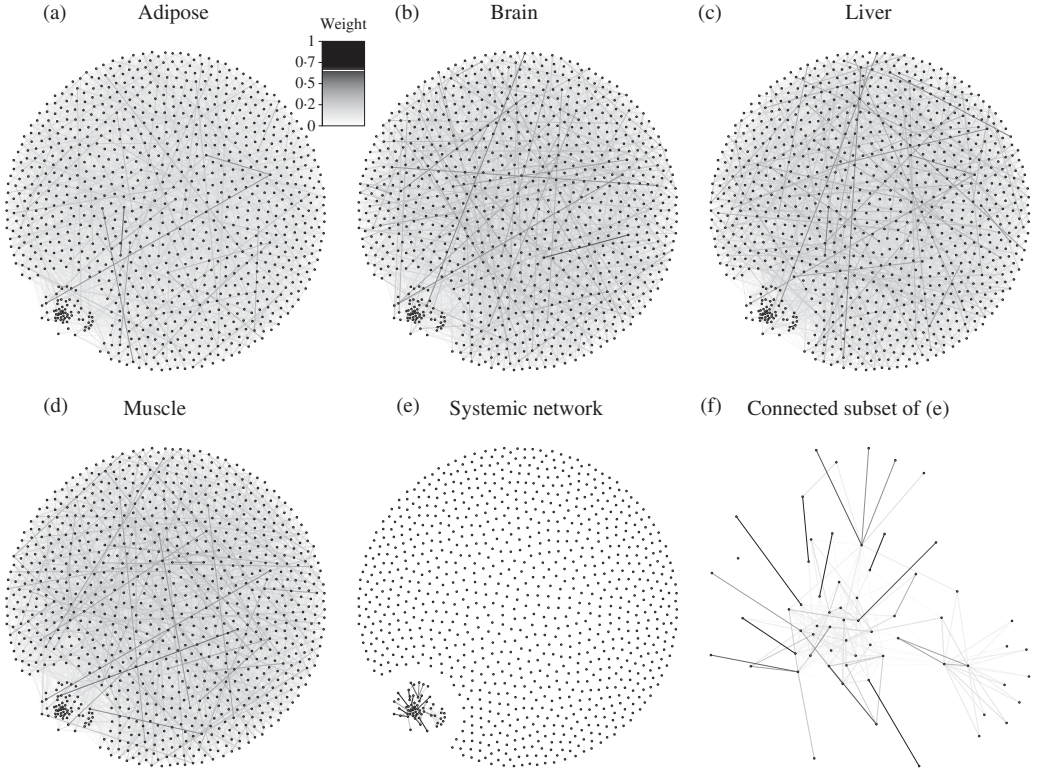

Fig. 6. Topology of co-expression networks inferred by the EM method applied to measurements of the 1000 genes with highest within-tissue variance in a population of F_2 mice. Panels (a)–(d) show category-specific networks estimated for adipose, hypothalamus, liver and muscle tissue. Panel (e) shows the structure of the estimated systemic network, describing across-tissue dependencies, with panel (f) showing a zoomed-in view of the connected subset of nodes in this graph.

6. Discussion

In this paper we consider joint estimation of a two-layer Gaussian graphical model that is different from, but related to, the single-layer model. In our setting, the single-layer model estimates an aggregate graph Ω_{Y_k} by imposing sparsity on Ω_{Y_k} directly. Our model, by contrast, estimates the two graphical layers that compose the aggregate, namely Ω_k and Ω_0, and imposes sparsity on each. This can imply an aggregate graph Ω_{Y_k} that is less sparse; but this is appropriate because in our setting Ω_{Y_k} is a by-product and, as such, is a secondary subject of inference. Importantly, our two-layer model includes the single-layer model as a special case, since in the absence of an appreciable systemic dependence, when $\Sigma_Z = 0$, the two-layer model reduces to a single layer.

Our model lends itself to several immediate extensions. First, we currently assume that the systemic graph affects all tissues equally, but, as suggested by one reviewer, we can extend our model to allow the influence of the systemic layer to vary among tissues. For example, since muscle and adipose tissue are both developed from the mesoderm, we might expect them to be more closely related to each other as compared with the pancreas, which is developed from the endoderm. We can accommodate such variation in our model as

$$Y_{k,i} = X_{k,i} + \alpha_k Z_i \quad (k = 1, \ldots, K;\ i = 1, \ldots, n),$$

where α_k quantifies the level of systemic influence in each tissue k. Our EM algorithm can also be modified to calculate α_k and Ω_k. More details can be found in the Supplementary Material.

Second, we can extend the ℓ_1-penalized maximum likelihood framework to other nonconvex penalties such as the truncated ℓ_1-function (Shen et al., 2012) and the smoothly clipped absolute deviation penalty (Fan & Li, 2001). Furthermore, we believe it would be both practicable and useful to extend these methods beyond the Gaussian assumption (Cai & Liu, 2011; Liu et al., 2012; Xue & Zou, 2012).

ACKNOWLEDGEMENT

The authors thank the editor, the associate editor and two reviewers for their helpful suggestions. This work was supported in part by the U.S. National Institutes of Health and the National Science Foundation. Yuying Xie is also affiliated with the Department of Statistics and Probability at Michigan State University. Yufeng Liu is affiliated with the Department of Genetics and Carolina Center for Genome Sciences, and both he and William Valdar are also affiliated with the Department of Biostatistics and the Lineberger Comprehensive Cancer Center at the University of North Carolina.

SUPPLEMENTARY MATERIAL

Supplementary material at *Biometrika* online includes technical proofs of the theorems, extra simulation results, R code for the EM algorithm, and model diagnosis for the real-data analysis.

REFERENCES

AMBROISE, C., CHIQUET, J. & MATIAS, C. (2009). Inferring sparse Gaussian graphical models with latent structure. *Electron. J. Statist.* **3**, 205–38.

BANERJEE, O., GHAOUI, L. E. & D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation. *J. Mach. Learn. Res.* **9**, 485–516.

CAI, T. & LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Am. Statist. Assoc.* **106**, 672–84.

CHEN, J. & CHEN, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759–71.

CROWLEY, J., ZHABOTYNSKY, V., SUN, W., HUANG, S., PAKATCI, I. K., KIM, Y., WANG, J., MORGAN, A., CALAWAY, J., AYLOR, D. ET AL. (2015). Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nature Genet.* **47**, 353–60.

DANAHER, P., WANG, P. & WITTEN, D. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Statist. Soc.* B **76**, 373–97.

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., LAIRD, M. N. & RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *J. R. Statist. Soc.* B **39**, 1–38.

DOBRIN, R., ZHU, J., MOLONY, C., ARGMAN, C., PARRISH, M., CARLSON, S., ALLAN, M., POMP, D. & SCHADT, E. (2009). Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. *Genome Biol.* **10**, 55.

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. H., HASTIE, T. J. & TIBSHIRANI, R. J. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–41.

GUO, J., LEVINA, E., MICHAILIDIS, G. & ZHU, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98**, 1–15.

HONORIO, J. & SAMARAS, D. (2010). Multi-task learning of Gaussian graphical models. In *Proc. 27th Int. Conf. Mach. Learn. (ICML*-10), J. Fürnkranz & T. Joachims, eds. Haifa, Israel: Omnipress.

HSIEH, C.-J., SUSTIK, M., DHILLON, I. S. & RAVIKUMAR, P. (2014). Quic: Quadratic approximation for sparse inverse covariance estimation. *J. Mach. Learn. Res.* **15**, 2911–47.

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

Lee, W. & Liu, Y. (2015). Estimation of multiple graphical models with common structures. *J. Mach. Learn. Res* **16**, 1035–62.

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

Liu, H., Han, F., Yuan, M., Lafferty, J. & Wasserman, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40**, 2293–326.

Lumeng, C. (2013). Innate immune activation in obesity. *Molec. Aspects Med.* **34**, 12–29.

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

Milner, J. & Beck, M. (2012). The impact of obesity on the immune response to infection. *Proc. Nutr. Soc.* **71**, 298–306.

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

Ravikumar, P., Wainwright, M. J., Raskutti, G. & Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Statist.* **5**, 935–80.

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

Shamir, R., Maron-Katz, A., Tanay, A., Linhart, C., Steinfeld, I., Sharan, R., Shiloh, Y. & Elkon, R. (2005). Expander – an integrative program suite for microarray data analysis. *BMC Bioinformatics* **6**, 232.

Shen, X., Pan, W. & Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *J. Am. Statist. Assoc.* **107**, 223–32.

Varoquaux, G., Gramfort, A., Poline, J.-B. & Thirion, B. (2010). Brain covariance selection: Better individual functional connectivity models using population prior. In *Advances in Neural Information Processing Systems* 23, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel & A. Culotta, eds. New York: Curran Associates, Inc.

Xu, M. & Shao, H. (2012). Solving the matrix nearness problem in the maximum norm by applying a projection and contraction method. *Adv. Oper. Res.* **2012**, 1–15.

Xue, L. & Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.* **40**, 2541–71.

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

Zhou, S., Lafferty, J. D. & Wasserman, L. A. (2010). Time varying undirected graphs. *Mach. Learn.* **80**, 295–319.

[*Received September* 2014. *Revised May* 2016]