*Biometrika* (2013), **100**, 3, *pp.* 655–670
doi: 10.1093/biomet/ast009
Advance Access publication 21 April 2013


# High-dimensional semiparametric bigraphical models

BY YANG NING

*Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada*
yning@jhsph.edu

AND HAN LIU

*Department of Operations Research and Financial Engineering, Princeton University, New Jersey 08544, U.S.A.*
hanliu@princeton.edu

SUMMARY

In multivariate analysis, a Gaussian bigraphical model is commonly used for modelling matrix-valued data. In this paper, we propose a semiparametric extension of the Gaussian bigraphical model, called the nonparanormal bigraphical model. A projected nonparametric rank-based regularization approach is employed to estimate sparse precision matrices and produce graphs under a penalized likelihood framework. Theoretically, our semiparametric procedure achieves the parametric rates of convergence for both matrix estimation and graph recovery. Empirically, our approach outperforms the parametric Gaussian model for non-Gaussian data and is competitive with its parametric counterpart for Gaussian data. Extensions to the categorical bigraphical model and the missing data problem are discussed.

*Some key words*: Bigraphical model; High dimensionality; Matrix-normal distribution; Rank-based statistic.

## 1. INTRODUCTION

The Gaussian bigraphical model, also called the matrix-normal graphical model (Dawid, 1981; Gupta & Nagar, 1999) or the Gaussian Kronecker graphical model (Werner et al., 2008), is commonly used for modelling matrix-valued data. The model assumes that a high-dimensional covariance matrix is separable as the Kronecker product of two low-dimensional component matrices which encode the dependence structures of row and column variables. Owing to its flexibility and interpretability, the Kronecker product covariance model has been widely used to analyse spatiotemporal data (Mardia & Goodall, 1993; Genton, 2007), multivariate data with repeated measurements (Naik & Rao, 2001) and genomic data (Teng & Huang, 2009). Estimation procedures for variance matrices include maximum likelihood estimation methods (Dutilleul, 1999; Lu & Zimmerman, 2005; Mitchell et al., 2005, 2006), empirical Bayes methods (Theobald & Wuttke, 2006) and Bayesian methods (Wang & West, 2009). Hoff (2011a, b) further extended the Bayesian approach to accommodate multi-dimensional data arrays. Most literature on matrix-valued data considers the classical setting, $n > p^2q^2$, where $n$ is the number of replicates, $p$ is the number of rows, and $q$ is the number of columns. Recently, Yin & Li (2012) proposed an $L_1$-penalized likelihood method to estimate sparse precision matrices in the high-dimensional Gaussian bigraphical model, which allows $p$ and $q$ to increase with $n$. They established rates of convergence and sparsistency of lasso-type estimators.

The popularity of the Gaussian bigraphical model is mainly due to its simplicity (Lauritzen, 1996, Ch. 5). However, the normality assumption is rather restrictive. To relax this assumption, Liu et al. (2009) proposed a semiparametric Gaussian copula model. Instead of assuming the data to be Gaussian, they assume that there exists a set of unknown transformations such that the transformed data follow a Gaussian distribution. To estimate the precision matrix, Liu et al. (2012) proposed a rank-based approach, which avoids the estimation of marginal transformations. The resulting estimator achieves the optimal parametric rates of convergence for both matrix estimation and graph recovery.

In this paper, we propose a semiparametric extension of the Gaussian bigraphical model, called the nonparanormal bigraphical model. We show that only row and column correlation matrices are estimable in the model. To infer the graph structure and estimate the precision matrices, we propose using a projected nonparametric rank-based regularization approach under a penalized likelihood framework, without estimating the marginal transformations. A novel projection procedure is introduced to guarantee positive definiteness of the rank-based correlation estimators. From a computational point of view, the full data are summarized by a single rank-based correlation matrix, whereas currently available algorithms for maximizing the likelihood require the full dataset as input. To calculate our estimators, we develop an iterative algorithm based on a new representation proposition. The convergence properties of the proposed algorithm are established. We also obtain the rates of convergence of the proposed matrix estimators, which are identical to the parametric rates obtained from the Gaussian bigraphical model. In addition, we show that our method yields faster rates of convergence than the best results in Yin & Li (2012) for the Gaussian bigraphical model. As a by-product, we give the rates of convergence for estimating the composite precision matrix in both the Frobenius norm and the spectral norm. The sparsistency of the proposed estimator is established. Furthermore, we illustrate how to extend our method to the categorical bigraphical model. An EM algorithm of normal-score type is proposed for missing data imputation, which extends the algorithm developed by Allen & Tibshirani (2010) for the Gaussian bigraphical model.

## 2. BACKGROUND

### 2·1. *Notation*

We adopt the following notation throughout this paper. For $v = (v_1, \ldots, v_d)^{\mathrm{T}} \in \mathbb{R}^d$ and $1 \leqslant q \leqslant \infty$, we define $\|v\|_q = (\sum_{i=1}^d |v_i|^q)^{1/q}$ and $\|v\|_\infty = \max_{1 \leqslant i \leqslant d} |v_i|$. For any $p \times q$ matrix $M = (M_{jk})$, let $M^{\mathrm{T}}$ denote the matrix transpose of $M$ and $\mathrm{vec}(M)$ the vectorization of $M$, i.e., $\mathrm{vec}(M) = (M_{11}, \ldots, M_{p1}, M_{12}, \ldots, M_{p2}, \ldots, M_{1q}, \ldots, M_{pq})^{\mathrm{T}}$. Let $f = (f_{jk})$ be a matrix of functions with elements $f_{jk}$, and let $f(X)$ be a matrix with elements $f_{jk}(X_{jk})$. Let $M_{i*}$ denote the $i$th row of $M$, $M_{*j}$ the $j$th column of $M$, $M_{-\{i,j\}*}$ the submatrix of $M$ with $i$th and $j$th rows removed, and $M_{*-\{i,j\}}$ the submatrix of $M$ with $i$th and $j$th columns removed. The matrix spectral norm, elementwise maximum norm and Frobenius norm of $M$ are defined, respectively, by $\|M\|_{\mathrm{s}} = \max(\|Mx\|_2/\|x\|_2 : x \in \mathbb{R}^q, x \neq 0)$, $\|M\|_{\max} = \max(|M_{ij}|)$ and $\|M\|_{\mathrm{F}} = (\sum_{i,j} M_{ij}^2)^{1/2}$. For a square matrix $M$, the smallest and largest eigenvalues of $M$ are denoted by $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$. Let $A \otimes B$ be the Kronecker product and $A \circ B$ the Hadamard product of matrices $A$ and $B$. The $d \times d$ identity matrix is denoted by $I_d$.

### 2·2. *The matrix-normal distribution and Gaussian bigraphical model*

A $p \times q$ random matrix $X$ follows a matrix-normal distribution $\mathrm{MN}(M; U, V)$, with mean matrix $M$ and row and column component covariance matrices $U$ and $V$, if and only if the density of $X$ is $\mathrm{pr}(X) = k(U, V) \exp[-\mathrm{tr}\{(X - M)^{\mathrm{T}} U^{-1} (X - M) V^{-1}/2\}]$, where

$k(U, V) = (2\pi)^{-pq/2}|U|^{-q/2}|V|^{-p/2}$. An equivalent representation of $X \sim \mathrm{MN}(M; U, V)$ is that $\mathrm{vec}(X) \sim N_{pq}\{\mathrm{vec}(M), V \otimes U\}$. Let $A = U^{-1}$ and $B = V^{-1}$ be the precision matrices of the row and column variables, respectively. By Chapter 5 of Lauritzen (1996), the precision matrices $A$ and $B$ encode the conditional independence structures of row and column variables, respectively; that is, the $i$th and $j$th rows $X_{i*}$ and $X_{j*}$ are independent, given the remaining rows $X_{-\{i,j\}*}$, if and only if $A_{ij} = 0$. Similarly, the $i$th and $j$th columns $X_{*i}$ and $X_{*j}$ are independent, given the remaining columns $X_{*-\{i,j\}}$, if and only if $B_{ij} = 0$.

### 2·3. *Rank-based estimation in the Gaussian graphical model*

As a semiparametric extension of the Gaussian graphical model, Liu et al. (2009) introduced the nonparanormal graphical model. A random vector $X = (X_1, \ldots, X_d)^{\mathrm{T}}$ satisfies a nonparanormal distribution, $X \sim \mathrm{NPN}(0, \Sigma, f)$, if and only if there exists a set of monotonic transformations $f = (f_j)_{j=1}^d$ such that $f(X) = \{f_1(X_1), \ldots, f_d(X_d)\}^{\mathrm{T}} \sim N_d(0, \Sigma)$ with $\mathrm{diag}(\Sigma) = (1, \ldots, 1)$. Given $n$ independent observations $X_1, \ldots, X_n$ where $X_i = (X_{i1}, \ldots, X_{id}) \sim \mathrm{NPN}(0, \Sigma, f)$, the aim is to estimate the precision matrix $\Omega = \Sigma^{-1}$ which encodes the conditional independence structure. To this end, Liu et al. (2009) suggested a normal-score method; however, the rate of convergence obtained for estimating $\Omega$ is not optimal. The same model was also considered by Hoff (2007), who proposed a Bayesian approach based on the marginal rank likelihood, which is free of the nuisance transformations $f(\cdot)$ but does not have an analytical form. Liu et al. (2012) proposed a rank-based approach, with which rank-based correlations such as Spearman's rho and Kendall's tau are used to estimate $\Sigma$ directly, by virtue of their invariance under monotonic transformations. We define Spearman's rho and Kendall's tau as

$$\hat{\rho}_{jk} = \frac{\sum_{i=1}^n (r_{ij} - \bar{r}_j)(r_{ik} - \bar{r}_k)}{\{\sum_{i=1}^n (r_{ij} - \bar{r}_j)^2 \sum_{i=1}^n (r_{ik} - \bar{r}_k)^2\}^{1/2}}, \tag{1}$$

$$\hat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \leqslant i < i' \leqslant n} \mathrm{sign}\{(X_{ij} - X_{i'j})(X_{ik} - X_{i'k})\}, \tag{2}$$

where $r_{ij}$ is the rank of $X_{ij}$ among $X_{1j}, \ldots, X_{nj}$ and $\bar{r}_j = n^{-1} \sum_{i=1}^n r_{ij} = (n+1)/2$. The correlation matrix $\Sigma$ can be estimated by $\hat{R}^\rho = (\hat{R}^\rho_{jk})$ or $\hat{R}^\tau = (\hat{R}^\tau_{jk})$, where

$$\hat{R}^\rho_{jk} = \begin{cases} 2 \sin\left(\frac{\pi}{6}\hat{\rho}_{jk}\right), & j \neq k, \\ 1, & j = k, \end{cases} \qquad \hat{R}^\tau_{jk} = \begin{cases} \sin\left(\frac{\pi}{2}\hat{\tau}_{jk}\right), & j \neq k, \\ 1, & j = k. \end{cases} \tag{3}$$

Once an estimate of $\Sigma$ has been obtained, it can be inserted into any matrix estimation procedure for the Gaussian graphical model (Yuan, 2010; Cai et al., 2011; Friedman et al., 2008). Liu et al. (2012) showed that such a procedure achieves the optimal parametric rates for parameter estimation and graph recovery.

## 3. Nonparanormal bigraphical model

### 3·1. *Definition and identifiability condition*

We start with the definition of a matrix-nonparanormal distribution.

Definition 1. *A $p \times q$ random matrix $X$ follows a matrix-nonparanormal distribution* $\mathrm{MNPN}(M; U, V; f)$*, with mean matrix $M$, row covariance component matrix $U$ and column*

*covariance component matrix $V$, if and only if there exists a set of monotonic transformations $f = (f_{jk})$ such that*

$$\text{vec}\{f(X)\} = \text{vec}[\{f_{jk}(X_{jk})\}] \sim N_{pq}\{\text{vec}(M), V \otimes U\}.$$

The choices $f(x) = x$ and $f(x) = \log(x)$ yield the matrix-normal distribution and the matrix-lognormal distribution, respectively. Since we only require that the $f(\cdot)$ be monotone, the matrix-nonparanormal distribution provides a much richer family of distributions than does the matrix-normal distribution. Indeed, the matrix-nonparanormal distribution can be viewed as a latent-variable model, where the latent variables $f(X)$ follow a matrix-normal distribution and must be symmetric, while the observed variables $X$ need not be symmetric. Let $A = U^{-1}$ and $B = V^{-1}$ be the precision matrices of the row and column variables, respectively. Following arguments similar to those of Yin & Li (2012), we can show that the sparsity patterns in $A$ and $B$ represent the conditional independence structures of the row and column variables.

PROPOSITION 1. *Let $X \sim \text{MNPN}(M; U, V; f)$, and let $A = U^{-1}$ and $B = V^{-1}$. The $i$th and $j$th rows $X_{i*}$ and $X_{j*}$ are independent given the remaining rows $X_{-\{i,j\}*}$ if and only if $A_{ij} = 0$. Similarly, the $i$th and $j$th columns $X_{*i}$ and $X_{*j}$ are independent given the remaining columns $X_{*-\{i,j\}}$ if and only if $B_{ij} = 0$.*

If $f(\cdot)$ is differentiable, the joint probability density function of $X$ is

$$\begin{aligned}&\text{pr}(X \mid M, U, V, f)\\ &\quad = k(U, V) \exp\left(-\frac{1}{2}\text{tr}[\{f(X) - M\}^{\mathrm{T}} U^{-1} \{f(X) - M\} V^{-1}]\right) \prod_{i=1}^{p} \prod_{r=1}^{q} |f'_{ir}(X_{ir})|,\end{aligned}$$

where $k(U, V) = (2\pi)^{-pq/2} |U|^{-q/2} |V|^{-p/2}$.

The model in Definition 1 is not identifiable. The distribution remains the same if $f(X)$ and $M$ are replaced by $f(X) - K$ and $M - K$, respectively, with $K \in \mathbb{R}^{p \times q}$. To make the model identifiable, we impose the constraint that $M = 0$. We get the same distribution if $\text{vec}\{f(\cdot)\}_j$ and $\text{diag}(V \otimes U)_j$ are replaced by $c\,\text{vec}\{f(\cdot)\}_j$ and $c^{-2}\,\text{diag}(V \otimes U)_j$ where $c$ is any positive scalar, so we can let $\text{diag}(V \otimes U) = (1, \ldots, 1)$. Moreover, $\text{pr}(X \mid U, V, f) = \text{pr}(X \mid c^2 U, V, cf) = \text{pr}(X \mid U, c^2 V, cf)$. We then set $V_{11} = 1$. These two conditions together imply that $\text{diag}(V) = (1, \ldots, 1)$ and $\text{diag}(U) = (1, \ldots, 1)$. Hence we can assume that $U$ and $V$ are correlation matrices. With these identifiability conditions, the matrix-nonparanormal distribution is denoted by $\text{MNPN}(U, V; f)$ with $\text{diag}(U) = (1, \ldots, 1)$ and $\text{diag}(V) = (1, \ldots, 1)$.

### 3·2. *Estimation*

We now consider estimation of the precision matrices $A = U^{-1}$ and $B = V^{-1}$ based on $n$ independent matrix-valued random variables $X_1, \ldots, X_n$, where $X_i \sim \text{MNPN}(U, V; f)$. We enforce sparsity on $A$ and $B$ by regularization, so $A$ and $B$ can be estimated by minimizing the $L_1$-penalized negative loglikelihood

$$\begin{aligned}w\{A, B, f(\cdot)\} = -q \log|A| - p \log|B| + \frac{1}{n} \sum_{i=1}^{n} \text{tr}\{f(X_i)^{\mathrm{T}} A f(X_i) B\}\\ + \lambda \sum_{i \neq j} |A_{ij}| + \gamma \sum_{i \neq j} |B_{ij}|,\end{aligned}$$

where $\lambda$ and $\gamma$ are tuning parameters. To obtain fast rates of convergence, we do not penalize the diagonal elements of $A$ and $B$ (Rothman et al., 2008). The dependence of $w\{A, B, f(\cdot)\}$ on the functions $f(\cdot)$ complicates the minimization procedure. To avoid estimation of $f(\cdot)$, we extend the rank-based approach of Liu et al. (2012). Let $Y_i = (Y_{i1}, \ldots, Y_{id}) = \mathrm{vec}(X_i)$, where $d = pq$. Spearman's rho and Kendall's tau statistics are given by (1) and (2), where $r_{ij}$ is the rank of $Y_{ij}$ among $Y_{1j}, \ldots, Y_{nj}$. The correlation matrix $\Sigma = V \otimes U$ can be estimated by $\hat{R} = \hat{R}^{\rho}$ or $\hat{R}^{\tau}$ as in (3). To obtain an estimate of $(A, B)$ without estimating $f(\cdot)$, one can minimize the objective function $\phi(A, B) = -q \log|A| - p \log|B| + \mathrm{tr}\{(B \otimes A)\hat{R}\} + \lambda \sum_{i \neq j} |A_{ij}| + \gamma \sum_{i \neq j} |B_{ij}|$, where $\hat{R}$ is either $\hat{R}^{\rho}$ or $\hat{R}^{\tau}$. However, one potential problem with the rank-based estimator is that $\hat{R}$ may not be positive definite. Since we do not penalize the diagonal elements of $A$ and $B$, the diagonal elements of the minimizer of $\phi(A, B)$ can diverge to infinity. To further regularize the estimator, we propose a new projection procedure. We project $\hat{R}$ to the space of positive-definite matrices:

$$\hat{R}_{\mathrm{p}} = \underset{R \in \mathcal{P}_{pq}}{\arg\min} \|\hat{R} - R\|_{\max}, \tag{4}$$

where $\mathcal{P}_d$ denotes the space of $d \times d$ positive-definite matrices. The calculation of $\hat{R}_{\mathrm{p}}$ can be based on a smoothed approximation method; see Nesterov (2005) for details. Given the projected rank-based estimator $\hat{R}_{\mathrm{p}}$, we suggest the projected $L_1$-penalized negative loglikelihood

$$\phi_{\mathrm{p}}(A, B) = -q \log|A| - p \log|B| + \mathrm{tr}\{(B \otimes A)\hat{R}_{\mathrm{p}}\} + \lambda \sum_{i \neq j} |A_{ij}| + \gamma \sum_{i \neq j} |B_{ij}|.$$

For the Gaussian bigraphical model, Yin & Li (2012) proposed an iterative algorithm to minimize the $L_1$-penalized negative loglikelihood with respect to matrices $A$ and $B$, but their algorithm required the full set of data $X_1, \ldots, X_n$ as input. Here the data are summarized by the projected rank-based estimator $\hat{R}_{\mathrm{p}}$ in $\phi_{\mathrm{p}}(A, B)$, and the matrices $A$ and $B$ are entangled together in the Kronecker product. To minimize $\phi_{\mathrm{p}}(A, B)$, the following representation proposition is crucial.

PROPOSITION 2. *Let $K_\ell$ be a $pq \times q$ matrix whose $\{\ell + p(j-1), j\}$th element is* 1 *and other elements are* 0*, where $j = 1, \ldots, q$ and $\ell = 1, \ldots, p$. Let $L_\ell$ be a $p \times pq$ matrix whose $\ell$th $p \times p$ submatrix is $I_p$ and other elements are* 0*, where $\ell = 1, \ldots, q$. Then*

$$\mathrm{tr}\{(B \otimes A)\hat{R}_{\mathrm{p}}\} = \mathrm{tr}(\hat{R}_B A) = \mathrm{tr}(\hat{R}_A B),$$

*where $\hat{R}_B$ is a $p \times p$ matrix whose $(\ell, m)$th element is $\mathrm{tr}(K_\ell B K_m^{\mathrm{T}} \hat{R}_{\mathrm{p}})$ and $\hat{R}_A$ is a $q \times q$ matrix whose $(\ell, m)$th element is $\mathrm{tr}(L_\ell^{\mathrm{T}} A L_m \hat{R}_{\mathrm{p}})$.*

The proof is presented in the Appendix. From Proposition 2 we develop the following projected rank-based bigraphical lasso algorithm.

*Step* 1. Calculate $\hat{R}_{\mathrm{p}}$ in (4) using the method of Nesterov (2005).

*Step* 2. Set $\hat{B}^{(1)} = I_q$ and $k = 1$.

*Step* 3. Given the current estimate $\hat{B}^{(k)}$, the estimate of $A$ is

$$\hat{A}^{(k+1)} = \underset{A}{\arg\min} \left\{ \{-q \log|A| + \mathrm{tr}(\hat{R}_1^{(k)} A) + \lambda \sum_{i \neq j} |A_{ij}| \right\},$$

where the $(\ell, m)$th element of $\hat{R}_1^{(k)}$ is $\mathrm{tr}(K_\ell \hat{B}^{(k)} K_m^{\mathrm{T}} \hat{R}_{\mathrm{p}})$ $(\ell, m = 1, \ldots, p)$.

*Step* 4. Given the current estimate $\hat{A}^{(k+1)}$, the estimate of $B$ is

$$\hat{B}^{(k+1)} = \arg\min_{B} \Big\{ -p \log|B| + \mathrm{tr}(\hat{R}_2^{(k+1)} B) + \gamma \sum_{i \neq j} |B_{ij}| \Big\},$$

where the $(\ell, m)$th element of $\hat{R}_2^{(k+1)}$ is $\mathrm{tr}(L_\ell^{\mathrm{T}} \hat{A}^{(k+1)} L_m \hat{R}_{\mathrm{p}})$ $(\ell, m = 1, \ldots, q)$.

*Step* 5. Repeat Steps 3 and 4 until $\|\hat{A}^{(k+1)} - \hat{A}^{(k)}\|_{\mathrm{F}} + \|\hat{B}^{(k+1)} - \hat{B}^{(k)}\|_{\mathrm{F}} < \epsilon$, where $\epsilon$ is a small positive number. Then set $(\hat{A}, \hat{B}) = (\hat{A}^{(k)}, \hat{B}^{(k)})$.

The minimizations in Steps 3 and 4 can be solved using the R functions glasso (Friedman et al., 2008; Witten et al., 2011) and huge (Zhao et al., 2012). Since $\phi_{\mathrm{p}}(A, B)$ is not a convex function for $(A, B)$ jointly, the algorithm is not guaranteed to reach the global minimum. The convergence properties of the algorithm are shown in Theorem 1, whose proof is given in the Supplementary Material.

THEOREM 1. *For $k = 1, 2, \ldots$, the sequence $(\hat{A}^{(k)}, \hat{B}^{(k)})$ generated from the projected rank-based bigraphical lasso algorithm satisfies*

$$\lim_{k \to \infty} \Big( \|\hat{A}^{(k+1)} - \hat{A}^{(k)}\|_{\mathrm{F}} + \|\hat{B}^{(k+1)} - \hat{B}^{(k)}\|_{\mathrm{F}} \Big) = 0.$$

*Moreover, the accumulation point of $(\hat{A}^{(k)}, \hat{B}^{(k)})$ is a stationary point of $\phi_{\mathrm{p}}(A, B)$.*

Following arguments like those in the proof of Theorem 1, we can also establish the convergence properties of the algorithm in Yin & Li (2012).

## 4. ASYMPTOTIC PROPERTIES

Since $\phi_{\mathrm{p}}(A, B)$ is not convex, in this section we establish the existence of a local minimizer with a certain rate of convergence. Let $U_0$ and $V_0$ be the true row and column correlation matrices, and let $A_0 = U_0^{-1} = (A_{ij}^{(0)})$ and $B_0 = V_0^{-1} = (B_{ij}^{(0)})$ be the true row and column precision matrices. Moreover, we write $S_A = \{(i, j) : A_{ij}^{(0)} \neq 0\}$ and $S_B = \{(i, j) : B_{ij}^{(0)} \neq 0\}$ for the supports of the true row and column precision matrices, respectively. Let $s_1$ and $s_2$ be the number of nonzero off-diagonal elements of $A_0$ and $B_0$. The magnitudes of $s_1$ and $s_2$ represent the degrees of sparsity of $A_0$ and $B_0$. Let $\Sigma_0 = V_0 \otimes U_0$. The concentration result for the projected rank-based estimator $\hat{R}_{\mathrm{p}}$ is given in Theorem 2.

THEOREM 2. *Given the projected rank-based estimator $\hat{R}_{\mathrm{p}}$ in* (4)*, for n large enough and $t > 0$ we have*

$$\mathrm{pr}\Big( \|\hat{R}_{\mathrm{p}} - \Sigma_0\|_{\max} \leqslant 16\pi t \Big) \geqslant 1 - p^2 q^2 \exp(-nt^2).$$

This theorem implies that $\|\hat{R}_{\mathrm{p}} - \Sigma_0\|_{\max} = O_{\mathrm{p}}[\{\log(pq)/n\}^{1/2}]$. Hereafter, we assume the following regularity conditions.

*Condition* 1. There exist constants $\delta_1$ and $\delta_2$ such that $0 < \delta_1 < \lambda_{\min}(A_0) \leqslant \lambda_{\max}(A_0) < \delta_2 < \infty$.

*Condition* 2. There exist constants $\delta_3$ and $\delta_4$ such that $0 < \delta_3 < \lambda_{\min}(B_0) \leqslant \lambda_{\max}(B_0) < \delta_4 < \infty$.

*Condition* 3. The tuning parameter $\lambda$ satisfies

$$\lambda = O\left[q\left(1+\frac{s_2}{q}+\frac{ps_2}{qs_1}\right)^{1/2}\left\{\frac{\log(pq)}{n}\right\}^{1/2}\right], \quad q\left(1+\frac{s_2}{q}\right)^{1/2}\left\{\frac{\log(pq)}{n}\right\}^{1/2} = O(\lambda).$$

*Condition* 4. The tuning parameter $\gamma$ satisfies

$$\gamma = O\left[p\left(1+\frac{s_1}{p}+\frac{qs_1}{ps_2}\right)^{1/2}\left\{\frac{\log(pq)}{n}\right\}^{1/2}\right], \quad p\left(1+\frac{s_1}{p}\right)^{1/2}\left\{\frac{\log(pq)}{n}\right\}^{1/2} = O(\gamma).$$

Conditions 1 and 2 provide upper and lower bounds for the eigenvalues of $A_0$ and $B_0$. Similar assumptions were made by Yin & Li (2012) for analysing the Gaussian bigraphical model, and by Lam & Fan (2009) and Rothman et al. (2008) for analysing the Gaussian graphical model. Conditions 3 and 4 provide upper and lower bounds for $\lambda$ and $\gamma$. We find that $\lambda$ and $\gamma$ cannot be too large, or the estimator will be substantially biased and even inconsistent due to the presence of the $L_1$ penalty. On the other hand, the tuning parameters cannot be too small, or the resulting estimator will not be sparse. The rates of convergence are given in Theorem 3, whose proof is outlined in the Appendix.

THEOREM 3. *Under Conditions* 1–4*, there exists a local minimizer* $(\hat{A}, \hat{B})$ *of* $\phi_{\mathrm{p}}(A, B)$ *such that as* $n \to \infty$*,*

$$\|\hat{A} - A_0\|_{\mathrm{F}} = O_{\mathrm{p}}\left\{\left(\frac{\log p + \log q}{n}\right)^{1/2}\left(\frac{s_1s_2 + ps_2 + qs_1}{q}\right)^{1/2}\right\},$$

$$\|\hat{B} - B_0\|_{\mathrm{F}} = O_{\mathrm{p}}\left\{\left(\frac{\log p + \log q}{n}\right)^{1/2}\left(\frac{s_1s_2 + ps_2 + qs_1}{p}\right)^{1/2}\right\} \tag{5}$$

*if* $(s_1s_2 + ps_2 + qs_1)(nq)^{-1}\log(pq) = o(1)$ *and* $(s_1s_2 + ps_2 + qs_1)(np)^{-1}\log(pq) = o(1)$.

The rates of convergence in the Gaussian bigraphical model, as a special case of the nonparanormal bigraphical model, were considered by Yin & Li (2012). They assume $X_1, \ldots, X_n \sim \mathrm{MN}(0, A_{\mathrm{c}}^{-1}, B_{\mathrm{c}}^{-1})$, where $A_{\mathrm{c}}$ and $B_{\mathrm{c}}$ are the inverse row and column covariance matrices. The estimates of $A_{\mathrm{c}}$ and $B_{\mathrm{c}}$ are obtained by minimizing $w(A, B, f)$, with $f$ consisting of identity functions. As in Theorem 3, we can establish the rates of convergence in the Gaussian bigraphical model.

COROLLARY 1. *Let* $X_1, \ldots, X_n \sim \mathrm{MN}(0, A_{\mathrm{c}}^{-1}, B_{\mathrm{c}}^{-1})$*. Under Conditions* 1–4*, there exists a local minimizer* $(\tilde{A}, \tilde{B})$ *of* $w(A, B, f)$*, with* $f$ *consisting of identity functions, such that as* $n \to \infty$*,*

$$\|\tilde{A} - A_{\mathrm{c}}\|_{\mathrm{F}} = O_{\mathrm{p}}\left[\left(\frac{\log p + \log q}{n}\right)^{1/2}\left\{\frac{(p+s_1)(q+s_2)}{q}\right\}^{1/2}\right],$$

$$\|\tilde{B} - B_{\mathrm{c}}\|_{\mathrm{F}} = O_{\mathrm{p}}\left[\left(\frac{\log p + \log q}{n}\right)^{1/2}\left\{\frac{(p+s_1)(q+s_2)}{p}\right\}^{1/2}\right], \tag{6}$$

*if* $(p+s_1)(q+s_2)(nq)^{-1}\log(pq) = o(1)$ *and* $(p+s_1)(q+s_2)(np)^{-1}\log(pq) = o(1)$.

*Remark* 1. Recall that the rate derived by Yin & Li (2012) for estimating $A$ in the Gaussian bigraphical model is

$$\|\tilde{A} - A_{\mathrm{c}}\|_{\mathrm{F}} = O_{\mathrm{p}}\left\{\left(\frac{\log p + \log q}{n}\right)^{1/2}(p + s_1)^{1/2}q^{1/2}\right\}, \tag{7}$$

which is equivalent to (6) when $s_2 = O(q^2)$. However, when $B$ is sparse, the convergence rate in (6) is much faster. For instance, when $s_2 = o(q)$, the rate in (6) is $O_{\mathrm{p}}(q^{1/2})$ faster than that in (7). Similar results hold for the estimate of $B$.

*Remark* 2. The estimation of $(A, B)$ in the nonparanormal bigraphical model achieves the same rate of convergence as that in the Gaussian bigraphical model.

*Remark* 3. Comparing (5) with (6), we find that a factor of order $O_{\mathrm{p}}[\{p\log(pq)/n\}^{1/2}]$ disappears. This is because in (5) we only estimate the inverse correlation matrix, rather than the inverse covariance matrix.

*Remark* 4. When $q$ is fixed, (5) reduces to $\|\hat{A} - A_0\|_{\mathrm{F}} = O_{\mathrm{p}}[\{(p + s_1)\log p/n\}^{1/2}]$. As shown by Lam & Fan (2009), in the Gaussian graphical model the rate of convergence for estimating the inverse correlation matrix is $O_{\mathrm{p}}\{(s_1 \log p/n)^{1/2}\}$. The same rate of convergence was established by Liu et al. (2012) in the nonparanormal graphical model. An extra term of order $O_{\mathrm{p}}\{(p\log p/n)^{1/2}\}$ appears in the bigraphical model, arising from the fact that estimation of $A$ and $B$ is intertwined, as has been shown in the computational algorithm. Even though $A$ is the inverse of a correlation matrix, we must estimate all of its diagonal elements, since they are convolved with the elements of $B$. Estimating the nonparanormal bigraphical model is thus more challenging than estimating the nonparanormal graphical model.

*Remark* 5. As $p, q \to \infty$ and when $A$ and $B$ are sparse, in the sense that $s_1 = o(p)$ and $s_2 = o(q)$, (5) reduces to $\|\hat{A} - A_0\|_{\mathrm{F}} = O_{\mathrm{p}}[\{n^{-1}(s_1 + ps_2/q)\log(pq)\}^{1/2}]$. Hence, the magnitude of $ps_2/q$ characterizes the impact of dimensionality and sparsity of $B$ on the estimation of $A$. Furthermore, if $s_1$ and $s_2$ are finite and $p$ and $q$ are of the same order, then we can allow $p, q \gg n$ without violating the consistency property of $\hat{A}$ and $\hat{B}$, provided that $(\log p + \log q)/n = o(1)$. In this case, the contribution of high dimensionality is merely of a logarithmic factor.

*Remark* 6. The estimation error in the matrix spectral norm, $\|\hat{A} - A_0\|_{\mathrm{s}}$, has the same rate of convergence as $\|\hat{A} - A_0\|_{\mathrm{F}}$, since $\|\hat{A} - A_0\|_{\mathrm{s}} \leqslant \|\hat{A} - A_0\|_{\mathrm{F}}$.

Let $\Omega_0 = B_0 \otimes A_0$ and $\hat{\Omega} = \hat{B} \otimes \hat{A}$ be the true and estimated composite precision matrices. The rate of convergence of $\hat{\Omega}$ to $\Omega_0$ is given in the next corollary, whose proof is deferred to the Supplementary Material.

COROLLARY 2. *Under the conditions in Theorem* 3*, as* $n \to \infty$ *we have*

$$\|\hat{\Omega} - \Omega_0\|_{\mathrm{F}} = O_{\mathrm{p}}\left[\left\{\frac{(s_1s_2 + ps_2 + qs_1)(\log p + \log q)}{n}\right\}^{1/2}\right],$$

$$\|\hat{\Omega} - \Omega_0\|_{\mathrm{s}} = O_{\mathrm{p}}\left[\left\{\frac{(s_1s_2 + ps_2 + qs_1)(\log p + \log q)}{n}\right\}^{1/2}\left(\frac{1}{p} + \frac{1}{q}\right)^{1/2}\right].$$

The following theorem provides the sparsistency result. The proof is similar to that of Theorem 4 in Yin & Li (2012) and is therefore omitted.

THEOREM 4. *Under the conditions in Theorem* 3*, let* $(\hat{A}, \hat{B})$ *be any local minimizer of* $\phi_{\mathrm{p}}(A, B)$ *satisfying the rate of convergence given in Theorem* 3 *and such that* $\|\hat{A} - A_0\|_{\mathrm{s}} = O_{\mathrm{p}}(\eta_A)$ *and* $\|\hat{B} - B_0\|_{\mathrm{s}} = O_{\mathrm{p}}(\eta_B)$ *for some* $\eta_A, \eta_B \to 0$*. Then, with probability tending to* 1*, we have that* $\hat{A}_{ij} = 0$ *for any* $(i, j) \in S_A^{\mathrm{c}}$ *and* $\hat{B}_{ij} = 0$ *for any* $(i, j) \in S_B^{\mathrm{c}}$*, given the following conditions:*

$$q\eta_A + q^{1/2}\left\{1 + \left(\frac{q\log q}{n}\right)^{1/2}\right\}\left(\frac{\log p + \log q}{n}\right)^{1/2}\left(\frac{s_1 s_2 + p s_2 + q s_1}{q}\right)^{1/2} = O(\lambda), \quad (8)$$

$$p\eta_B + p^{1/2}\left\{1 + \left(\frac{p\log p}{n}\right)^{1/2}\right\}\left(\frac{\log p + \log q}{n}\right)^{1/2}\left(\frac{s_1 s_2 + p s_2 + q s_1}{p}\right)^{1/2} = O(\gamma). \quad (9)$$

*Remark* 7. The conditions (8) and (9) give lower bounds for $\lambda$ and $\gamma$. To check whether the lower bounds and upper bounds in Conditions 3 and 4 are compatible, we consider the worst-case scenario, where $\eta_A = \|\hat{A} - A_0\|_{\mathrm{F}}$ and $\eta_B = \|\hat{B} - B_0\|_{\mathrm{F}}$, and the best-case scenario, where $\eta_A = \|\hat{A} - A_0\|_{\mathrm{F}}/p^{1/2}$ and $\eta_B = \|\hat{B} - B_0\|_{\mathrm{F}}/q^{1/2}$. After some algebra, we can show that in the worst-case scenario, we need $s_1 = O(1)$ and $s_2 = O(1)$ to ensure compatibility. Similarly, in the best-case scenario, we need $s_1 = O\{q(1 + q\log q/n)^{-1}\}$ and $s_2 = O\{p(1 + p\log p/n)^{-1}\}$, which reduce to $s_1 = o(n)$ and $s_2 = o(n)$ when $q\log q/n = O(1)$ and $p\log p/n = O(1)$.

## 5. NUMERICAL RESULTS

### 5·1. *Simulation studies*

In simulation studies we adopt the same data-generating procedures as in Liu et al. (2012). To generate the inverse row correlation matrix $A$, we set $A_{jj} = 1$ and $A_{jk} = tb_{jk}$ if $j \neq k$, where $t$ is a constant which guarantees the positive definiteness of $A$ and $b_{jk}$ is a Bernoulli random variable with success probability $p_{jk} = (2\pi)^{-1/2}\exp\{\|z_j - z_k\|_2^2/(2s_1)\}$; here each $z_j = (z_j^{(1)}, z_j^{(2)})$ is independently generated from a bivariate uniform $[0, 1]$ distribution, and $s_1$ determines the sparsity of $A$. Similar procedures can be used to generate the precision matrix $B$, whose sparsity parameter is $s_2$. We rescale $A$ and $B$ such that the diagonal elements of $A^{-1}$ and $B^{-1}$ are 1; see the Supplementary Material for details. We sample $X_1, \ldots, X_n$ from MN$(0; A^{-1}, B^{-1})$, MNPN$(0; A^{-1}, B^{-1}; f)$ and MT$(0; A^{-1}, B^{-1}, e)$, where MT$(0; A^{-1}, B^{-1}, e)$ represents the matrix-$t$ distribution with $e$ degrees of freedom and $A$ and $B$ are row and column precision matrices. The definition of the matrix-$t$ distribution is given in the Supplementary Material. In the matrix-nonparanormal distribution, for any $i \in \{1, \ldots, p\}$ and $j \in \{1, \ldots, q\}$, $f_{ij}(t) = f_{\mathrm{uni}}(t)$ with $f_{\mathrm{uni}}^{-1}(t) = g_0(t)\{\int g_0^2(s)\phi(s)\,\mathrm{d}s\}^{-1/2}$, where $\phi(\cdot)$ is the standard Gaussian density function and $g_0(t) = \mathrm{sign}(t)|t|^{\alpha}$. We take $\alpha = 3$ in MNPN$(0; A^{-1}, B^{-1}; f)$ and $e = 3$ in MT$(0; A^{-1}, B^{-1}, e)$, representing moderate deviations from normal distributions.

The following scenarios with different dimensions, sample sizes and degrees of sparsity are considered. Scenario (i) has $n = 100$, $p = 30$, $q = 30$ and $s_1 = 1, s_2 = 1$. Scenario (ii) has $n = 100$, $p = 100$, $q = 100$ and $s_1 = 2, s_2 = 2$. Scenario (iii) has $n = 50$, $p = 100$, $q = 50$ and $s_1 = 2, s_2 = 1$. Scenario (iv) has $n = 30$, $p = 200$, $q = 200$ and $s_1 = 8, s_2 = 8$. Scenario (i) is an example in which $n$ is larger than $p$ and $q$. In scenario (ii), $p$ and $q$ are comparable to $n$. The numbers of rows and columns are different in scenario (iii). Scenario (iv) has $p, q \gg n$.

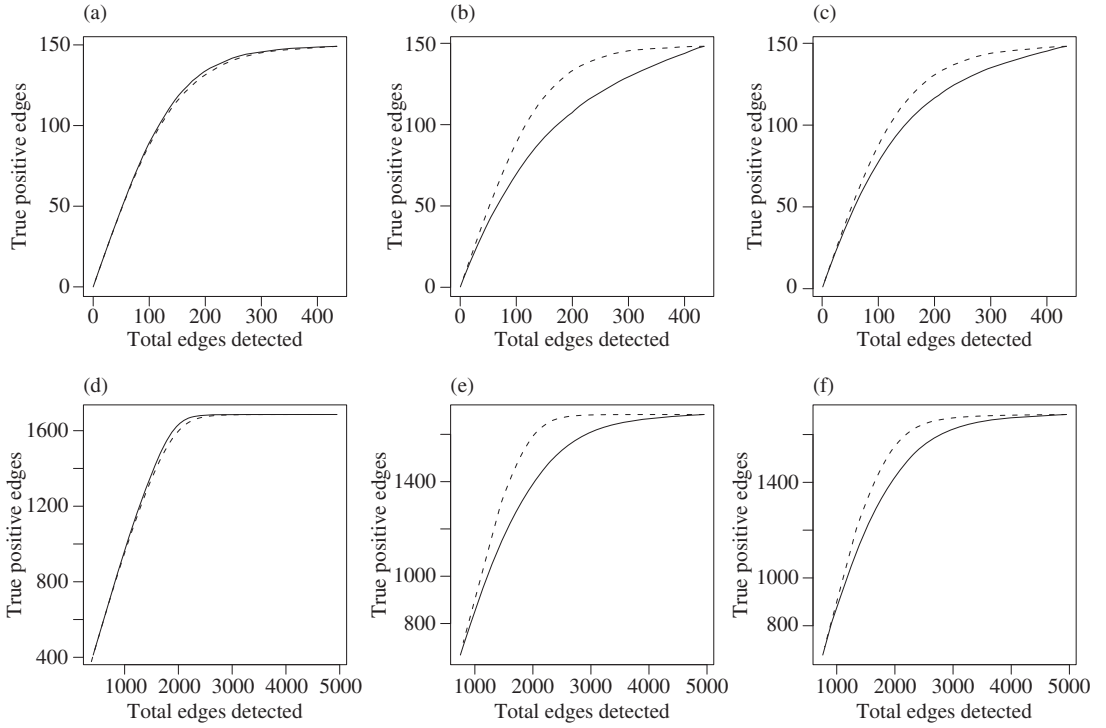


Fig. 1. Plots of the mean number of true positive edges against the mean total number of edges detected for various tuning parameters $\lambda$, based on 100 replicates of simulation scenarios (i) and (ii), using the method of Yin & Li (2012) (solid lines) and our proposed method (dashed lines): (a) matrix-normal data in scenario (i); (b) matrix-nonparanormal data in scenario (i); (c) matrix-$t$ data in scenario (i); (d) matrix-normal data in scenario (ii); (e) matrix-nonparanormal data in scenario (ii); (f) matrix-$t$ data in scenario (ii).

We conducted 100 replicate simulations. To simplify the selection of tuning parameters, we took $\lambda/p = \gamma/q$ (Allen & Tibshirani, 2010). Simulation results based on Kendall's tau and Spearman's rho were almost identical; hence we only present the results based on Kendall's tau.

For each simulated dataset, we applied our proposed method and the $L_1$-penalized Gaussian likelihood method of Yin & Li (2012). To examine the performance of these two methods with respect to graph recovery, we plotted the number of true positive edges against the total number of edges detected for different tuning parameters $\lambda$; here the number of true positive edges refers to the number of lower off-diagonal elements $(i, j)$ such that $A_{ij} \neq 0$ and the estimated $A_{ij}$ is also nonzero, and the total number of edges detected refers to the number of estimated nonzero lower off-diagonal elements. In simulation scenario (i), the mean number of true edges was 149, whereas in scenario (ii) it was 1685.

Figure 1 shows the plot based on 100 replicates in scenarios (i) and (ii). Simulation results for scenarios (iii) and (iv) are given in the Supplementary Material. In scenarios (i) and (ii), the row and column precision matrices are symmetric with the same dimension and sparsity. To save space, we only present the plot for the row precision matrix. For the matrix-normal data, our method performs as well as that of Yin & Li (2012), although the latter method shows a slight advantage in the sense that, given the same total number of edges detected, it identifies more true positive edges than our method. The performance of the method of Yin & Li (2012) gets worse when the data-generating distribution is not Gaussian. For the matrix-nonparanormal data, our method outperforms that of Yin & Li (2012), as expected. While both Gaussian and nonparanormal bigraphical models are misspecified for matrix-$t$ data, our method still performs better than that of Yin & Li (2012). The same conclusions hold in simulation scenarios (iii) and (iv).

Table 1. *Estimation errors of our method and the method of Yin & Li (2012) for* $\Delta_A = \hat{A} - A$ *and* $\Delta_B = \hat{B} - B$*, as measured by the spectral and Frobenius norms, together with associated optimal tuning parameters, based on* 100 *replications. Numbers in parentheses are the simulation standard errors. All values have been multiplied by* 100

| Data distribution | Scenario | Method | $\|\Delta_A\|_{\mathrm{F}}$ | $\|\Delta_B\|_{\mathrm{F}}$ | $\lambda_{\mathrm{F}}^*$ | $\|\Delta_A\|_{\mathrm{s}}$ | $\|\Delta_B\|_{\mathrm{s}}$ | $\lambda_{\mathrm{s}}^*$ |
|---|---|---|---|---|---|---|---|---|
| MN | (i) | PR | 49(2) | 49(2) | 25 | 19(1) | 19(1) | 11 |
| | | BL | 48(2) | 48(2) | 29 | 18(1) | 18(1) | 16 |
| | (ii) | PR | 123(3) | 124(3) | 27 | 26(2) | 26(2) | 12 |
| | | BL | 102(2) | 102(2) | 26 | 24(1) | 24(1) | 11 |
| | (iii) | PR | 201(3) | 99(3) | 47 | 49(2) | 25(1) | 10 |
| | | BL | 182(3) | 85(2) | 44 | 47(2) | 23(1) | 10 |
| | (iv) | PR | 116(3) | 116(3) | 352 | 21(1) | 21(1) | 230 |
| | | BL | 104(3) | 104(3) | 382 | 20(1) | 19(1) | 243 |
| MNPN | (i) | PR | 48(2) | 49(2) | 41 | 18(1) | 18(1) | 20 |
| | | BL | 65(2) | 64(3) | 22 | 30(3) | 29(3) | 18 |
| | (ii) | PR | 124(2) | 123(2) | 44 | 35(2) | 34(2) | 18 |
| | | BL | 190(2) | 189(2) | 30 | 57(2) | 56(2) | 15 |
| | (iii) | PR | 189(3) | 91(2) | 43 | 57(3) | 43(2) | 7 |
| | | BL | 245(4) | 122(3) | 29 | 86(3) | 66(3) | 8 |
| | (iv) | PR | 94(2) | 95(2) | 598 | 23(1) | 23(1) | 410 |
| | | BL | 131(3) | 131(2) | 544 | 35(2) | 36(2) | 370 |
| MT | (i) | PR | 68(5) | 64(4) | 64 | 35(3) | 37(3) | 21 |
| | | BL | 81(6) | 82(6) | 60 | 48(4) | 48(4) | 25 |
| | (ii) | PR | 138(4) | 140(5) | 85 | 61(4) | 63(4) | 24 |
| | | BL | 161(5) | 164(5) | 97 | 80(4) | 83(4) | 31 |
| | (iii) | PR | 195(7) | 119(3) | 165 | 94(4) | 56(4) | 52 |
| | | BL | 239(7) | 132(3) | 184 | 116(4) | 77(4) | 74 |
| | (iv) | PR | 118(3) | 118(3) | 928 | 32(2) | 31(2) | 879 |
| | | BL | 142(3) | 146(3) | 951 | 42(2) | 42(2) | 924 |

MN, matrix-normal distribution; MNPN, matrix-nonparanormal distribution; MT, matrix-*t* distribution; PR, our projected rank-based lasso estimator; BL, the bigraphical lasso estimator of Yin & Li (2012).

Table 1 reports the mean estimation errors of $\hat{A} - A$ and $\hat{B} - B$ in terms of the spectral and Frobenius norms, together with associated optimal tuning parameters. The optimal tuning parameters $\lambda_{\mathrm{F}}^*$ and $\lambda_{\mathrm{s}}^*$ for the estimator $(\hat{A}, \hat{B})$ are defined as $\lambda_{\mathrm{F}}^* = \arg\min_\lambda(\|\hat{A} - A\|_{\mathrm{F}} + \|\hat{B} - B\|_{\mathrm{F}})$ and $\lambda_{\mathrm{s}}^* = \arg\min_\lambda(\|\hat{A} - A\|_{\mathrm{s}} + \|\hat{B} - B\|_{\mathrm{s}})$. For the Gaussian data, the estimation error for the method of Yin & Li (2012) is only 4% to 10% smaller than that for our method. In contrast, the estimation errors for our method are up to 40% smaller than those for the method of Yin & Li (2012) when $X$ follows the matrix-nonparanormal distribution, or up to 25% smaller when $X$ follows the matrix-$t$ distribution. In summary, our method is more robust with respect to the data-generating distribution.

### 5·2. *Genomic data*

In this section, we present the results of applying our method and that of Yin & Li (2012) to the atlas of gene expression in the mouse aging project dataset (Zahn et al., 2007), which contains gene expression values for 8932 genes in 16 tissues. Yin & Li (2012) showed that the gene expression levels in different tissues are correlated. To identify statistically significant genes, we need to take into account gene and tissue dependence structures (Allen & Tibshirani, 2012). In

Table 2. *Total number of edges identified by the graph estimators and the number of different edges between the estimated graphs based on our method and the method of Yin & Li (2012), for* 37 *genes and* 8 *tissues*

| | Gene network | | | | Tissue network | | |
|---|---|---|---|---|---|---|---|
| Number of edges | 20 | 27 | 50 | 100 | 10 | 15 | 20 |
| Number of different edges | 3 | 4 | 10 | 18 | 2 | 3 | 4 |

addition, the correlation structures of genes and tissues are often of interest in their own right. For simplicity, we only focus on a subset of 37 genes belonging to the mouse vascular endothelial growth factor signalling pathway in 8 tissues. The number of replicates is $n = 40$.

Applying the model diagnostic procedure described in the Supplementary Material, we find that the Kronecker correlation assumption is reasonable for this dataset. Furthermore, the quantile-quantile plot in the Supplementary Material shows that many gene expression levels may not be normally distributed. Hence, our method potentially produces more accurate estimates of gene and tissue dependence graphs. To compare our method with the method of Yin & Li (2012), the tuning parameters were selected separately so that the number of edges identified by the two methods are identical. As summarized in Table 2, about 20% of the edges identified by the two methods are different. Given the degrees of sparsity of the graphs, our findings are potentially of biological interest.

For a similar dataset, Yin & Li (2012) identified a gene graph with 27 edges and a tissue graph with 15 edges. As a comparison, we present the graphs with the same numbers of edges in the Supplementary Material. Many important association patterns are revealed by both methods. For instance, it has long been recognized that a group of PLC-$\gamma$ genes in the PKC-dependent pathway is crucial for ERK phosphorylation and proliferation (Holmes et al., 2007). We observe that the dependence of genes *Plcg2*, *Pla2g6* and *Ptk2* in this pathway is recovered by both methods. Similarly, several genes related to the migration of endothelial cells, such as *Mapk13*, *Mapk14* and *Mapkapk2* are also identified in both graphs. In the tissue network, kidney, lung and adrenal glands belonging to the vascular tissue group are connected. Many neural tissues, such as the spinal cord, hippocampus and cerebrum, are correlated as well. As far as the graph differences are concerned, the genes *Mapk3* and *Mapkapk2*, which are likely to be functionally dependent (Christodoulou et al., 2006), are shown to be connected by our method, although not by that of Yin & Li (2012). In addition, it is commonly believed that the function of the thymus is directly associated with the functions of lung and adrenal tissues (Healy et al., 1983); the corresponding correlations are only identified by our method. In summary, the gene and tissue dependence graphs generated by our method seem to be more biologically meaningful than those generated by the method of Yin & Li (2012).

## 6. EXTENSIONS

### 6·1. *Binary bigraphical model*

We consider the following binary bigraphical model: for a $p \times q$ binary matrix-valued random variable $Z$, we assume that there exists an underlying matrix-valued variable $X \sim \mathrm{MNPN}(U, V; f)$ such that $Z_{jk} = I(X_{jk} > C_{jk})$, where $C = \mathrm{vec}\{(C_{jk})\}$ is a $pq \times 1$ vector of constants. Given $n$ independent copies of $Z$, say $Z_1, \ldots, Z_n$, the aim is to infer the conditional independence structure of the latent random variable $X$, which is encoded by the sparsity patterns in the precision matrices $A$ and $B$, where $A = U^{-1}$ and $B = V^{-1}$.

Let $\Delta_{jk} = f_{jk}(C_{jk})$. Assume that the $f(\cdot)$ functions are monotonically increasing and that $Z_{jk} = I\{f_{jk}(X_{jk}) > \Delta_{jk}\}$ for $j = 1, \ldots, p$ and $k = 1, \ldots, q$. Then $E(Z_{jk}) = 1 - \Phi(\Delta_{jk})$, where $\Phi(x)$ is the standard normal cumulative distribution function. Thus $\Delta_{jk}$ can be estimated by $\Phi^{-1}(1 - \bar{Z}_{jk})$, where $\bar{Z}_{jk} = \sum_{i=1}^{n}(Z_i)_{jk}/n$.

As shown in the Supplementary Material, the underlying correlation matrix $\Sigma = V \otimes U$ can be recovered by Kendall's tau. Once $\Sigma$ is estimated by a rank-based estimator, we can similarly project it to the space of positive-definite matrices. The precision matrices $A$ and $B$ can be estimated by minimizing the projected $L_1$-penalized negative loglikelihood of the latent random variables $X_1, \ldots, X_n$. We also consider the case where some of the observed variables are binary and some are continuous; see the Supplementary Material.

### 6·2. *Missing data*

Missing data is an important challenge with matrix-valued variates. Allen & Tibshirani (2010) proposed an EM algorithm for missing-data imputation, applicable when $X$ follows a matrix-normal distribution. In this section, we extend our estimation method to data with missing values.

Compared to the Gaussian bigraphical model, our nonparanormal bigraphical model is more complicated, due to the presence of the nuisance functions $f(\cdot)$. To extend the EM algorithm of Allen & Tibshirani (2010), the $f(\cdot)$ must be estimated. For complete data, the rank-based correlation is invariant under monotonic transformations. This invariance property is not preserved when missing data are imputed from the observed data, so we propose a normal-score-type EM algorithm to estimate $\{f(\cdot), A, B\}$ simultaneously. We first estimate the transformed data $f(X)$ when they are missing, then update the functions $f(\cdot)$ using a normal-score method similar to that of Liu et al. (2009) based on the imputed data, and finally minimize the penalized negative conditional loglikelihood to obtain estimates of $A$ and $B$. The algorithm is iterated until convergence is achieved. Details of the EM algorithm are given in the Supplementary Material. The Bayesian approach of Hoff (2011a) can also be generalized to handle missing data. Given a prior distribution for the correlation matrix, the posterior distribution can be computed using Markov chain Monte Carlo simulation.

## Acknowledgement

This research was supported by the National Institutes of Health and the National Science Foundation. The authors thank Hongzhe Li and Jianxin Yin for making available their processed atlas of gene expression in the mouse aging project dataset; they also thank Nancy Reid, Mark Woodward, the referees, the associate editor and the editor for helpful comments.

## Supplementary material

Supplementary material available at *Biometrika* online includes proofs, further simulation results, figures for the genomic data, more details on the categorical bigraphical model, an EM algorithm for handling missing data, and discussions of the model diagnostic procedure and the nonparanormal bigraphical model with $n = 1$.

## Appendix

### *Proof of Proposition* 2

Since $\hat{R}_{\mathrm{p}}$ is positive definite, we consider the Cholesky decomposition $\hat{R}_{\mathrm{p}} = TT^{\mathrm{T}}$, where $T = (t_1, \ldots, t_{pq})$ is a lower triangular matrix and $t_i$ is a $pq \times 1$ vector. Let $t_i = \mathrm{vec}(W_i)$, where $W_i$ is a $p \times q$

matrix. We have

$$\begin{aligned}\operatorname{tr}\{(B\otimes A)\hat{R}_{\mathrm{p}}\} &= \operatorname{tr}\{T^{\mathrm{T}}(B\otimes A)T\} = \sum_{i=1}^{pq} t_i^{\mathrm{T}}(B\otimes A)t_i \\ &= \sum_{i=1}^{pq}\operatorname{vec}(W_i)^{\mathrm{T}}(B\otimes A)\operatorname{vec}(W_i) = \sum_{i=1}^{pq}\operatorname{tr}(W_i^{\mathrm{T}}AW_iB).\end{aligned}$$

Then, the $(\ell, m)$th element of $\sum_{i=1}^{pq} W_i^{\mathrm{T}}AW_i$ is

$$\begin{aligned}\left(\sum_{i=1}^{pq} W_i^{\mathrm{T}}AW_i\right)_{\ell m} &= \sum_{i=1}^{pq}(W_i)_{*\ell}^{\mathrm{T}}A(W_i)_{*m} = \sum_{i=1}^{pq}\operatorname{vec}(W_i)^{\mathrm{T}}L_\ell^{\mathrm{T}}AL_m\operatorname{vec}(W_i) \\ &= \sum_{i=1}^{pq} t_i^{\mathrm{T}}L_\ell^{\mathrm{T}}AL_mt_i = \operatorname{tr}(T^{\mathrm{T}}L_\ell^{\mathrm{T}}AL_mT) = \operatorname{tr}\left(L_\ell^{\mathrm{T}}AL_m\hat{R}_{\mathrm{p}}\right),\end{aligned}$$

where $L_m$ is as given in Proposition 2. Similarly, the $(\ell, m)$th element of $\sum_{i=1}^{pq} W_iBW_i^{\mathrm{T}}$ is

$$\begin{aligned}\left(\sum_{i=1}^{pq} W_iBW_i^{\mathrm{T}}\right)_{\ell m} &= \sum_{i=1}^{pq}(W_i)_{\ell *}B(W_i)_{m*} = \sum_{i=1}^{pq}\operatorname{vec}(W_i)^{\mathrm{T}}K_\ell BK_m^{\mathrm{T}}\operatorname{vec}(W_i) \\ &= \sum_{i=1}^{pq} t_i^{\mathrm{T}}K_\ell BK_m^{\mathrm{T}}t_i = \operatorname{tr}(T^{\mathrm{T}}K_\ell BK_m^{\mathrm{T}}T) = \operatorname{tr}\left(K_\ell BK_m^{\mathrm{T}}\hat{R}_{\mathrm{p}}\right),\end{aligned}$$

where $K_m$ is as given in Proposition 2. The proof is complete.

*Proof of Theorem* 3

The main idea of the proof follows from Yin & Li (2012), Lam & Fan (2009) and Rothman et al. (2008). Let $\Delta_1 = \alpha_n U_1$, where $U_1$ is a symmetric matrix of size $p$. Let $\Delta_2 = \beta_n U_2$, where $U_2$ is a symmetric matrix of size $q$. Let

$$\alpha_n = \left\{\frac{\log(pq)}{n}\frac{(s_1s_2 + ps_2 + qs_1)}{q}\right\}^{1/2}, \quad \beta_n = \left\{\frac{\log(pq)}{n}\frac{(s_1s_2 + ps_2 + qs_1)}{p}\right\}^{1/2}.$$

The aim is to show that there exists a local minimizer $(\hat{A}, \hat{B})$ of $\phi_{\mathrm{p}}(A, B)$ in

$$\mathcal{A} = \{(A_0 + \Delta_1, B_0 + \Delta_2) : \|\Delta_1\|_{\mathrm{F}} < C_1\alpha_n,\ \|\Delta_2\|_{\mathrm{F}} < C_2\beta_n\},$$

where $C_1$ and $C_2$ are large enough constants. Hence, $\|\hat{A} - A_0\|_{\mathrm{F}} = O_{\mathrm{p}}(\alpha_n)$ and $\|\hat{B} - B_0\|_{\mathrm{F}} = O_{\mathrm{p}}(\beta_n)$. Let $\partial\mathcal{A} = \{(A_0 + \Delta_1, B_0 + \Delta_2) : \|\Delta_1\|_{\mathrm{F}} = C_1\alpha_n,\ \|\Delta_2\|_{\mathrm{F}} = C_2\beta_n\}$. It suffices to show that

$$\operatorname{pr}\left\{\inf_{(A_0+\Delta_1, B_0+\Delta_2)\in\partial\mathcal{A}} \phi_{\mathrm{p}}(A_0 + \Delta_1, B_0 + \Delta_2) > \phi_{\mathrm{p}}(A_0, B_0)\right\} \to 1 \tag{A1}$$

for sufficiently large constants $C_1$ and $C_2$. Let $A_1 = A_0 + \Delta_1 = (A_{ij}^{(1)})$, $B_1 = B_0 + \Delta_2 = (B_{ij}^{(1)})$ and $\Omega_1 = B_1 \otimes A_1$. Then

$$\begin{aligned}\phi_{\mathrm{p}}(A_1, B_1) - \phi_{\mathrm{p}}(A_0, B_0) &= -q(\log|A_1| - \log|A_0|) - \operatorname{pr}(\log|B_1| - \log|B_0|) + \operatorname{tr}\{\hat{R}(\Omega_1 - \Omega_0)\} \\ &\quad + \lambda\sum_{i\neq j}(|A_{ij}^{(1)}| - |A_{ij}^{(0)}|) + \gamma\sum_{i\neq j}(|B_{ij}^{(1)}| - |B_{ij}^{(0)}|) \\ &= I_1 + I_2 + I_3 + I_4 + I_5.\end{aligned}$$

Let $I_4 = \lambda\sum_{(i,j)\in S_A^{\mathrm{c}}}|A_{ij}^{(1)}| + \lambda\sum_{i\neq j,(i,j)\in S_A}(|A_{ij}^{(1)}| - |A_{ij}^{(0)}|) = I_{41} + I_{42}$. Using Taylor's expansion,

$$I_1 = -q\operatorname{tr}(A_0^{-1}\Delta_1) + q\operatorname{vec}(\Delta_1)^{\mathrm{T}}\left\{\int_0^1 (1-v)g(v, A_v)\,\mathrm{d}v\right\}\operatorname{vec}(\Delta_1), \tag{A2}$$

where $g(v, A_v) = A_v^{-1} \otimes A_v^{-1}$. A similar expansion holds for $I_2$. The bilinearity of the Kronecker product yields

$$I_3 = \mathrm{tr}\{(\hat{R}_\mathrm{p} - \Sigma_0)(\Delta_2 \otimes A_0)\} + \mathrm{tr}\{(\hat{R}_\mathrm{p} - \Sigma_0)(B_0 \otimes \Delta_1)\} + \mathrm{tr}\{(\hat{R}_\mathrm{p} - \Sigma_0)(\Delta_2 \otimes \Delta_1)\} + \mathrm{tr}\{\Sigma_0(\Delta_2 \otimes A_0)\} + \mathrm{tr}\{\Sigma_0(B_0 \otimes \Delta_1)\} + \mathrm{tr}\{\Sigma_0(\Delta_2 \otimes \Delta_1)\}. \quad \text{(A3)}$$

Following from the facts that $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ and $\mathrm{tr}(A \otimes B) = \mathrm{tr}(A)\mathrm{tr}(B)$, we have

$$\mathrm{tr}\{\Sigma_0(\Delta_2 \otimes A_0)\} = \mathrm{tr}\{(B_0^{-1}\Delta_2) \otimes I\} = p\,\mathrm{tr}(B_0^{-1}\Delta_2). \quad \text{(A4)}$$

Combining equations (A2), (A3) and (A4), we have

$$\begin{aligned} I_1 + I_2 + I_3 &= q\,\mathrm{vec}(\Delta_1)^\mathrm{T} \left\{ \int_0^1 (1-v) g(v, A_v)\,\mathrm{d}v \right\} \mathrm{vec}(\Delta_1) \\ &\quad + p\,\mathrm{vec}(\Delta_2)^\mathrm{T} \left\{ \int_0^1 (1-v) g(v, B_v)\,\mathrm{d}v \right\} \mathrm{vec}(\Delta_2) + \mathrm{tr}\{(\hat{R}_\mathrm{p} - \Sigma_0)(\Delta_2 \otimes A_0)\} \\ &\quad + \mathrm{tr}\{(\hat{R}_\mathrm{p} - \Sigma_0)(B_0 \otimes \Delta_1)\} + \mathrm{tr}\{(\hat{R}_\mathrm{p} - \Sigma_0)(\Delta_2 \otimes \Delta_1)\} + \mathrm{tr}(A_0^{-1}\Delta_1)\mathrm{tr}(B_0^{-1}\Delta_2) \\ &= K_1 + K_2 + K_3 + K_4 + K_5 + K_6. \end{aligned}$$

Arguments similar to those in Lam & Fan (2009) yield $K_1 \geqslant 2^{-1} q C_1^2 \alpha_n^2 \{\delta_2 + o(1)\}^{-2}$ and $K_2 \geqslant 2^{-1} p C_2^2 \beta_n^2 \{\delta_4 + o(1)\}^{-2}$. Next, we will show that $|K_4|$ is dominated by $K_1 + I_{41}$, i.e., that $|K_4| < K_1 + I_{41}$ for sufficiently large $n$ and $C_1$. Then

$$\begin{aligned} |K_4| &\leqslant \sum_{i \neq k} |(\hat{R}_\mathrm{p} - \Sigma_0)_{ik}(B_0 \otimes \Delta_1)_{ik}| \\ &\leqslant \|\hat{R}_\mathrm{p} - \Sigma_0\|_\mathrm{max} \sum_{i \neq j} |(\Delta_1)_{ij}| \sum_{k \neq \ell} |(B_0)_{k\ell}| + \|\hat{R}_\mathrm{p} - \Sigma_0\|_\mathrm{max} \sum_{i=j} |(\Delta_1)_{ij}| \sum_{k \neq \ell} |(B_0)_{k\ell}| \\ &\quad + \|\hat{R}_\mathrm{p} - \Sigma_0\|_\mathrm{max} \sum_{i \neq j} |(\Delta_1)_{ij}| \sum_{k=\ell} |(B_0)_{k\ell}| = K_{41} + K_{42} + K_{43}. \end{aligned}$$

Let us consider $K_{41}$, $K_{42}$ and $K_{43}$ separately. By Theorem 2, $\|\hat{R}_\mathrm{p} - \Sigma_0\|_\mathrm{max} = O_\mathrm{p}[\{\log(pq)/n\}^{1/2}]$. For $K_{41}$, we have

$$K_{41} \leqslant O_\mathrm{p}\left[ \left\{ \frac{\log(pq)}{n} \right\}^{1/2} \right] \left\{ (s_1 s_2 q)^{1/2} \|B_0\|_\mathrm{s} C_1 \alpha_n + (s_2 q)^{1/2} \|B_0\|_\mathrm{s} \sum_{i \neq j, (i,j) \in S_A^\mathrm{c}} |A_{ij}^{(1)}| \right\}.$$

Since $\|B_0\| \leqslant \delta_4$ by Condition 2, one can show that $|K_{41}|$ is dominated by $K_1 + I_{41}$. Following similar steps, $K_{43}$ is also shown to be dominated by $K_1 + I_{41}$. By the Cauchy–Schwartz inequality, we have

$$K_{42} \leqslant O_\mathrm{p}\left[ \left\{ \frac{\log(pq)}{n} \right\}^{1/2} \right] p^{1/2} \|\Delta_1\|_\mathrm{F} (s_2 q)^{1/2} \|B_0\|_\mathrm{s} \leqslant O_\mathrm{p}(q\alpha_n^2).$$

Combining the upper bounds for $K_{41}$, $K_{42}$ and $K_{43}$, we know that $|K_4|$ is dominated by $K_1 + I_{41}$. As shown in the Supplementary Material, $|K_3|$ and $|K_5|$ are also controlled, and $|K_6|$ is bounded above by $K_1 + K_2$. Furthermore, $|I_{42}| \leqslant \lambda \sum_{i \neq j, (i,j) \in S_A} |A_{ij}^{(1)} - A_{ij}^{(0)}| \leqslant \lambda s_1^{1/2} \|\Delta_1\|_\mathrm{F} \leqslant O_\mathrm{p}(q\alpha_n^2)$, where the last step follows from Condition 3. Then $I_{42}$ is dominated by $K_1 > 0$. We have shown that $\phi_\mathrm{p}(A_1, B_1) - \phi_\mathrm{p}(A_0, B_0)$ is bounded from below by a positive constant independent of $\Delta_1$ and $\Delta_2$. Therefore (A1) holds, which completes the proof.

## References


Allen, G. I. & Tibshirani, R. J. (2010). Transposable regularized covariance models with an application to missing data imputation. *Ann. Appl. Statist.* **42**, 764–90.

Allen, G. I. & Tibshirani, R. J. (2012). Inference with transposable data: Modelling the effects of row and column correlations. *J. R. Statist. Soc.* B **74**, 721–43.

CAI, T., LIU, W. & LUO, X. (2011). A constrained $\ell_1$ minimization approach to sparse precision matrix estimation. *J. Am. Statist. Assoc.* **106**, 594–607.

CHRISTODOULOU, I., BUTTERY, L. D., TAI, G., HENCH, L. L. & POLAK, J. M. (2006). Characterization of human fetal osteoblasts by microarray analysis following stimulation with 58s bioactive gel-glass ionic dissolution products. *J. Biomed. Materials Res.* **77**, 431–46.

DAWID, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68**, 265–74.

DUTILLEUL, P. (1999). The MLE algorithm for the matrix normal distribution. *J. Statist. Comp. Simul.* **64**, 105–23.

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

GENTON, M. G. (2007). Spatial-temporal analysis of multivariate environmental monitoring data. *Environmetrics* **18**, 681–95.

GUPTA, A. & NAGAR, D. (1999). *Matrix Variate Distributions*. Boca Raton, FL: Chapman & Hall.

HEALY, D. L., HODGEN, G. D., SCHULTE, H. M., CHROUSOS, D. L., LORIAUX, D. L., HALL, N. R. & GOLDSTEIN, A. L. (1983). The thymus-adrenal connection: Thymosin has corticotropin-releasing activity in primates. *Science* **222**, 1353–5.

HOFF, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Statist.* **1**, 265–83.

HOFF, P. D. (2011a). Hierarchical multilinear models for multiway data. *Comp. Statist. Data Anal.* **55**, 530–43.

HOFF, P. D. (2011b). Separable covariance arrays via the Tucker product, with applications to multivariate relational data. *Bayesian Anal.* **6**, 179–96.

HOLMES, K., ROBERTS, O. L., THOMAS, A. M. & CROSS, M. J. (2007). Vascular endothelial growth factor receptor-2: Structure, function, intracellular signalling and therapeutic inhibition. *Cell Signal.* **19**, 2003–12.

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

LAURITZEN, S. L. (1996). *Graphical Models*. Oxford: Clarendon Press.

LIU, H., HAN, F., YUAN, M., LAFFERTY, J. D. & WASSERMAN, L. A. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40**, 2293–326.

LIU, H., LAFFERTY, J. D. & WASSERMAN, L. A. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10**, 2295–328.

LU, N. & ZIMMERMAN, D. L. (2005). The likelihood ratio test for a separable covariance matrix. *Statist. Prob. Lett.* **73**, 449–57.

MARDIA, K. V. & GOODALL, C. (1993). Spatial-temporal analysis of multivariate environmental monitoring data. *Environmetrics* **6**, 347–85.

MITCHELL, M. W., GENTON, M. G. & GUMPERTZ, M. L. (2005). Testing for separability of space-time covariances. *Environmetrics* **64**, 819–31.

MITCHELL, M. W., GENTON, M. G. & GUMPERTZ, M. L. (2006). A likelihood ratio test for separability of covariances. *J. Mult. Anal.* **97**, 1025–43.

NAIK, D. N. & RAO, S. S. (2001). Analysis of multivariate repeated measures data with a Kronecker product structured covariance matrix. *J. Appl. Statist.* **29**, 91–105.

NESTEROV, Y. (2005). Smooth minimization of non-smooth functions. *Math. Program.* **103**, 127–52.

ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. & ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Statist.* **2**, 494–515.

TENG, S. & HUANG, H. (2009). A statistical framework to infer functional gene relationships from biologically interrelated microarray experiments. *J. Am. Statist. Assoc.* **104**, 465–73.

THEOBALD, D. L. & WUTTKE, D. S. (2006). Empirical Bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian procrustes problem. *Proc. Nat. Acad. Sci.* **103**, 18521–7.

WANG, H. & WEST, M. (2009). Bayesian analysis of matrix normal graphical models. *Biometrika* **96**, 821–34.

WERNER, K., JANSSON, M. & STOICA, P. (2008). On estimation of covariance matrices with Kronecker product structure. *IEEE Trans. Sig. Proces.* **56**, 478–91.

WITTEN, D. M., FRIEDMAN, J. H. & SIMON, N. (2011). New insights and faster computations for the graphical lasso. *J. Comp. Graph. Statist.* **20**, 892–900.

YIN, J. & LI, H. (2012). Model selection and estimation in the matrix normal graphical model. *J. Mult. Anal.* **107**, 119–40.

YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11**, 2261–86.

ZAHN, J. M., POOSALA, S., OWEN, A. B., INGRAM, D. K., LUSTIG, A., CARTER, A., WEERARATNA, A. T., TAUB, D. D., GOROSPE, M., MAZAN-MAMCZARZ, K., LAKATTA, E. G., BOHELER, K. R., XU, X., MATTSON, M. P., FALCO, G., KO, M. S. H., SCHLESSINGER, D., FIRMAN, J., KUMMERFELD, S. K., WOOD, W. H. III, et al. (2007). AGEMAP: a gene expression database for aging in mice. *PLoS Genetics* **3**, 2326–37.

ZHAO, T., LIU, H., ROEDER, K. E., LAFFERTY, J. D. & WASSERMAN, L. A. (2012). The huge package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.* **13**, 1059–62.

[*Received June* 2012. *Revised January* 2013]

*Biometrika* (0000), **00**, 00, *pp*. 1–16

Advance Access publication on 15 December 2012
*Printed in Great Britain*

# Supplementary Material to: High Dimensional Semiparametric Bigraphical Models

BY YANG NING

*Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada.*

yning@jhsph.edu

HAN LIU

*Department of Operation Research and Financial Engineering, Princeton University, New Jersey 08544, U.S.A.*

hanliu@princeton.edu

SUMMARY

This supplementary file includes the proofs, further simulation results, figures for the genomic data, more details on the categorical bigraphical model, an EM algorithm for handling missing data, and discussions of the model diagnostic procedure and the nonparanormal bigraphical model with $n = 1$.

## PROOF OF THEOREM 1

To show Theorem 1, we begin with the following lemmas.

LEMMA 1. *If* $(\hat{A}^{(k)}, \hat{B}^{(k)}) \neq (\hat{A}^{(k+1)}, \hat{B}^{(k+1)})$, *then* $\phi_p(\hat{A}^{(k+1)}, \hat{B}^{(k+1)}) < \phi_p(\hat{A}^{(k)}, \hat{B}^{(k)})$.

*Proof of Lemma 1.* Note that $-\log|A|$ is a convex function of $A$. The quantity $\mathrm{tr}\{(B \otimes A)\hat{R}_p\}$ is a linear function of $A$ by fixing $B$, and the penalty term $\sum_{i \neq j} |A_{ij}|$ is also convex by the triangle inequality. Then $\phi_p(A, B)$ is strictly convex as a function of $A$, assuming $B$ is fixed, and similarly it is also strictly convex as a function of $B$, assuming $A$ is fixed. Hence, the minimization in steps (iii) and (iv) of the algorithm yields non-increasing values of $\phi_p(A, B)$, i.e.,

$$\phi_p(\hat{A}^{(k+1)}, \hat{B}^{(k+1)}) \leq \phi_p(\hat{A}^{(k+1)}, \hat{B}^{(k)}) \leq \phi_p(\hat{A}^{(k)}, \hat{B}^{(k)}).$$

If $(\hat{A}^{(k)}, \hat{B}^{(k)}) \neq (\hat{A}^{(k+1)}, \hat{B}^{(k+1)})$, at least one of above inequalities is strict. Then $\phi_p(\hat{A}^{(k+1)}, \hat{B}^{(k+1)}) < \phi_p(\hat{A}^{(k)}, \hat{B}^{(k)})$.

LEMMA 2. *There exist compact convex sets* $\mathcal{A} \subseteq \mathbb{R}^{p \times p}$ *and* $\mathcal{B} \subseteq \mathbb{R}^{q \times q}$, *such that the sequence* $(\hat{A}^{(k)}, \hat{B}^{(k)})$ *is contained in* $(\mathcal{A}, \mathcal{B})$.

*Proof of Lemma 2.* We only need to show that $\|\hat{A}^{(k)}\|_F$ and $\|\hat{B}^{(k)}\|_F$ are bounded, say by a constant $K$. Then we can take $\mathcal{A} = [-K, K]^{p \times p} \cap \mathcal{P}_p$ and $\mathcal{B} = [-K, K]^{q \times q} \cap \mathcal{P}_q$, where $\mathcal{P}_d$ is the space of $d \times d$ positive definite matrices. If the statement is false, then there exists a subsequence indexed by $k_i$ such that $\|\hat{A}^{(k_i)}\|_F + \|\hat{B}^{(k_i)}\|_F$ diverges to infinity. For notational simplicity, we assume $\|\hat{A}^{(k)}\|_F + \|\hat{B}^{(k)}\|_F \to \infty$. By the arguments as in Lemma 1 and as $k \to \infty$, we

get

$$\lim_{k\to\infty} \phi_p(\hat{A}^{(k)}, \hat{B}^{(k)}) \leq \phi_p(\hat{A}^{(2)}, \hat{B}^{(1)}). \tag{1}$$

Next, we will show that, $\phi_p(\hat{A}^{(k)}, \hat{B}^{(k)})$ is unbounded, i.e., $\phi_p(\hat{A}^{(k)}, \hat{B}^{(k)}) \to +\infty$, as $k \to \infty$. Since $\hat{B}^{(k)} \otimes \hat{A}^{(k)}$ and $\hat{R}_p$ are both positive definite, we can consider their Cholesky decompositions. Write $\hat{B}^{(k)} \otimes \hat{A}^{(k)} = T^{(k)}T^{(k)T}$ and $\hat{R}_p = S^{(k)}S^{(k)T}$, where $S^{(k)} = (s_{ij}^{(k)})$ and $T^{(k)} = (t_{ij}^{(k)})$ are lower triangular matrices. Note that if all the diagonal elements of $\hat{A}^{(k)}$ can become arbitrarily small, i.e., $\|\text{diag}(\hat{A}^{(k)})\|_F \to 0$ as $k \to \infty$, all the eigenvalues of $\hat{A}^{(k)}$ approach 0. It implies $-q \log |\hat{A}^{(k)}| \to +\infty$, and then $\phi_p(\hat{A}^{(k)}, \hat{B}^{(k)}) \to +\infty$. Similarly, if all the diagonal elements of $\hat{B}^{(k)}$ are arbitrarily small, $-p \log |\hat{B}^{(k)}| \to +\infty$, and then $\phi_p(\hat{A}^{(k)}, \hat{B}^{(k)}) \to +\infty$. The remaining situation is that at least one of the diagonal elements of $\hat{B}^{(k)}$ and at least one of the diagonal elements of $\hat{A}^{(k)}$ are bounded from below by a positive constant. It implies that $\|A^{(k)}\|_F$ and $\|B^{(k)}\|_F$ are both bounded from below. Note that $\|T^{(k)}T^{(k)T}\|_F = \|A^{(k)}\|_F \|B^{(k)}\|_F$. Then $\|A^{(k)}\|_F + \|B^{(k)}\|_F \to \infty$, implies $\|T^{(k)}T^{(k)T}\|_F \to \infty$, and then there exists at least one $t_{im}^{(k)}$ diverging to infinity.

The function $\phi_p(A^{(k)}, B^{(k)})$ can be writen as a function of $S^{(k)}$ and $T^{(k)}$ as

$$\begin{aligned}
\phi_p(A^{(k)}, B^{(k)}) &= -2\sum_{j=1}^{pq} \log t_{jj}^{(k)} + \sum_{i,j=1}^{pq} \left( \sum_{m=1}^{pq} t_{im}^{(k)} s_{mj}^{(k)} \right)^2 + \lambda \sum_{i\neq j} |A_{ij}^{(k)}| + \gamma \sum_{i\neq j} |B_{ij}^{(k)}| \\
&\geq -2\sum_{j=1}^{pq} \log t_{jj}^{(k)} + \sum_{i,j=1}^{pq} \left( \sum_{m=1}^{pq} t_{im}^{(k)} s_{mj}^{(k)} \right)^2 .
\end{aligned}$$

Denote

$$Q(T^{(k)}) = -2\sum_{j=1}^{pq} \log t_{jj}^{(k)} + \sum_{i,j=1}^{pq} \left( \sum_{m=1}^{pq} t_{im}^{(k)} s_{mj}^{(k)} \right)^2 .$$

If the off-diagonal element $t_{im}^{(k)}$ goes to infinity, $Q(T^{(k)})$ is dominated by $(t_{im}^{(k)})^2 \{\sum_{j=1}^{pq} (s_{mj}^{(k)})^2\}$. Since $s_{mm}^{(k)} > 0$, $\phi_p(A^{(k)}, B^{(k)}) \to +\infty$, as $k \to \infty$. Likewise, if the diagonal element $t_{jj}^{(k)}$ goes to infinity, $Q(T^{(k)})$ is dominated by $(t_{jj}^{(k)})^2 (s_{jj}^{(k)})^2$. Again, $\phi_p(A^{(k)}, B^{(k)}) \to +\infty$.

By (1), we then deduce that $\phi_p(\hat{A}^{(2)}, \hat{B}^{(1)}) = +\infty$. However, according to the definition of $\hat{A}^{(2)}$, we know $\phi_p(\hat{A}^{(2)}, \hat{B}^{(1)}) \leq \phi_p(I_p, I_q)$. Then, $\phi_p(\hat{A}^{(2)}, \hat{B}^{(1)})$ must be finite, which yields a contradiction.

LEMMA 3. *The sequence $\phi_p(\hat{A}^{(k)}, \hat{B}^{(k)})$ converges monotonically, and $(\hat{A}^{(k)}, \hat{B}^{(k)})$ has at least one accumulation point. For any accumulation point $(A^*, B^*)$, it is a stationary point of $\phi_p(A, B)$.*

*Proof of Lemma 3.* Since $\mathcal{A}$ and $\mathcal{B}$ in Lemma 2 are compact and $\phi_p(A, B)$ is continuous, $\phi_p(A, B)$ has a finite minimum and therefore is bounded from below. Together with Lemma 1, the sequence $\phi_p(\hat{A}^{(k)}, \hat{B}^{(k)})$ converges monotonically to a limit value. Also from the compactness of $\mathcal{A}$ and $\mathcal{B}$, $(\hat{A}^{(k)}, \hat{B}^{(k)})$ has at least one accumulation point. Next, we will show that

$$\phi_p(A^*, B^*) \leq \phi_p(A, B^*), \;\; \phi_p(A^*, B^*) \leq \phi_p(A^*, B), \tag{2}$$

for any $A \in \mathcal{A}$ and $B \in \mathcal{B}$. We know that there exists a convergent subsequence $(\hat{A}^{(k_i)}, \hat{B}^{(k_i)})$ with a limit point $(A^*, B^*)$. Since $\phi_p(A, B)$ is continuous, we get

$$\phi_p(A^*, B^*) = \lim_{k_i \to \infty} \phi_p(\hat{A}^{(k_i+1)}, \hat{B}^{(k_i)}) \leq \lim_{k_i \to \infty} \phi_p(A, \hat{B}^{(k_i)}) = \phi_p(A, B^*),$$

for any $A \in \mathcal{A}$. Similarly, $\phi_p(A^*, B^*) \leq \phi_p(A^*, B)$, for any $B \in \mathcal{B}$. Since (2) is true, the partial derivatives of $\phi_p(A, B)$ at $(A^*, B^*)$ are 0. Then $(A^*, B^*)$ is a stationary point of $\phi_p(A, B)$.

*Proof of Theorem 1.* We use the method of proof by contradiction. Assume that there exist $\delta > 0$ and infinitely many $k_i$, such that $\|\hat{A}^{(k_i+1)} - \hat{A}^{(k_i)}\|_F + \|\hat{B}^{(k_i+1)} - \hat{B}^{(k_i)}\|_F > \delta$. Since $\mathcal{A}$ is compact, we can select a convergent subsequence in the sequence $(\hat{A}^{(k_i+1)}, \hat{A}^{(k_i)})$. Among the selected subsequence, we can select a further convergent subsequence in $(\hat{B}^{(k_i+1)}, \hat{B}^{(k_i)})$. For notational simplicity, we assume that the sequences $(\hat{A}^{(k_i)}, \hat{B}^{(k_i)})$ and $(\hat{A}^{(k_i+1)}, \hat{B}^{(k_i+1)})$ converge to $(A_1, B_1)$ and $(A_2, B_2)$ with $\|A_1 - A_2\|_F + \|B_1 - B_2\|_F \geq \delta$. Since $(\hat{A}^{(k_i)}, \hat{B}^{(k_i)}) \neq (\hat{A}^{(k_i+1)}, \hat{B}^{(k_i+1)})$, by Lemma 1 we have $\phi_p(\hat{A}^{(k_i+1)}, \hat{B}^{(k_i+1)}) < \phi_p(\hat{A}^{(k_i)}, \hat{B}^{(k_i)})$. Then

$$\phi_p(A_1, B_1) = \lim_{k_i \to \infty} \phi_p(\hat{A}^{(k_i)}, \hat{B}^{(k_i)}) \geq \lim_{k_i \to \infty} \phi_p(\hat{A}^{(k_i+1)}, \hat{B}^{(k_i+1)}) = \phi_p(A_2, B_2).$$

Since $k_{i-2} + 1 < k_i$, likewise, we have

$$\phi_p(A_1, B_1) = \lim_{k_i \to \infty} \phi_p(\hat{A}^{(k_i)}, \hat{B}^{(k_i)}) \leq \lim_{k_i \to \infty} \phi_p(\hat{A}^{(k_{i-2}+1)}, \hat{B}^{(k_{i-2}+1)}) = \phi_p(A_2, B_2).$$

Then we derive $\phi_p(A_1, B_1) = \phi_p(A_2, B_2)$. According to the definition, $\phi_p(\hat{A}^{(k_i+1)}, \hat{B}^{(k_i+1)}) \leq \phi_p(\hat{A}^{(k_i+1)}, B)$ for any $B \in \mathcal{B}$, and $\phi_p(\hat{A}^{(k_i+1)}, \hat{B}^{(k_i)}) \leq \phi_p(A, \hat{B}^{(k_i)})$ for any $A \in \mathcal{A}$. As $k_i \to \infty$,

$$\phi_p(A_2, B_2) \leq \phi_p(A_2, B), \;\; \phi_p(A_2, B_1) \leq \phi_p(A, B_1), \tag{3}$$

for any $B \in \mathcal{B}$, and $A \in \mathcal{A}$. Since $\phi_p(A_2, B)$ is strictly convex as a function of $B$ for fixed $A$ and as a function of $A$ for fixed $B$, (3) becomes an equality only if $B = B_2$ and $A = A_1$. Thus $\phi_p(A_1, B_1) = \phi_p(A_2, B_2)$ implies $A_1 = A_2$ and $B_1 = B_2$, which yields a contradiction. By Lemma 3, the accumulation point of $(\hat{A}^{(k)}, \hat{B}^{(k)})$ is a stationary point of $\phi_p(A, B)$. The proof is complete.

## PROOF OF THEOREM 2

Liu et al. (2012) showed the following concentration inequality of the rank-based estimator $\hat{R}$.

LEMMA 4. *Given the rank-based estimator $\hat{R}$, for n large enough and $t > 0$*

$$\text{pr}\left(\|\hat{R} - \Sigma_0\|_{\max} \leq 8\pi t\right) \geq 1 - p^2 q^2 \exp\left(-nt^2\right).$$

*Proof of Theorem 2*. According to the definition of $\hat{R}_p$,

$$\|\hat{R}_p - \Sigma_0\|_{\max} \leq \|\hat{R}_p - \hat{R}\|_{\max} + \|\hat{R} - \Sigma_0\|_{\max} \leq 2\|\hat{R} - \Sigma_0\|_{\max}. \tag{4}$$

By Lemma 4, we have

$$\text{pr}\left(\|\hat{R}_p - \Sigma_0\|_{\max} \leq 16\pi t\right) \geq \text{pr}\left(\|\hat{R} - \Sigma_0\|_{\max} \leq 8\pi t\right),$$

which completes the proof.

## PROOF OF THEOREM 3

*Proof of Theorem 3*. The main idea of the proof follows from Yin & Li (2012); Lam & Fan (2009); Rothman et al. (2008). Let $\Delta_1 = \alpha_n U_1$, where $U_1$ is a symmetric matrix of size $p$. Let $\Delta_2 = \beta_n U_2$, where $U_2$ is a symmetric matrix of size $q$. Let

$$\alpha_n = \left\{ \frac{\log(pq)}{n} \frac{(s_1 s_2 + p s_2 + q s_1)}{q} \right\}^{1/2}, \quad \beta_n = \left\{ \frac{\log(pq)}{n} \frac{(s_1 s_2 + p s_2 + q s_1)}{p} \right\}^{1/2}.$$

The aim is to show that there exists a local minimizer $(\hat{A}, \hat{B})$ of $\phi_p(A, B)$ in

$$\mathcal{A} := \{(A_0 + \Delta_1, B_0 + \Delta_2) : \|\Delta_1\|_F < C_1 \alpha_n, \|\Delta_2\|_F < C_2 \beta_n\}, \tag{5}$$

where $C_1$ and $C_2$ are large enough constants. Hence, $\|\hat{A} - A_0\|_F = O_p(\alpha_n)$ and $\|\hat{B} - B_0\|_F = O_p(\beta_n)$. Let

$$\partial\mathcal{A} = \{(A_0 + \Delta_1, B_0 + \Delta_2) : \|\Delta_1\|_F = C_1 \alpha_n, \|\Delta_2\|_F = C_2 \beta_n\}.$$

It suffices to show that

$$\text{pr}\left\{ \inf_{(A_0+\Delta_1, B_0+\Delta_2) \in \partial\mathcal{A}} \phi_p(A_0 + \Delta_1, B_0 + \Delta_2) > \phi_p(A_0, B_0) \right\} \to 1, \tag{6}$$

for sufficiently large constants $C_1$ and $C_2$. The reason is as follows. The function $\phi_p(A, B)$ is continuous and therefore attains the minimum in the closure of $\mathcal{A}$. Since (6) implies that $\phi_p(A_0, B_0)$ is smaller than the infimum of $\phi_p(A, B)$ over $\partial\mathcal{A}$, the infimum is attained in the interior of $\mathcal{A}$. Then there exists at least one local minimizer of $\phi_p(A, B)$ in the region $\mathcal{A}$.

Let $A_1 = A_0 + \Delta_1 = (A_{ij}^{(1)})$, $B_1 = B_0 + \Delta_2 = (B_{ij}^{(1)})$, and $\Omega_1 = B_1 \otimes A_1$. Then,

$$\begin{aligned}
\phi_p(A_1, B_1) - \phi_p(A_0, B_0) &= \underbrace{-q(\log|A_1| - \log|A_0|)}_{I_1} \underbrace{-p(\log|B_1| - \log|B_0|)}_{I_2} + \underbrace{\text{tr}\{\hat{R}(\Omega_1 - \Omega_0)\}}_{I_3} \\
&\quad \underbrace{\lambda \sum_{i \neq j} (|A_{ij}^{(1)}| - |A_{ij}^{(0)}|)}_{I_4} + \underbrace{\gamma \sum_{i \neq j} (|B_{ij}^{(1)}| - |B_{ij}^{(0)}|)}_{I_5} \\
&= I_1 + I_2 + I_3 + I_4 + I_5.
\end{aligned}$$

Let

$$I_4 = \lambda \sum_{(i,j) \in S_A^c} |A_{ij}^{(1)}| + \lambda \sum_{i \neq j, (i,j) \in S_A} (|A_{ij}^{(1)}| - |A_{ij}^{(0)}|) = I_{41} + I_{42},$$

$$I_5 = \gamma \sum_{(i,j) \in S_B^c} |B_{ij}^{(1)}| + \gamma \sum_{i \neq j, (i,j) \in S_B} (|B_{ij}^{(1)}| - |B_{ij}^{(0)}|) = I_{51} + I_{52}.$$

Using Taylor's expansion with integral remainder, we have

$$I_1 = -q\text{tr}(A_0^{-1}\Delta_1) + q\text{vec}(\Delta_1)^T \Big\{ \int_0^1 (1 - v) g(v, A_v) dv \Big\} \text{vec}(\Delta_1), \tag{7}$$

$$I_2 = -p\text{tr}(B_0^{-1}\Delta_2) + p\text{vec}(\Delta_2)^T \Big\{ \int_0^1 (1 - v) g(v, B_v) dv \Big\} \text{vec}(\Delta_2), \tag{8}$$

where $g(v, A_v) = A_v^{-1} \otimes A_v^{-1}$, $A_v = A_0 + v\Delta_1$ and $g(v, B_v)$ can be similarly defined. By the bilinearity property of the Kronecker product, we have

$$I_3 = \text{tr}\{(\hat{R}_p - \Sigma_0)(\Delta_2 \otimes A_0)\} + \text{tr}\{(\hat{R}_p - \Sigma_0)(B_0 \otimes \Delta_1)\} + \text{tr}\{(\hat{R}_p - \Sigma_0)(\Delta_2 \otimes \Delta_1)\} + \text{tr}\{\Sigma_0(\Delta_2 \otimes A_0)\} + \text{tr}\{\Sigma_0(B_0 \otimes \Delta_1)\} + \text{tr}\{\Sigma_0(\Delta_2 \otimes \Delta_1)\}. \quad (9)$$

Following from the fact that $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$, and $\text{tr}(A \otimes B) = \text{tr}(A)\text{tr}(B)$, we have

$$\text{tr}\{\Sigma_0(\Delta_2 \otimes A_0)\} = \text{tr}\{(B_0^{-1}\Delta_2) \otimes I\} = p\text{tr}(B_0^{-1}\Delta_2). \quad (10)$$

Combining equations (7), (8), (9) and (10), we have

$$\begin{aligned} I_1 + I_2 + I_3 = & \underbrace{q\text{vec}(\Delta_1)^T \left\{ \int_0^1 (1-v) g(v, A_v) dv \right\} \text{vec}(\Delta_1)}_{K_1} \\ & + \underbrace{p\text{vec}(\Delta_2)^T \left\{ \int_0^1 (1-v) g(v, B_v) dv \right\} \text{vec}(\Delta_2)}_{K_2} + \underbrace{\text{tr}\{(\hat{R}_p - \Sigma_0)(\Delta_2 \otimes A_0)\}}_{K_3} \\ & + \underbrace{\text{tr}\{(\hat{R}_p - \Sigma_0)(B_0 \otimes \Delta_1)\}}_{K_4} + \underbrace{\text{tr}\{(\hat{R}_p - \Sigma_0)(\Delta_2 \otimes \Delta_1)\}}_{K_5} + \underbrace{\text{tr}(A_0^{-1}\Delta_1)\text{tr}(B_0^{-1}\Delta_2)}_{K_6} \\ = & K_1 + K_2 + K_3 + K_4 + K_5 + K_6. \end{aligned}$$

By similar arguments as in Lam & Fan (2009), $K_1$ and $K_2$ can be bounded from below,

$$\begin{aligned} K_1 &\geq (q\|\Delta_1\|_F^2/2) \min_{0 \leq v \leq 1} \lambda_{\max}^{-2}(A_v) \\ &\geq (q\|\Delta_1\|_F^2/2)(\|A_0\|_s + \|\Delta_1\|_s)^{-2} \\ &\geq \frac{1}{2} q C_1^2 \alpha_n^2 \{\delta_2 + o(1)\}^{-2}, \end{aligned}$$

where we use $\|\Delta_1\| \leq \|\Delta_1\|_F = o(1)$. Similarly,

$$K_2 \geq \frac{1}{2} p C_2^2 \beta_n^2 \{\delta_4 + o(1)\}^{-2}.$$

Next, we will show that $|K_4|$ is dominated by $K_1 + I_{41}$, i.e., $|K_4| < K_1 + I_{41}$ for sufficiently large $n$ and $C_1$. In other words, we say $|K_4|$ is dominated by $K_1 + I_{41}$. Then

$$\begin{aligned} |K_4| &\leq \sum_{i \neq k} |(\hat{R}_p - \Sigma_0)_{ik}(B_0 \otimes \Delta_1)_{ik}| \\ &\leq \underbrace{\|\hat{R}_p - \Sigma_0\|_{\max} \sum_{i \neq j} |(\Delta_1)_{ij}| \sum_{k \neq \ell} |(B_0)_{k\ell}|}_{K_{41}} + \underbrace{\|\hat{R}_p - \Sigma_0\|_{\max} \sum_{i = j} |(\Delta_1)_{ij}| \sum_{k \neq \ell} |(B_0)_{k\ell}|}_{K_{42}} \\ &\quad + \underbrace{\|\hat{R}_p - \Sigma_0\|_{\max} \sum_{i \neq j} |(\Delta_1)_{ij}| \sum_{k = \ell} |(B_0)_{k\ell}|}_{K_{43}} = K_{41} + K_{42} + K_{43}. \end{aligned}$$

Let us consider $K_{41}$, $K_{42}$ and $K_{43}$ separately. By Theorem 2, we know that

$$\|\hat{R}_p - \Sigma_0\|_{\max} = O_p\left\{\left(\frac{\log p + \log q}{n}\right)^{1/2}\right\}.$$

For $K_{41}$, we have

$$\begin{aligned}
K_{41} &= O_p\left\{\left(\frac{\log p + \log q}{n}\right)^{1/2}\right\} \sum_{i\neq j} |(\Delta_1)_{ij}| \sum_{k\neq \ell} |(B_0)_{k\ell}| \\
&= O_p\left\{\left(\frac{\log p + \log q}{n}\right)^{1/2}\right\} \left\{\sum_{i\neq j,(i,j)\in S_A} |(\Delta_1)_{ij}| + \sum_{i\neq j,(i,j)\in S_A^c} |(\Delta_1)_{ij}|\right\} \sum_{k\neq \ell} |(B_0)_{k\ell}| \\
&\leq O_p\left\{\left(\frac{\log p + \log q}{n}\right)^{1/2}\right\} \left(s_1^{1/2}\|\Delta_1\|_F + \sum_{i\neq j,(i,j)\in S_A^c} |A_{ij}^{(1)}|\right) s_2^{1/2}\|B_0\|_F \\
&\leq O_p\left\{\left(\frac{\log p + \log q}{n}\right)^{1/2}\right\} (s_1 s_2 q)^{1/2}\|B_0\|_s C_1\alpha_n \\
&\quad + O_p\left\{\left(\frac{\log p + \log q}{n}\right)^{1/2}\right\} (s_2 q)^{1/2}\|B_0\|_s \sum_{i\neq j,(i,j)\in S_A^c} |A_{ij}^{(1)}|.
\end{aligned}$$

Since $\|B_0\| \leq \delta_4$ by condition (C2), we can show that the first term is dominated by $K_1$ for large enough $C_1$. The second term is also dominated by $I_{41}$, by condition (C3),

$$\frac{\log p + \log q}{n}(s_2 q) = O(\lambda^2).$$

Hence $|K_{41}|$ is dominated by $K_1 + I_{41}$. Similarly for $K_{43}$, we have

$$\begin{aligned}
K_{43} &= O_p\left\{\left(\frac{\log p + \log q}{n}\right)^{1/2}\right\} \sum_{i\neq j} |(\Delta_1)_{ij}| \sum_{k=\ell} |(B_0)_{k\ell}| \\
&= O_p\left\{\left(\frac{\log p + \log q}{n}\right)^{1/2}\right\} \left\{\sum_{i\neq j,(i,j)\in S_A} |(\Delta_1)_{ij}| + \sum_{i\neq j,(i,j)\in S_A^c} |(\Delta_1)_{ij}|\right\} \sum_{k=\ell} |(B_0)_{k\ell}| \\
&\leq O_p\left\{\left(\frac{\log p + \log q}{n}\right)^{1/2}\right\} \left(s_1^{1/2}\|\Delta_1\|_F + \sum_{i\neq j,(i,j)\in S_A^c} |A_{ij}^{(1)}|\right) q^{1/2}\|B_0\|_F \\
&\leq O_p\left\{\left(\frac{\log p + \log q}{n}\right)^{1/2}\right\} s_1^{1/2} q\|B_0\|_s C_1\alpha_n \\
&\quad + O_p\left\{\left(\frac{\log p + \log q}{n}\right)^{1/2}\right\} q\|B_0\|_s \sum_{i\neq j,(i,j)\in S_A^c} |A_{ij}^{(1)}|.
\end{aligned}$$

Similarly, the first term is dominated by $K_1$. The second term is also dominated by $I_{41}$, by condition (C3). For $K_{42}$, by the Cauchy–Schwartz inequality, we have

$$K_{42} \leq O_p\left\{\left(\frac{\log p + \log q}{n}\right)^{1/2}\right\} p^{1/2}\|\Delta_1\|_F (s_2 q)^{1/2}\|B_0\|_s \leq O_p(q\alpha_n^2).$$

Combining the upper bounds for $K_{41}$, $K_{42}$ and $K_{43}$, we know that $|K_4|$ is dominated by $K_1 + I_{41}$. Likewise, $|K_3|$ is also dominated by $K_2 + I_{51}$. The next step is to provide the bound for $|K_5|$. Note that

$$|K_5| \le \underbrace{\|\hat{R}_p - \Sigma_0\|_{\max} \sum_{i\neq j} |(\Delta_1)_{ij}| \sum_{k\neq \ell} |(\Delta_2)_{k\ell}|}_{K_{51}} + \underbrace{\|\hat{R}_p - \Sigma_0\|_{\max} \sum_{i=j} |(\Delta_1)_{ij}| \sum_{k\neq \ell} |(\Delta_2)_{k\ell}|}_{K_{52}}$$
$$\underbrace{+\|\hat{R}_p - \Sigma_0\|_{\max} \sum_{i\neq j} |(\Delta_1)_{ij}| \sum_{k=\ell} |(\Delta_2)_{k\ell}|}_{K_{53}} = K_{51} + K_{52} + K_{53}.$$

Then

$$K_{51} + K_{53} \le O_p\left\{\left(\frac{\log p + \log q}{n}\right)^{1/2}\right\}\left({s_1}^{1/2}\|\Delta_1\|_F + \sum_{i\neq j,(i,j)\in S_A^c} |A_{ij}^{(1)}|\right)\sum_{k,\ell}|(\Delta_2)_{k\ell}|$$
$$\le O_p\left\{\left(\frac{\log p + \log q}{n}\right)^{1/2}\right\}\left(s_1^{1/2}\alpha_n + \sum_{i\neq j,(i,j)\in S_A^c} |A_{ij}^{(1)}|\right)q\beta_n.$$

It is straightforward to verify that

$$\left(\frac{\log p + \log q}{n}\right){s_1}^{1/2} q\alpha_n\beta_n \le O_p(q\alpha_n^2 + p\beta_n^2).$$

Moreover, the second term is dominated by $I_{41}$, i.e.,

$$q\beta_n\left(\frac{\log p + \log q}{n}\right) \le O_p(\lambda).$$

Thus, $K_{51} + K_{53}$ is dominated by $K_1 + I_{41}$. Symmetrically, $K_{51} + K_{52}$ is dominated by $K_2 + I_{51}$. Then, $|K_5|$ is dominated by $K_1 + I_{41} + K_2 + I_{51}$.

By the Cauchy–Schwartz inequality, we get,

$$|K_6| \le \left\{pq\text{tr}(\Delta_1 A_0^{-1}\Delta_1 A_0^{-1})\text{tr}(\Delta_2 B_0^{-1}\Delta_2 B_0^{-1})\right\}^{1/2}$$
$$\le \frac{q}{2}\text{tr}(\Delta_1 A_0^{-1}\Delta_1 A_0^{-1}) + \frac{p}{2}(\Delta_2 B_0^{-1}\Delta_2 B_0^{-1}). \tag{11}$$

To show $|K_6|$ is bounded by $K_1 + K_2$, we need the fact that $A_v^{-1} = (A_0 + v\Delta_1)^{-1} = A_0^{-1} + O_p(\Delta_1)$, and $\text{tr}(A^T BCD^T) = \text{vec}(A)^T(D \otimes B)\text{vec}(C)$. Then

$$K_1 = q\int_0^1 (1-v)\text{tr}(\Delta_1 A_v^{-1}\Delta_1 A_v^{-1})dv$$
$$\ge \frac{q}{2}\min_{0\le v\le 1}\text{tr}(\Delta_1 A_v^{-1}\Delta_1 A_v^{-1})$$
$$= \frac{q}{2}\text{tr}(\Delta_1 A_0^{-1}\Delta_1 A_0^{-1})\{1 + o_p(1)\}. \tag{12}$$

Similarly, $K_2 \ge p\text{tr}(\Delta_2 B_0^{-1}\Delta_2 B_0^{-1})\{1 + o_p(1)\}/2$. Therefore, from (11) and (12), $|K_6|$ is bounded above by $K_1 + K_2$.

Up to now, we have shown that $I_1 + I_2 + I_3$ is dominated by $K_1 + I_{41} + K_2 + I_{51}$. If we can show that $I_{42}$ and $I_{52}$ are dominated by $K_1 > 0$ and $K_2 > 0$ respectively, then $\phi_p(A_1, B_1) -$

$\phi_p(A_0, B_0)$ is bounded from below by a positive constant independent of $\Delta_1$ and $\Delta_2$. Therefore, (6) is true. Note that

$$|I_{42}| \leq \lambda \sum_{i \neq j, (i,j) \in S_A} |A_{ij}^{(1)} - A_{ij}^{(0)}| \leq \lambda s_1^{1/2} \|\Delta_1\|_F \leq O_p(q\alpha_n^2),$$

where the last step follows from the condition (C3). Similarly, $|I_{52}|$ is dominated by $K_2$. This completes the proof.

## PROOF OF COROLLARY 2

By the property of Kronecker product, $\|C \otimes D\|_F = \|C\|_F \|D\|_F$, and $\|C \otimes D\|_s = \|C\|_s \|D\|_s$, and Theorem 3, we have

$$\begin{aligned}
&\|\hat{\Omega} - \Omega_0\|_F \\
&= \|\hat{B} \otimes \hat{A} - B_0 \otimes \hat{A} + B_0 \otimes \hat{A} - B_0 \otimes A_0\|_F \\
&\leq \|(\hat{B} - B_0) \otimes \hat{A}\|_F + \|B_0 \otimes (\hat{A} - A_0)\|_F \\
&= \|\hat{B} - B_0\|_F \|\hat{A}\|_F + \|B_0\|_F \|\hat{A} - A_0\|_F \\
&\leq \|\hat{B} - B_0\|_F \|\hat{A} - A_0\|_F + \|\hat{B} - B_0\|_F \|A_0\|_F + \|B_0\|_F \|\hat{A} - A_0\|_F \\
&\leq \|\hat{B} - B_0\|_F \|\hat{A} - A_0\|_F + \|\hat{B} - B_0\|_F \|A_0\| p^{1/2} + \|B_0\| q^{1/2} \|\hat{A} - A_0\|_F \\
&= O_p\left\{\frac{(s_1 s_2 + p s_2 + q s_1)\log(pq)}{n(pq)^{1/2}}\right\} + O_p\left[\left\{\frac{(s_1 s_2 + p s_2 + q s_1)(\log p + \log q)}{n}\right\}^{1/2}\right] \\
&= O_p\left[\left\{\frac{(s_1 s_2 + p s_2 + q s_1)(\log p + \log q)}{n}\right\}^{1/2}\right].
\end{aligned}$$

Similarly, in terms of the spectral norm, we have

$$\begin{aligned}
&\|\hat{\Omega} - \Omega_0\|_s \\
&= \|\hat{B} \otimes \hat{A} - B_0 \otimes \hat{A} + B_0 \otimes \hat{A} - B_0 \otimes A_0\|_s \\
&\leq \|(\hat{B} - B_0) \otimes \hat{A}\|_s + \|B_0 \otimes (\hat{A} - A_0)\|_s \\
&= \|\hat{B} - B_0\|_s \|\hat{A}\|_s + \|B_0\|_s \|\hat{A} - A_0\|_s \\
&\leq \|\hat{B} - B_0\|_s \|\hat{A} - A_0\|_s + \|\hat{B} - B_0\|_s \|A_0\|_s + \|B_0\|_s \|\hat{A} - A_0\|_s \\
&= O_p\left\{\frac{(s_1 s_2 + p s_2 + q s_1)\log(pq)}{n(pq)^{1/2}}\right\} + O_p\left[\left\{\frac{(s_1 s_2 + p s_2 + q s_1)(\log p + \log q)}{n}\right\}^{1/2}\left(\frac{1}{p} + \frac{1}{q}\right)\right] \\
&= O_p\left[\left\{\frac{(s_1 s_2 + p s_2 + q s_1)(\log p + \log q)}{n}\right\}^{1/2}\left(\frac{1}{p} + \frac{1}{q}\right)\right].
\end{aligned}$$

## SIMULATION RESULTS

Using the procedure described in our main paper, we can generate two precision matrices $A$ and $B$. Since only the correlation matrices in the nonparanormal bigraphical model are estimable, we need to rescale $A$ and $B$ such that the diagonal elements of $A^{-1}$ and $B^{-1}$ are 1. Given the inverse covariance matrices $A$ and $B$, we first calculate $A^{-1} = (a_{ij})$ and $B^{-1} = (b_{ij})$, then

calculate the correlation matrices $U = \{a_{ij}/(a_{ii}a_{jj})^{1/2}\}$ and $V = \{b_{ij}/(b_{ii}b_{jj})^{1/2}\}$, and finally obtain the precision matrices $U^{-1}$ and $V^{-1}$.

The definition of the matrix-t distribution is as follows.

DEFINITION 1. *A $p \times q$ random matrix $X$ follows a matrix-t distribution* $\text{MT}(M; U, V; e)$ *with mean matrix $M$, row covariance component matrix $U$, column covariance component matrix $V$ and $e$ degrees of freedom, if and only if the density of $X$ is*

$$p(X) = K(U,V)|I_p + U^{-1}(X-M)V^{-1}(X-M)^T|^{-(e+p+q-1)/2}, \tag{13}$$

*where*

$$K(U,V) = |U|^{-q/2}|V|^{-p/2}\frac{\Gamma_q\{(e+p+q-1)/2\}}{(e\pi)^{pq/2}\Gamma_q\{(e+q-1)/2\}}.$$

*Here $\Gamma_q$ is the multivariate gamma function.*

Similar to the matrix-normal distribution, the matrix-t distribution is also a special case of the multivariate t-distribution whose variance has a Kronecker product structure. Hence, $X_1, ..., X_n \sim \text{MT}(0; A^{-1}, B^{-1}, e)$ can be simulated using the R function `rmvt`. Once $X_1, ..., X_n$ are simulated, one can apply our estimation procedure and that of Yin & Li (2012) to the data. The simulation results under scenarios (a) and (b) are given in section 5.1 of our main paper.

Under simulation scenario (d), the mean total number of non-zero off-diagonal elements in the precision matrix is 630. Under simulation scenario (c), the mean total number of non-zero off-diagonal elements in the row and column precision matrices are 1685 and 590, respectively. Figures 1 and 2 show the plot of the mean number of true positive edges against the mean total number of edges detected for different tuning parameters $\lambda$ based on 100 replicates under simulation scenarios (c) and (d).

## FIGURES FOR THE GENOMIC DATA

Figure 3 presents the quantile-quantile plot for some gene expression levels to examine the normality assumption in Yin & Li (2012). Figure 4 and 5 show the estimated gene graphs with 27 edges and tissue graphs with 15 edges, based on our method and the method of Yin & Li (2012) respectively.

## BINARY BIGRAPHICAL MODEL

Let $Y_i = \text{vec}(Z_i)$. For $n$ independent copies of $(Y_{ij}, Y_{ik})$, $(i = 1, ..., n)$, Kendall's tau is defined as $\hat{\tau}_{jk} = \{n(n-1)\}^{-1}\sum_{i\neq i'}(Y_{ij} - Y_{i'j})(Y_{ik} - Y_{i'k})$. Let $\Phi_2(u, v, t)$ be the cumulative distribution function of a standard bivariate normal distribution with correlation $t$. The following theorem shows that the underlying correlation matrix $\Sigma = V \otimes U$, can be recovered by Kendall's tau.

THEOREM 1. *Kendall's tau $\hat{\tau}_{jk}$ is a consistent estimator of $F(\Sigma_{jk})$, where the function $F(t)$ is*

$$F(t) = 2\{\Phi_2(\Delta_j, \Delta_k, t) - \Phi(\Delta_j)\Phi(\Delta_k)\}.$$

*When $\Delta_j = \Delta_k = 0$, $F(t)$ can be simplified to $F(t) = \pi^{-1}\sin^{-1}t$ and $\Sigma_{jk}$ can be consistently estimated by* $\sin(\pi\hat{\tau}_{jk})$.

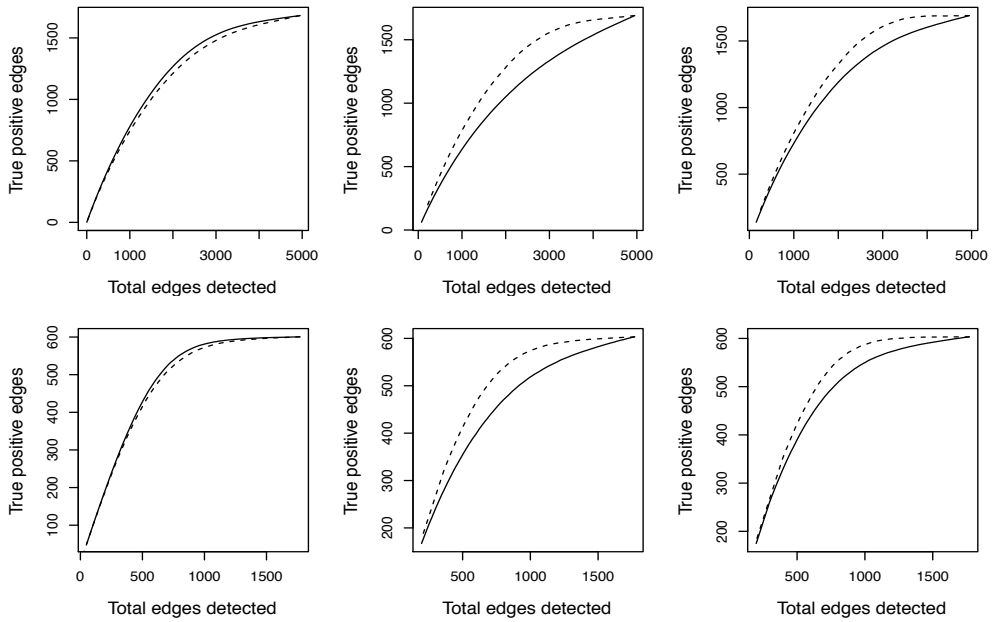


Fig. 1. Plots of the mean number of true positive edges against the mean total number of edges detected for different tuning parameters $\lambda$ based on 100 replicates of simulation scenario (c). The top left panel is for estimating the row matrix in matrix-normal data, the top middle panel is for estimating the row matrix in matrix-nonparanormal data, the top right panel is for estimating the row matrix in matrix-t data, the bottom left panel is for estimating the column matrix in matrix-normal data, the bottom middle panel is for estimating the column matrix in matrix-nonparanormal data and the bottom right panel is for estimating the column matrix in matrix-t data. The solid and dashed lines represent the method of Yin & Li (2012) and our method.

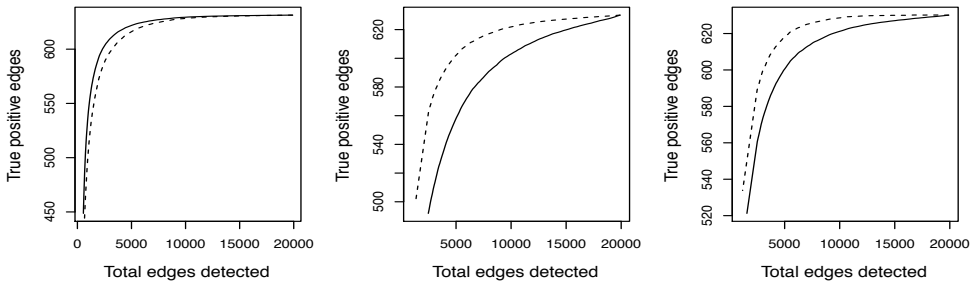


Fig. 2. Plots of the mean number of true positive edges against the mean total number of edges detected for different tuning parameters $\lambda$ based on 100 replicates of simulation scenario (d). The top left panel is for estimating the row matrix in matrix-normal data, the top middle panel is for estimating the row matrix in matrix-nonparanormal data, the top right panel is for estimating the row matrix in matrix-t data, the bottom left panel is for estimating the column matrix in matrix-normal data, the bottom middle panel is for estimating the column matrix in matrix-nonparanormal data and the bottom right panel is for estimating the column matrix in matrix-t data. The solid and dashed lines represent the method of Yin & Li (2012) and our method.

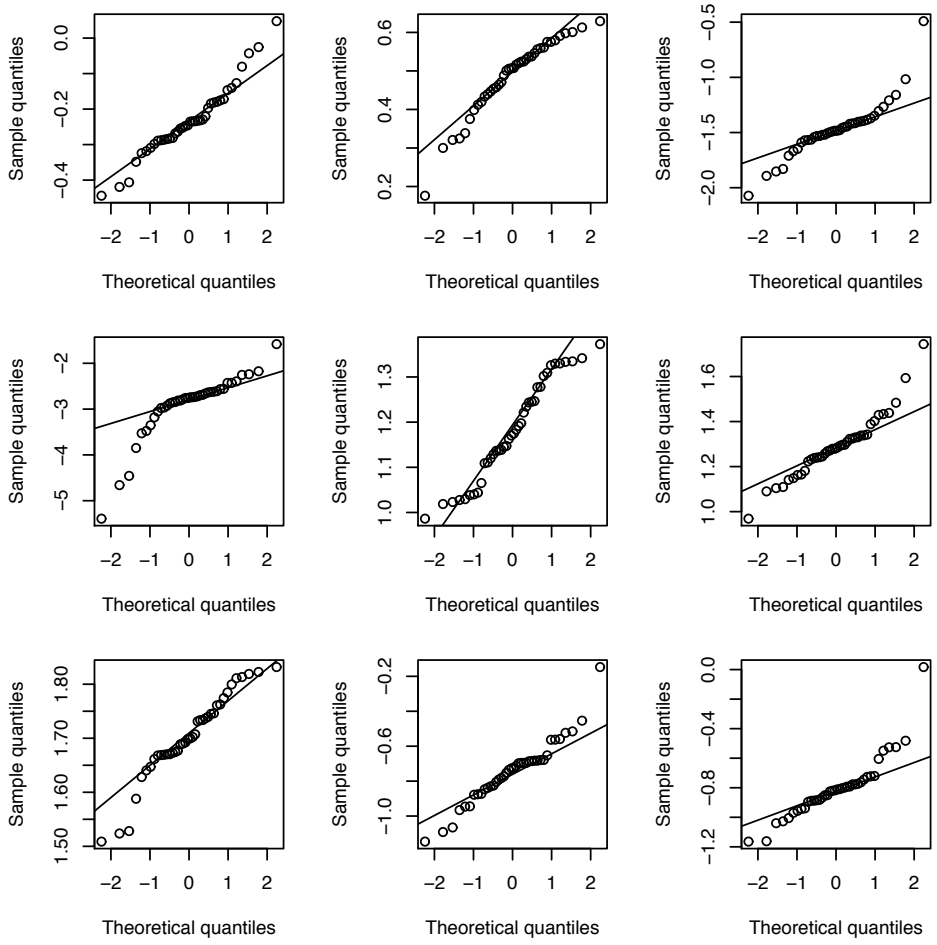


Fig. 3. Examples of the quantile-quantile plot for the gene expression data.

*Proof.* For $n$ independent pairs of binary data $(Y_{ij}, Y_{ik})$, where $i = 1, ..., n$, Kendall's tau reduces to

$$\begin{aligned}\hat{\tau}_{jk} &= \{n(n-1)\}^{-1} \sum_{i \neq i'} (Y_{ij} - Y_{i'j})(Y_{ij} - Y_{i'j}) \\ &= 2(ad - bc)/\{n(n-1)\},\end{aligned}$$

where $a, b, b, d$ are the total number of pairs $(1,1)$, $(0,1)$, $(1,0)$ and $(1,1)$ in $(Y_{ij}, Y_{ik})$. By the law of large numbers, $\hat{\tau}_{jk}$ is consistent for $F(\Sigma_{jk})$, where

$$\begin{aligned}F(t) &= 2[L(\Delta_j, \Delta_k, t)\Phi_2(\Delta_j, \Delta_k, t) - \{\Phi(\Delta_j) - \Phi_2(\Delta_j, \Delta_k, t)\}\{\Phi(\Delta_k) - \Phi_2(\Delta_j, \Delta_k, t)\}] \\ &= 2\{\Phi_2(\Delta_j, \Delta_k, t) - \Phi(\Delta_j)\Phi(\Delta_k)\},\end{aligned}$$

where $L(\Delta_j, \Delta_k, t) = 1 - \Phi(\Delta_j) - \Phi(\Delta_k) + \Phi_2(\Delta_j, \Delta_k, t)$. When $\Delta_j = \Delta_k = 0$, $F(t) = 2\Phi_2(0, 0, t) - 1/2 = \pi^{-1} \sin^{-1} t$, where the last step follows from the Sheppard's theorem (Sheppard, 1899). □

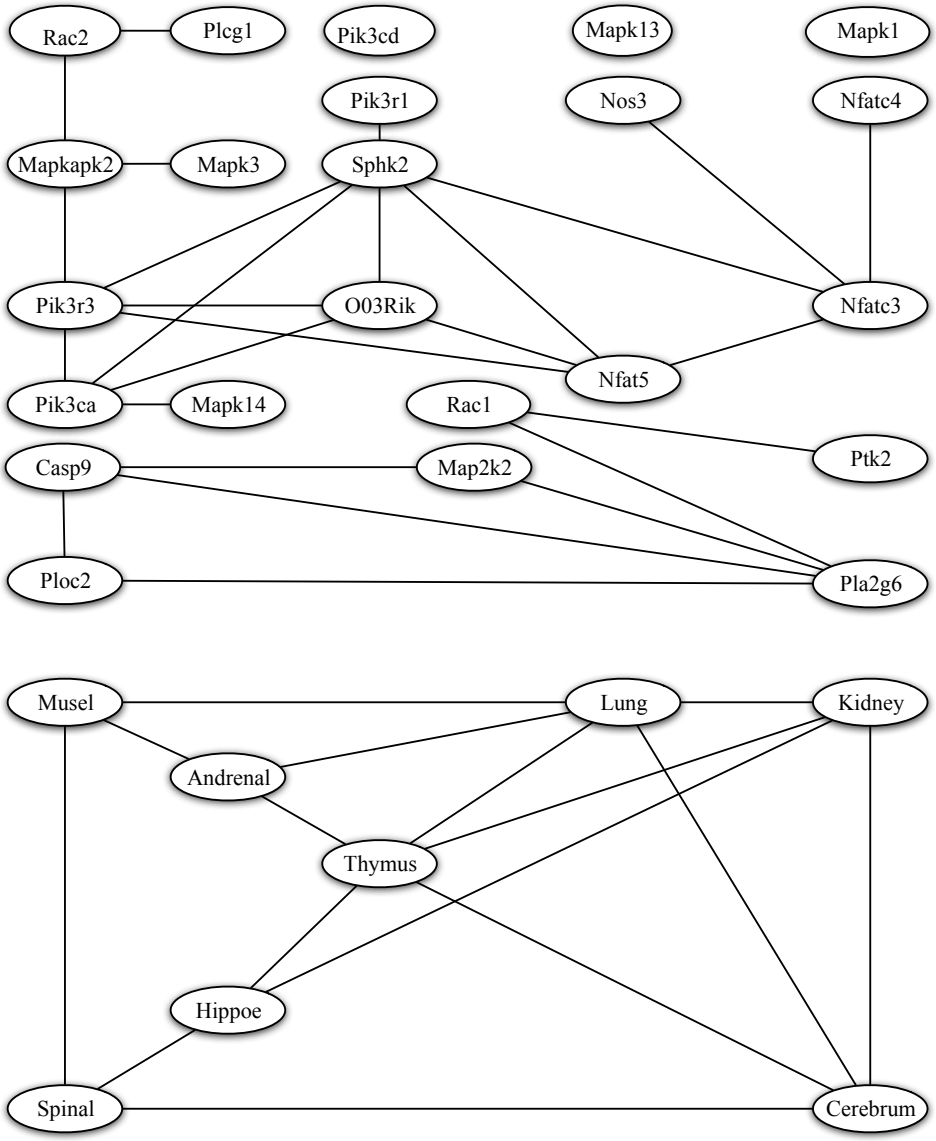


Fig. 4. Estimated gene network (top panel) and tissue network (bottom panel) for the gene expression data based on our method. For the gene network, many isolated nodes are not plotted.

## MIXED BINARY AND CONTINUOUS BIGRAPHICAL MODEL

In the section, we consider the bigraphical model with both binary and continuous observations. For a $p \times q$ matrix-valued random variable $Z$, we assume there exists an underlying matrix-valued random variable $X \sim \text{MNPN}(U, V; f)$, satisfying $Z_{jk} = I(X_{jk} > C_{jk})$ if $Z_{jk}$ is binary, and $Z_{jk} = X_{jk}$ if $Z_{jk}$ is continuous, where $C_{jk}$ is a constant. Given $n$ independent copies of matrix-valued data $Z_1, ..., Z_n$, our aim is to estimate the precision matrices $A = U^{-1}$ and $B = V^{-1}$ respectively.

For binary $Z_{jk}$, we can similarly estimate $\Delta_{jk} = f_{jk}(C_{jk})$. Denote $Y_i = \text{vec}(Z_i)$. If $(Y_{ij}, Y_{ik})$ are both binary, Kendall's tau can be used to estimate $\Sigma = V \otimes U$ as shown in Theorem 5. The rank-based estimators are consider by Liu et al. (2012), if $(Y_{ij}, Y_{ik})$ are both continuous. The

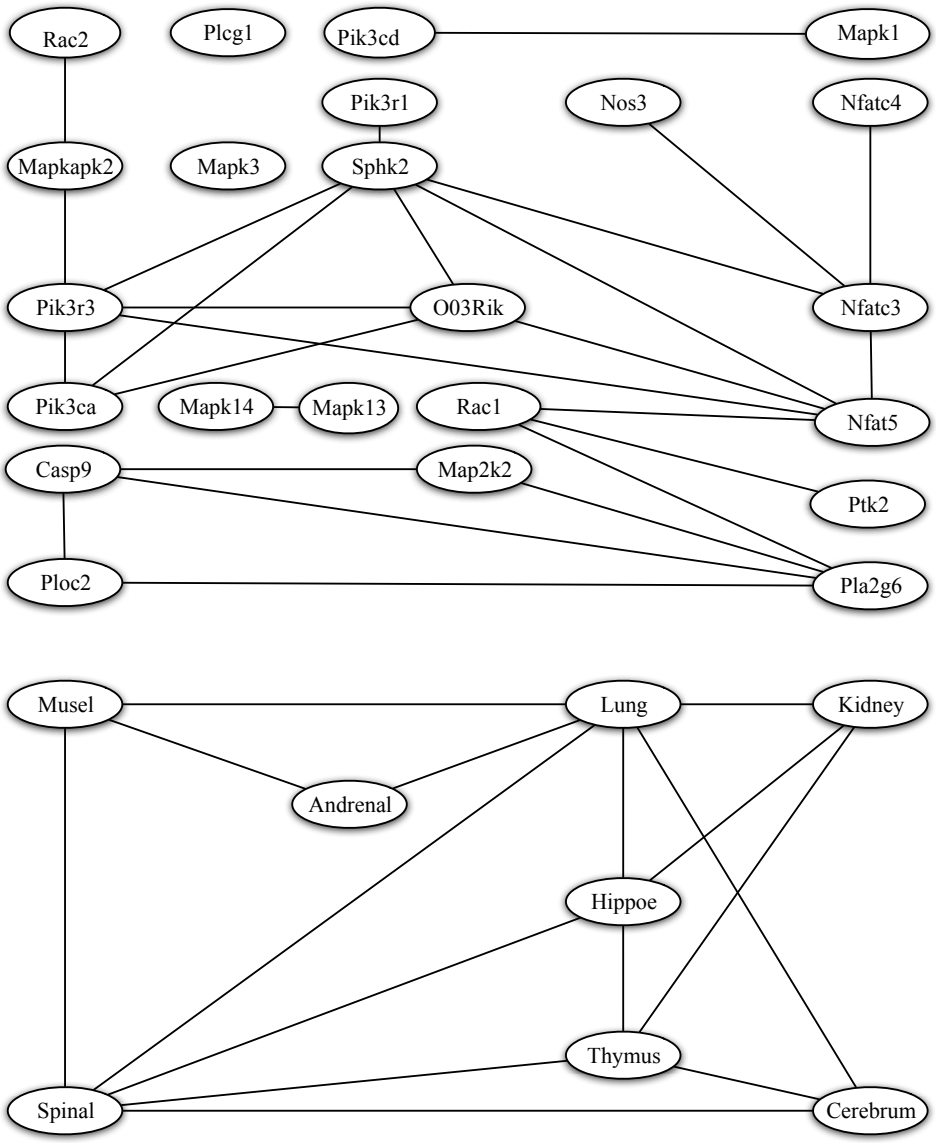


Fig. 5. Estimated gene network (top panel) and tissue network (bottom panel) for the gene expression data based on the method of Yin & Li (2012). For the gene network, many isolated nodes are not plotted.

following proposition considers the case where one element of $(Y_{ij}, Y_{ik})$ is binary and the other one is continuous.

PROPOSITION 1. *Assume that $Y_{ij}$ is binary and $Y_{ik}$ is continuous. Kendall's tau, $\hat{\tau}_{jk}$ is a consistent estimator of $H(\Sigma_{jk})$, where*

$$H(t) = 4\Phi_2(\Delta_j, 0, t/2^{1/2}) - 2\Phi(\Delta_j),$$

*where $\Delta = \text{vec}\{(\Delta_{jk})\}$. If $\Delta_j = 0$, then $H(t) = 2\pi^{-1}\sin^{-1}(2^{-1/2}t)$, and $\Sigma_{jk}$ can be consistently estimated by $2^{1/2}\sin(\pi\hat{\tau}_{jk}/2)$.*

AN EM ALGORITHM FOR MISSING DATA IMPUTATION

We follow similar notations to Allen & Tibshirani (2010). The subscripts $o$ and $m$ indicate the observed part and missing part respectively. For instance, $X_{i,o}$ is the observe part of $X_i$. Let $T_i = f(X_i)$, $i = 1, ..., n$. Recall that the penalized negative log-likelihood is

$$w(A, B, f) = -q \log |A| - p \log |B| + \frac{1}{n} \sum_{i=1}^{n} \text{tr}(T_i^T A T_i B) + \lambda \sum_{i \neq j} |A_{ij}| + \gamma \sum_{i \neq j} |B_{ij}|.$$

Given the current parameter $\theta' = (f', A', B')$, in the E step of the EM algorithm, we calculate $Q(\theta \mid X_o, \theta')$, the conditional expectation of $w(A, B, f)$ given the observed data and the parameter $\theta'$. By Proposition 3 in Allen & Tibshirani (2010), we know

$$\begin{aligned} E\{\text{tr}(T^T A T B) \mid X_o, \theta'\} &= \text{tr}[\{\hat{T}^T A \hat{T} + G(A)\}B] \\ &= \text{tr}[\{\hat{T} B \hat{T}^T + F(B)\}A], \end{aligned}$$

where $\hat{T} = E(T \mid X_o, \theta')$, $G(A)$ is a $q \times q$ matrix $G(A) = \{\text{tr}(C^{(ij)} A)\}$ with $C^{(ij)} = \text{cov}(T_{*i}, T_{*j} \mid X_o, \theta')$, and $F(B)$ is a $p \times p$ matrix $F(B) = \{\text{tr}(D^{(ij)} B)\}$ with $D^{(ij)} = \text{cov}(T_{i*}, T_{j*} \mid X_o, \theta')$. Note that $C^{(ij)}$ and $D^{(ij)}$ can be calculated using the formula

$$\text{cov}\{\text{vec}(T)_m, \text{vec}(T)_m \mid X_o, \theta'\} = \Sigma_{mm} - \Sigma_{mo} \Sigma_{oo}^{-1} \Sigma_{om}.$$

Let $m$ be the indices of the missing values of $\text{vec}(T)$ and $o$ be the observed. Since $T$ satisfies the matrix-normal distribution, we have

$$\text{vec}(\hat{T})_k = \begin{cases} \Sigma_{ko} \Sigma_{oo}^{-1} \text{vec}(T)_o, & \text{if } k \in m, \\ \text{T}_k, & \text{if } k \in o. \end{cases}$$

Then in the E step, we can calculate Q function as

$$\begin{aligned} Q(\theta \mid X_o, \theta') &= -q \log |A| - p \log |B| + \frac{1}{n} \sum_{i=1}^{n} \text{tr}[\{\hat{T}_i^T A \hat{T}_i + G(A)\}B] + \lambda \sum_{i \neq j} |A_{ij}| + \gamma \sum_{i \neq j} |B_{ij}| \\ &= -q \log |A| - p \log |B| + \frac{1}{n} \sum_{i=1}^{n} \text{tr}[\{\hat{T}_i B \hat{T}_i^T + F(B)\}A] + \lambda \sum_{i \neq j} |A_{ij}| + \gamma \sum_{i \neq j} |B_{ij}|. \end{aligned}$$

Now, let us consider the M step. Similar to Liu et al. (2009), we propose a normal-score method to estimate the unknown function $f_{jk}(\cdot)$. If $X$ is fully observed, $f(X) \sim N(0, 1)$ implies $\text{pr}\{f(X) < t^*\} = \Phi(t^*)$. Taking $t = f^{-1}(t^*)$, then $\text{pr}(X < t) = \Phi\{f(t)\}$. Replacing $\text{pr}(X < t)$ by the corresponding empirical cumulative distribution function, we get a normal-score estimate of $f(\cdot)$. Denote $\hat{T}_i^{(j\ell)}$ the $(j, \ell)$th element of $\hat{T}_i$. Since $\hat{T}_i^{(j\ell)}$ is imputed, given the current value of $f'_{j\ell}$, we can update $f_{j\ell}$ by

$$\hat{f}_{j\ell}(t) = \Phi^{-1} \left[ \frac{1}{n} \sum_{i=1}^{n} I \left\{ \hat{T}_i^{(j\ell)} < f'_{j\ell}(t) \right\} \right]. \tag{14}$$

To estimate $A$ and $B$, we have to iteratively minimize $Q(\theta \mid X_o, \theta')$ with respect to $A$ and $B$ separately. Given an estimator $\hat{B}$ and estimated functions $\hat{f}$,

$$Q(\hat{f}, A, \hat{B} \mid X_o, \theta') \propto -q \log |A| + \frac{1}{n} \sum_{i=1}^{n} \text{tr}[\{\hat{T}_i \hat{B} \hat{T}_i^T + F(\hat{B})\}A] + \lambda \sum_{i \neq j} |A_{ij}|. \tag{15}$$

Then the estimator of $A$ is obtained by minimizing $Q(\hat{f}, A, \hat{B} \mid X_o, \theta')$. Similarly, given an estimator $\hat{A}$ and estimated functions $\hat{f}$,

$$Q(\hat{f}, \hat{A}, B \mid X_o, \theta') \propto -p \log |B| + \frac{1}{n} \sum_{i=1}^{n} \text{tr}[\{\hat{T}_i^T \hat{A} \hat{T}_i + G(\hat{A})\} B] + \gamma \sum_{i \neq j} |B_{ij}|. \quad (16)$$

Then the estimator of $B$ is obtained by minimizing $Q(\hat{f}, \hat{A}, B \mid X_o, \theta')$.

In summary, we develop the following EM algorithm for handling missing data in our matrix-nonparanormal model.

1. Set $A = I_p$, $B = I_q$, $f(t) = t$ and $k = 1$.
2. Given the current parameter $(\hat{f}_{j\ell}^{(k)}, \hat{A}^{(k)}, \hat{B}^{(k)})$, for any $j = 1, ..., p$ and $\ell = 1, ..., q$,
$$\hat{f}_{j\ell}^{(k+1)}(t) = \Phi^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} I\{\hat{T}_i^{(j\ell)} < \hat{f}_{j\ell}^{(k)}(t)\}\right].$$
3. E step for estimating $A$: Calculate $\hat{T}_i \hat{B}^{(k)} \hat{T}_i^T + F(\hat{B}^{(k)})$, for any $i = 1, ..., n$.
4. M step for estimating $A$: Minimize $Q(\hat{f}^{(k+1)}, A, \hat{B}^{(k)} \mid X_o, \hat{f}^{(k+1)}, \hat{A}^{(k)}, \hat{B}^{(k)})$ in (15) with respect to $A$ to obtain $\hat{A}^{(k+1)}$.
5. E step for estimating $B$: Calculate $\hat{T}_i^T \hat{A}^{(k+1)} \hat{T}_i + G(\hat{A}^{(k+1)})$, for any $i = 1, ..., n$.
6. M step for estimating $B$: Minimize $Q(\hat{f}^{(k+1)}, \hat{A}^{(k+1)}, B \mid X_o, \hat{f}^{(k+1)}, \hat{A}^{(k+1)}, \hat{B}^{(k)})$ in (16) with respect to $B$ to obtain $\hat{B}^{(k+1)}$.
7. Repeat steps 2-6 until convergence or $k = K$, where $K$ is a specified integer.

As discussed by Allen & Tibshirani (2010), the EM algorithm for matrix-normal data can be computationally expensive, so they proposed a one-step approximation method. A similar technique can be applied to our normal-score type EM algorithm.

## DISCUSSION OF MODEL DIAGNOSTIC PROCEDURE AND NONPARANORMAL BIGRAPHICAL MODEL WITH $n = 1$

It is also of interest to check whether the matrix-nonparanormal model is appropriate for a specific dataset. Given $X \sim \text{MNPN}(U, V; f)$, we have $\text{cor}\{f_{ij}(X_{ij}), f_{i'j'}(X_{i'j'})\} = U_{ii'} V_{jj'}$, for any $i, i' = 1, ..., p$ and $j, j' = 1, ..., p$, which yields,

$$\text{cor}\{f_{ij}(X_{ij}), f_{i'j'}(X_{i'j'})\} = \text{cor}\{f_{ij}(X_{ij}), f_{ij'}(X_{ij'})\}\text{cor}\{f_{ij}(X_{ij}), f_{i'j}(X_{i'j})\}. \quad (17)$$

Without knowledge of the marginal transformations $f(\cdot)$, we can still estimate the correlation by the rank-based estimator. Denote by $\hat{r}_{(ij)(i'j')}$ the rank-based estimator for $\text{cor}\{f_{ij}(X_{ij}), f_{i'j'}(X_{i'j'})\}$. When the matrix-nonparanormal model holds, (17) implies that

$$\hat{r}_{(ij)(i'j')} \approx \hat{r}_{(ij)(ij')} \hat{r}_{(ij)(i'j)}. \quad (18)$$

A simple procedure for model checking is to calculate the rank-based estimator and check whether (18) holds approximately.

As pointed out by a referee, matrix-valued data may arise from applications in which $n = 1$. However, our nonparanormal model includes $pq$ unknown functions $f(\cdot)$ which are of the same dimension as $X$, when $n = 1$. Without any restriction on $f(\cdot)$, they are not estimable. A similar phenomenon is also observed by Allen & Tibshirani (2010) in the context of the Gaussian bigraphical model. One possible remedy is to assume that the functions $f(\cdot)$ are the same for

each row or column. Under this assumption, the number of unknown functions reduces to $p$ or $q$. Information can be pooled together to estimate $f(\cdot)$. It is also of interest to develop the rank-based method in this context.

## REFERENCES

ALLEN, G. I. & TIBSHIRANI, R. J. (2010). Transposable regularized covariance models with an application to missing data imputation. *Ann. Appl. Statist.* **42**, 764–790.

LAM, C. & FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37**, 4254–4278.

LIU, H., HAN, F., YUAN, M., LAFFERTY, J. D. & WASSERMAN, L. A. (2012). High dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.(Accepted)* .

LIU, H., LAFFERTY, J. D. & WASSERMAN, L. A. (2009). The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10**, 2295–2328.

ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. & ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Statist.* **2**, 494–515.

SHEPPARD, W. F. (1899). On the application of the theory of error to cases of normal distribution and normal correlation. *Philos. Trans. Roy. Soc. A* **192**, 101–531.

YIN, J. & LI, H. (2012). Model selection and estimation in the matrix normal graphical model. *J. Mult. Anal.* **107**, 119–140.

[*Received xxx* 2012. *Revised xxx* 2012]