Biometrika (2013), **100**, 1, *pp.* 139–156
© 2012 Biometrika Trust
Printed in Great Britain

doi: 10.1093/biomet/ass058
Advance Access publication 30 November 2012

Covariate-adjusted precision matrix estimation with an application in genetical genomics

By T. TONY CAI

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.

tcai@wharton.upenn.edu

HONGZHE LI

Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania 19104, U.S.A.

hongzhe@upenn.edu

WEIDONG LIU

Department of Mathematics, Institute of Natural Sciences and MOE-LSC, Shanghai Jiao Tong University, Shanghai 200240, China

liuweidong99@gmail.com

and JICHUN XIE

Department of Statistics, Fox School of Business, Temple University, Philadelphia, Pennsylvania 19122, U.S.A.

jichun@temple.edu

Summary

Motivated by analysis of genetical genomics data, we introduce a sparse high-dimensional multivariate regression model for studying conditional independence relationships among a set of genes adjusting for possible genetic effects. The precision matrix in the model specifies a covariate-adjusted Gaussian graph, which presents the conditional dependence structure of gene expression after the confounding genetic effects on gene expression are taken into account. We present a covariate-adjusted precision matrix estimation method using a constrained ℓ_1 minimization, which can be easily implemented by linear programming. Asymptotic convergence rates in various matrix norms and sign consistency are established for the estimators of the regression coefficients and the precision matrix, allowing both the number of genes and the number of the genetic variants to diverge. Simulation shows that the proposed method results in significant improvements in both precision matrix estimation and graphical structure selection when compared to the standard Gaussian graphical model assuming constant means. The proposed method is applied to yeast genetical genomics data for the identification of the gene network among a set of genes in the mitogen-activated protein kinase pathway.

Some key words: Constrained ℓ_1 penalization; Gaussian graphical model; High dimensionality; Multivariate regression.

1. Introduction

Genetical genomics experiments measure both genetic variants and gene expression data on the same subjects. Such data have provided important insights into gene expression regulations in both model organisms and humans (Brem & Kruglyak, 2005; Cheung & Spielman, 2002). For a given gene, a typical analysis of such datasets is to identify the genetic loci or single nucleotide polymorphisms that are linked or associated with the expression level of this gene. Depending on the locations of the genetic variants, they are often classified as distal trans-linked loci or proximal cis-linked loci. However, the genetic architecture for many gene expressions may be complex due to possible multiple genetic effects and gene-gene interactions, and poorly estimated genetic architecture may compromise inference on the dependency structures of genes at the transcriptional level. Although a single gene analysis can be effective in identifying the associated genetic variants, gene expressions of many genes are highly correlated due either to shared genetic variants or to other unmeasured common regulators. One important biological problem is to study the conditional independence among these genes at the expression level.

Gaussian graphical models have been applied to infer the relationship between genes at the transcriptional level (Segal et al., 2005; Li & Gui, 2006; Peng et al., 2009a), where the precision matrix for multivariate normal data has an interpretation of conditional dependence. Compared with marginal dependence, conditional dependence can capture the direct link between two variables when other variables are conditioned on. Since the expression variation of a gene can usually be explained by a small subset of other genes, the precision matrix for gene expression data is expected to be sparse. Estimation of high-dimensional Gaussian graphical models has been an active area of research in recent years. Meinshausen & Bühlmann (2006) proposed a neighbourhood selection procedure by identifying edges for each node in the graph using ℓ_1 penalized regression. This approach reduces the graphical model estimation problem to a collection of separate high-dimensional variable selection problems that have been well studied. Estimation of the precision matrix and the graphical structure can also be obtained through a penalized maximum likelihood approach; see, for example, Friedman et al. (2008), Rothman et al. (2008) and Yuan & Lin (2007). Friedman et al. (2008) proposed a fast block coordinate descent algorithm to solve the penalized likelihood maximization problem. Cai et al. (2011) proposed a constrained ℓ_1 minimization estimator for precision matrix and obtained results on convergence rates and sign consistency.

Although a direct application of the Gaussian graphical model to gene expression data alone provides some insights into gene regulation at the expression level, it ignores the effects of genetic variants on gene expression. One important observation from many genetical genomics experiments is that the gene expression level of many genes is inheritable and can be partially explained by genetic variation (Brem & Kruglyak, 2005; Cheung & Spielman, 2002). Since some genetic variants have effects on the expression of multiple genes and therefore may serve as confounders while detecting the association between the genes, ignoring the effects of genetic variants on the gene expression levels can lead to both false positive and false negative associations in the gene network graph. The effect of genetic variants on gene expression therefore needs to be adjusted in estimating the high-dimensional precision matrix.

The problem can be formulated as joint estimation of multiple regression coefficients and the precision matrix. Most of the available approaches use a groupwise regularization term where the multiple regressions can be fitted jointly (Turlach et al., 2005; Peng et al., 2009b; Obozinski et al., 2011). Rothman et al. (2010) focus on improving estimation of regression coefficients by incorporating the covariance information. Similarly, Yin & Li (2011) proposed a penalized estimation method for a sparse conditional Gaussian graphical model that iteratively estimates the regression coefficients and precision matrix. Li et al. (2012) developed a method that is based on a combination of a kernel-based estimate of the means and a regularized estimate of the precision matrix.

In this paper, we present a two-stage constrained ℓ_1 minimization approach for covariate-adjusted precision matrix estimation, where we use a constrained ℓ_1 minimization approach to first estimate the regression coefficient matrix and then estimate the precision matrix using the estimated regression coefficients in the first stage. Different from the approaches of Rothman et al. (2010) and Yin & Li (2011), our approach does not make the multivariate normal assumption on the error distribution. The method can be easily implemented by linear programming. An R package (R Development Core Team, 2012) implementing our method has been developed and is available on the CRAN website, http://cran.r-project.org/. We provide the rates of convergence and the estimation bounds for the estimates of both the regression coefficient matrix and the precision matrix in various matrix norms, allowing both the number of the covariates and the number of the response variables to diverge as the sample size approaches infinity. In addition, a simple thresholding on the estimated precision matrix is proposed to recover the support of the covariate-adjusted precision matrix and is shown to provide consistent support recovery.

2. Two-stage covariate-adjusted precision matrix estimation

2·1. *Covariate-adjusted Gaussian graphical model*

We consider a genetical genomics experiment. Let $y = (y_1, \ldots, y_p)^{\mathrm{T}}$ denote the random vector of expression levels for p genes, and let $x = (x_1, \ldots, x_q)^{\mathrm{T}}$ denote the random vector of the numerical values of q genetic markers. We consider the multivariate regression model

$$y = \Gamma_0 x + z, \tag{1}$$

where Γ_0 is a $p \times q$ unknown coefficient matrix, z is a $p \times 1$ random vector with mean zero, covariance matrix $\Sigma_0 = (\sigma_{ij}^0)$ and precision matrix $\Omega_0 = (\omega_{ij}^0) = \Sigma_0^{-1}$. We assume that x and z are independent and that we have n independent identically distributed observations (x_k, y_k) $(k = 1, \ldots, n)$ from (1).

In genetical genomics data, each row of Γ_0 is assumed to be sparse, since each gene is expected to have only a few genetic regulators. The precision matrix Ω_0 is also expected to be sparse, since typical genetic networks have limited links. If z follows a multivariate normal distribution, the conditional independence of z_i and z_j is equivalent to $\omega_{ij} = 0$, and the matrix Ω_0 has an interpretation of conditional dependence and can be used to construct a conditional dependence graph. To be more specific, let $G = (V, E)$ be a graph representing conditional independence relations between the components of y. The vertex set V has p components $y_1, \ldots, y_p$ and the edge set E consists of pairs (i, j), where $(i, j) \in E$ if there is an edge between y_i and y_j. The edge between y_i and y_j is excluded from E if and only if z_i and z_j are independent given all other z_k $(k \neq i, j)$. We are interested in detecting the nonzero entries of Ω_0 in order to construct a conditional independence graph for y after the effects of the covariates x on y are removed. Such a graphical model is called the covariate-adjusted Gaussian graphical model.

Estimation of Γ_0 in (1) in high-dimensional settings, where p and q can be larger than n, has been extensively studied. Most of the available approaches use a groupwise regularization term where the p regressions can be fitted jointly (Turlach et al., 2005; Peng et al., 2009b; Obozinski et al., 2011). Rothman et al. (2010) and Yin & Li (2011) developed ℓ_1-penalized estimation methods that iteratively estimate Γ_0 and Ω_0. Rothman et al. (2010) focus on improving estimation of Γ_0 by incorporating Ω_0. The work of Yin & Li (2011) aims to improve the estimate of Ω_0 after the effects of the covariates on the means are taken into account, and this is also the focus of the present paper.

2·2. *Estimation of* Γ_0

When $q = 1$, many methods have been developed for estimation of Γ_0, including those based on ℓ_1 minimization (Tibshirani, 1996) and the Dantzig selector (Candès & Tao, 2007). We propose to develop a method for estimating Γ_0 using a constrained ℓ_1 minimization that can be treated as a multivariate extension of the Dantzig selector. For a matrix $A = (a_{ij}) \in \mathbb{R}^{p\times q}$, define the elementwise ℓ_1 norm by $|A|_1 = \sum_{i=1}^{p}\sum_{j=1}^{q}|a_{ij}|$ and the elementwise ℓ_∞ norm by $|A|_\infty = \max_{i,j}|a_{ij}|$.

Let $\bar{y} = n^{-1}\sum_{k=1}^{n} y_k$, $\bar{x} = n^{-1}\sum_{k=1}^{n} x_k$ and $\bar{z} = n^{-1}\sum_{k=1}^{n} z_k$. Then

$$y_k - \bar{y} = \Gamma_0(x_k - \bar{x}) + z_k - \bar{z}. \tag{2}$$

Set $S_{xy} = n^{-1}\sum_{k=1}^{n}(y_k - \bar{y})(x_k - \bar{x})^{\mathrm{T}}$ and $S_{xx} = n^{-1}\sum_{k=1}^{n}(x_k - \bar{x})(x_k - \bar{x})^{\mathrm{T}}$. We propose to estimate Γ_0 by the solution to the optimization problem

$$\hat{\Gamma} \in \mathop{\arg\min}_{\Gamma\in R^{p\times q}}\{|\Gamma|_1 : |S_{xy} - \Gamma S_{xx}|_\infty \leqslant \lambda_n\}, \tag{3}$$

where λ_n is a tuning parameter. This is equivalent to the p optimization problems

$$\min|\gamma_i|_1, \text{ subject to } |S_{xy,i} - \gamma_i^{\mathrm{T}} S_{xx}|_\infty \leqslant \lambda_n \quad (i = 1, \ldots, p), \tag{4}$$

where $\Gamma = (\gamma_1, \ldots, \gamma_p)^{\mathrm{T}}$ and $S_{xy} = (S_{xy,1}, \ldots, S_{xy,p})^{\mathrm{T}}$. This is exactly the Dantzig selector formulation in the usual regression analysis for the ith regression and its solution can therefore be obtained by solving the corresponding linear programming problem. This simple observation is useful for the implementation and technical analysis. In this paper and the R package we developed, the procedure is implemented by a linear programming optimization using the primal dual and interior point algorithm.

2·3. *Estimation of* Ω_0

After inserting the estimator $\hat{\Gamma}$ given in (3) into equation (2), we can estimate Ω_0 by the method of constrained ℓ_1-minimization proposed in Cai et al. (2011). Let

$$S_{yy} = \frac{1}{n}\sum_{k=1}^{n}(y_k - \hat{\Gamma}x_k)(y_k - \hat{\Gamma}x_k)^{\mathrm{T}}.$$

The precision matrix Ω_0 is then estimated by the solution to the optimization problem

$$\hat{\Omega}_1 \in \mathop{\arg\min}_{\Omega\in R^{p\times p}}\{|\Omega|_1 : \ |I_{p\times p} - S_{yy}\Omega|_\infty \leqslant \tau_n\}, \tag{5}$$

where τ_n is a tuning parameter. Let $\hat{\Omega}_1 = (\hat{\omega}_{ij}^1)$ be a solution to (5). This constrained ℓ_1 minimization approach is the same as the one proposed in Cai et al. (2011), except that S_{yy} depends on the estimated coefficient matrix $\hat{\Gamma}$. Since no symmetry condition on $\hat{\Omega}_1$ is imposed, as a result, the solution may not be symmetrical in general. The final estimator of Ω_0, denoted by $\hat{\Omega} = (\hat{\omega}_{ij})$, is obtained by symmetrizing the estimator as follows:

$$\hat{\Omega} = (\hat{\omega}_{ij}), \text{ where } \hat{\omega}_{ij} = \hat{\omega}_{ji} = \hat{\omega}^1_{ij} I(|\hat{\omega}^1_{ij}| \leqslant |\hat{\omega}^1_{ji}|) + \hat{\omega}^1_{ji} I(|\hat{\omega}^1_{ij}| > |\hat{\omega}^1_{ji}|), \tag{6}$$

where $I(\cdot)$ is the indicator function. As in (4), the problem (5) can be decomposed into p optimization problems. For $i = 1, \ldots, p$, let $\hat{\omega}_i$ be the solution of the convex optimization problem

$$\min |\omega_i|_1 \text{ subject to } |e_i - S_{yy}\omega_i|_\infty \leqslant \tau_n,$$

where ω_i is a vector in $\mathbb{R}^p$, e_i is a standard unit vector in $\mathbb{R}^p$ with 1 in the ith coordinate and 0 in all other coordinates. This can also be solved using the primal dual and interior point algorithm.

2·4. *Tuning parameter selection*

Two tuning parameters λ_n and τ_n need to be selected. We tune these two parameters together via L-fold crossvalidation, where the Bregman divergence can be use to measure the model fit. Specifically, we divide all the samples in the training dataset into L disjoint subgroups, also known as folds, and denote the index of subjects in the lth fold by T_l for $l = 1, \ldots, L$. The L-fold crossvalidation score is defined as

$$\text{CV}(\lambda_n, \tau_n) = \sum_{l=1}^{L} [\log\det\{\hat{\Omega}_{-l}(\lambda_n, \tau_n)\} - \text{tr}\{S_{yyl}\hat{\Omega}_{-l}(\lambda_n, \tau_n)\}],$$

where n_l is the size of the lth fold T_l and

$$S_{yyl} = n_l^{-1} \sum_{k=1}^{n_l} \{y_k - \hat{\Gamma}_{-l}(\lambda_n)x_k\}\{y_k - \hat{\Gamma}_{-l}(\lambda_n)x_k\}^{\mathrm{T}}$$

with $\hat{\Omega}_{-l}(\lambda_n, \tau_n)$ and $\hat{\Gamma}_{-l}(\lambda_n)$ being the estimates of Ω and Γ based on the sample $(\bigcup_{l=1}^{L} T_l) \backslash T_l$ with λ_n and τ_n as the tuning parameters. Then, we choose $(\lambda_n^*, \tau_n^*) = \text{argmax CV}(\lambda_n, \tau_n)$ as the best tuning parameters, which are used to obtain the final estimates of the regression coefficients and precision matrix based on the whole training set. Here the maximization of $\text{CV}(\lambda_n, \tau_n)$ with respect to (λ_n, τ_n) is achieved via a grid search.

3. Rates of convergence of the estimators

3·1. *Convergence rates of* $\hat{\Gamma} - \Gamma_0$

In this section, we present theoretical properties of the estimators $\hat{\Gamma}$ and $\hat{\Omega}$. We first introduce the matrix norms used in the rest of the paper. For a matrix $A = (a_{ij}) \in \mathbb{R}^{p\times q}$, define the spectral norm as $\|A\|_2 = \max_{|x|_2=1} |Ax|_2$, the matrix L_∞ norm as $\|A\|_{L_\infty} = \max_{1\leqslant i\leqslant p} \sum_{j=1}^{q} |a_{ij}|$, and the Frobenius norm as $\|A\|_F = (\sum_{i,j} a_{ij}^2)^{1/2}$. The notation $A \succ 0$ means that A is positive definite. Write $x = (x_1, \ldots, x_q)^{\mathrm{T}}$, $z = (z_1, \ldots, z_p)^{\mathrm{T}}$ and $u = z^{\mathrm{T}}\Omega_0 = (u_1, \ldots, u_p)$. The following conditions are needed for establishing the rate of convergence.

Condition 1. Let $\log(p \vee q) = o(n)$. Suppose that there exist some $\eta > 0$ and $K > 0$ such that

$$E\{\exp(\eta x_i^2)\} \leqslant K, \quad E\{\exp(\eta z_j^2/\sigma^0_{jj})\} \leqslant K, \quad E\{\exp(\eta u_j^2/\omega^0_{jj})\} \leqslant K,$$

for all $i = 1, \ldots, q$ and $j = 1, \ldots, p$, and let $\max_{1\leqslant j\leqslant p} \sigma^0_{jj} \leqslant K$.

Condition 2. The regression coefficient matrix Γ_0 belongs to the following class with $0 \leqslant \delta_1 < 1$:

$$\mathcal{V}_{\delta_1} = \mathcal{V}_{\delta_1}\{s_1(q)\} = \left\{ \Gamma \in \mathbb{R}^{p\times q} : \max_{1\leqslant i\leqslant p} \sum_{j=1}^{q} |\gamma_{ij}|^{\delta_1} \leqslant s_1(q) \right\}.$$

Condition 3. The precision matrix $\Omega_0 = (\omega_{ij}^0)_{p\times q}$ belongs to the following class with $0 \leqslant \delta_2 < 1$:

$$\mathcal{U}_{\delta_2} = \mathcal{U}_{\delta_2}\{s_2(p)\} = \left\{ \Omega \succ 0 : \|\Omega\|_{L_\infty} \leqslant M_p, \quad \max_{1\leqslant i\leqslant p} \sum_{j=1}^{p} |\omega_{ij}|^{\delta_2} \leqslant s_2(p), \right.$$
$$\left. \lambda_{\max}(\Omega)/\lambda_{\min}(\Omega) \leqslant C_0 \right\}.$$

Condition 4. There exists some $N_q > 0$ such that the matrix l_∞ norm of Σ_x^{-1} satisfies $\|\Sigma_x^{-1}\|_{L_\infty} \leqslant N_q$, where $\Sigma_x = \text{cov}(x)$.

Condition 1 is a sub-Gaussian condition on x, z and $z^{\mathrm{T}}\Omega_0$, where the variance of u_j is ω_{jj}^0. The dimensions p and q are of order $\exp\{o(n)\}$. Conditions 2 and 3 assume the uniformity class of matrices for the regression coefficient matrix and the precision matrix, where $\mathcal{V}_0$ and $\mathcal{U}_0$ are classes of matrices with the sparsity measurements of $s_1(q)$ and $s_2(p)$, respectively. Similar parameter spaces have also been used in Bickel & Levina (2008) and Cai et al. (2011). Conditions 2 and 3 also bound the matrix L_∞ norm of Γ_0 and Ω_0. Finally, Condition 4 bounds the matrix L_∞ norm of the inverse covariance matrix of x.

The estimation error $\hat{\Gamma} - \Gamma_0$ can be measured by different matrix norms: the matrix L_∞ norm, the Frobenius norm and the entry-wise ℓ_∞ norm. The matrix L_∞ norm measures the accuracy of the estimation of Γ_0. The Frobenius norm is also a reasonable measure on the accuracy of the estimation of Γ_0 and can be viewed as the sum of squared errors for estimating individual rows. The elementwise ℓ_∞ norm can be used to recover the support of Γ_0 by a further thresholding step. We have the following rates of convergence for the estimator $\hat{\Gamma}$ in matrix L_∞ and the Frobenius norm.

Theorem 1. *Suppose Conditions* 1, 2 *and* 4 *hold. Let* $\Gamma_0 \in \mathcal{V}_{\delta_1}$ *and* $\lambda_n = C_1[\{\log(pq)\}/n]^{1/2}$, *where* $C_1 > 0$ *is a sufficiently large constant. If*

$$s_1(q) = o\left[N_q^{\delta_1 - 1} \left\{ \frac{n}{\log(pq)} \right\}^{(1-\delta_1)/2} \right], \tag{7}$$

then with probability greater than $1 - O\{(pq)^{-1}\}$, *we have*

$$\|\hat{\Gamma} - \Gamma_0\|_{L_\infty} \leqslant C N_q^{1-\delta_1} s_1(q) \left\{ \frac{\log(pq)}{n} \right\}^{(1-\delta_1)/2} \tag{8}$$

and

$$\frac{1}{p} \|\hat{\Gamma} - \Gamma_0\|_F^2 \leqslant C N_q^{2-\delta_1} s_1(q) \left\{ \frac{\log(pq)}{n} \right\}^{1-\delta_1/2} \tag{9}$$

for some constant $C > 0$.

Theorem 1 shows that the regression coefficients matrix Γ_0 can be estimated consistently under the Frobenius norm if the sparsity $s_1(q)$ of Γ_0 is of order $o[N_q^{\delta_1-2}\{n/\log(pq)\}^{1-\delta_1/2}]$. The requirement on the dimensions p and q is mild as they appear only in the logarithmic term. To see this, if $s_1(q) = O(n^{r_1})$ for some $r_1 < 1 - \delta_1/2$ and N_q is bounded, then p and q can be as large as $\exp(n^{r_2})$ for some $r_2 < 1 - \delta_1/2 - r_1$.

THEOREM 2. *Under the conditions of Theorem* 1, *with probability greater than* $1 - O\{(pq)^{-1}\}$, *we have*

$$|\hat{\Gamma} - \Gamma_0|_\infty \leqslant C_0 N_q \left\{\frac{\log(pq)}{n}\right\}^{1/2} \tag{10}$$

for some constant $C_0 > 0$.

The rate under the elementwise l_∞ norm is critical to the support recovery. Define $\tilde{\Gamma}_{\text{thr}} = (\tilde{\gamma}_{ij})$ with

$$\tilde{\gamma}_{ij} = \hat{\gamma}_{ij} I\left[|\hat{\gamma}_{ij}| \geqslant C_0 N_q \left\{\frac{\log(pq)}{n}\right\}^{1/2}\right],$$

where $(\hat{\gamma}_{ij}) = \hat{\Gamma}$. Let $S(\Gamma_0) = \{(i, j) : \gamma_{ij}^0 \neq 0\}$ be the true support of the coefficient matrix Γ_0 and $\gamma_{\min} = \min_{(i,j)\in S(\Gamma_0)} |\gamma_{ij}|$.

THEOREM 3. *Suppose the conditions in Theorem* 1 *hold and*

$$\gamma_{\min} \geqslant 2C_0 N_q \left\{\frac{\log(pq)}{n}\right\}^{1/2}. \tag{11}$$

Then with probability greater than $1 - O\{(pq)^{-1}\}$, *we have* $S(\tilde{\Gamma}_{\text{thr}}) = S(\Gamma_0)$.

The lower bound condition (11) requires that the magnitude of the nonzero entries in Γ_0 cannot be too small in order to recover the support.

3·2. *Convergence rates of* $\hat{\Omega} - \Omega_0$

We consider the rate of $\hat{\Omega} - \Omega_0$ under the spectral norm and the elementwise l_∞ norm. The rate under the spectral norm is important because it can lead to the consistency of the estimation of eigenvalues and eigenvectors and it is essentially needed in developing theoretical properties for various statistical inference problems when the estimator of the precision matrix is used.

THEOREM 4. *Suppose Conditions* 1–4 *and* (7) *hold. Let* $\Gamma_0 \in \mathcal{V}_{\delta_1}$, $\Omega_0 \in \mathcal{U}_{\delta_2}$ *and*

$$s_1(q) \leqslant C(1 + M_p)^{-1} N_q^{-2+\delta_1} \left\{\frac{n}{\log(pq)}\right\}^{(1-\delta_1)/2}. \tag{12}$$

Let $\tau_n = C_2[\{\log(pq)\}/n]^{1/2}$, *where* $C_2 > 0$ *is a sufficiently large constant. Then with probability greater than* $1 - O\{(pq)^{-1}\}$, *we have*

$$\|\hat{\Omega} - \Omega_0\|_2 \leqslant C M_p^{1-\delta_2} s_2(p) \left\{\frac{\log(pq)}{n}\right\}^{(1-\delta_2)/2} \tag{13}$$

for some constant $C > 0$.

The condition (12) on the sparsity $s_1(q)$ of Γ_0 ensures that Γ_0 can be well estimated with a certain rate so that $y - \hat{\Gamma}_0 x$ can be used to replace $y - \Gamma_0 x$. The convergence rate in (13) is optimal. In fact, as shown in an unpublished 2010 technical report available from the first author, even if $\Gamma_0 = 0$ or is known in advance, the minimax optimal rate of estimation of Ω_0 is still $O\{M_p^{1-\delta_2} s_2(p)(\log p/n)^{(1-\delta_2)/2}\}$. If $q = O(p)$, then the rate in (13) is the same as the oracle optimal rate and thus is also optimal.

The next theorem shows the convergence rate under the elementwise l_∞ norm, which is useful for the recovery of the support of Ω.

THEOREM 5. *If Conditions* 1–4 *and* (7) *hold, we have with probability greater than* $1 - O\{(pq)^{-1}\}$ *that,*

$$|\hat{\Omega} - \Omega_0|_\infty \leqslant C M_p \left\{ \frac{\log(pq)}{n} \right\}^{1/2}, \tag{14}$$

where $C > 0$ *is a constant.*

The proofs of Theorems 4 and 5 are given in the Appendix. The key is to account for the estimation error and uncertainty of $\hat{\Gamma}_0$ in evaluating the estimation error of $\hat{\Omega}_0$. This is in contrast to the estimation of Ω_0 in Cai et al. (2011) when Γ_0 is assumed to be zero. As shown in an unpublished 2010 technical report available from the first author, the minimax optimal rate under the elementwise l_∞ norm for estimating the precision matrix is $O\{M_p(\log p/n)^{1/2}\}$ when $\Gamma_0 = 0$ or is known. Hence covariate-adjusted ℓ_1 minimization can achieve the same optimal rate as the case that Γ_0 is known.

4. GRAPHICAL MODEL SELECTION CONSISTENCY

When the error term z in (1) is assumed to follow $N(0, \Omega_0^{-1})$, recovery of the support of the precision matrix Ω_0 is closely related to the covariate-adjusted graphical model selection. When $\Gamma_0 = 0$, the problem reduces to Gaussian graphical model selection. We consider the setting when Ω_0 belongs to $\mathcal{U}_0$ and are interested in estimating the support of Ω_0, $S(\Omega_0) = \{(i, j) : \omega_{ij}^0 \neq 0\}$ when $\Gamma_0 \neq 0$. Define $\theta_{\min} = \min_{(i,j)\in S(\Omega_0)} |\omega_{ij}^0|$. As long as $\theta_{\min} \geqslant 2M_p\tau_n$, using the rate under the elementwise ℓ_∞ norm given in Theorem 5, we have following result.

THEOREM 6. *Suppose Conditions* 1–4 *and* (7) *hold. Further suppose that* $\theta_{\min} > 2M_p\tau_n$. *Then for all* $\omega_{ij} \neq 0$, *the probability of* $\hat{\omega}_{ij} \neq 0$ *tends to one.*

Sign consistency can be achieved by further thresholding the entries of $\hat{\Omega}$. Let

$$\hat{\Omega}_r = (\hat{\omega}_{ij}^r), \quad \hat{\omega}_{ij}^r = \hat{\omega}_{ij} I(|\hat{\omega}_{ij}| \geqslant \tau_n'),$$

where τ_n' is a tuning parameter that will be specified later. Define $\Psi = \{\mathrm{sign}(\omega_{ij}^0) : i = 1, \ldots, p,\ j = 1, \ldots, p\}$ and let $\hat{\Psi} = \{\mathrm{sign}(\hat{\omega}_{ij}^r) : i = 1, \ldots, p,\ j = 1, \ldots, p\}$ be the vector of the signs of the elements of the true and the estimated precision matrix, where $\mathrm{sign}(t)$ is defined as

$$\mathrm{sign}(t) = \begin{cases} 1, & t > 0, \\ 0, & t = 0, \\ -1, & t < 0. \end{cases}$$

Table 1. *Four models and the parameters used in the simulations*

	Parameters
Model 1	$p = 60$, $q = 30$, $n = 100$, $\text{pr}(\Gamma_{ij} \neq 0) = 5/q$, $\text{pr}(\Omega_{ij} \neq 0 \mid i \neq j) = 5/p$
Model 2	$p = 200$, $q = 200$, $n = 200$, $\text{pr}(\Gamma_{ij} \neq 0) = 30/q$, $\text{pr}(\Omega_{ij} \neq 0 \mid i \neq j) = 5/p$
Model 3	$p = 200$, $q = 200$, $n = 100$, $\text{pr}(\Gamma_{ij} \neq 0) = 30/q$, $\text{pr}(\Omega_{ij} \neq 0 \mid i \neq j) = 5/p$
Model 4	$p = 800$, $q = 300$, $n = 200$, $\text{pr}(\Gamma_{ij} \neq 0) = 30/q$, $\text{pr}(\Omega_{ij} \neq 0 \mid i \neq j) = 10/p$

We have the following theorem on sign consistency of the estimator $\hat{\Psi}$, i.e., the estimator recovers not only the sparsity pattern of Ω_0, but also the signs of the nonzero elements.

THEOREM 7. *Let $\tau'_n = 4M_p\tau_n$. Suppose that $\theta_{\min} > 2\tau'_n$. Then under the conditions of Theorem 4, as n and p tend to infinity, $\hat{\Psi} = \Psi$ with probability tending to one.*

Theorem 7 shows that the support of Ω_0 can be recovered exactly if the minimum of the nonzero entries in Ω_0 has a lower bound that is not too small. The lower bound condition is necessary in order to recover the support exactly. In fact, as shown in an unpublished 2010 technical report available from the first author, suppose that $\Gamma_0 = 0$ or is known in advance, if $\theta_{\min} \leqslant c\tau'_n$ for a sufficiently small constant $c > 0$, then for any estimator of Ω_0, it is not possible to recover the support exactly uniformly over the class of $s_2(p)$ sparse precision matrices.

In practice, since the estimator obtained from (6) is already sparse, we do not further threshold the estimator. Although the sign consistency cannot be guaranteed, under weaker conditions, we can still get an estimator with its properties stated in Theorem 6.

5. SIMULATION RESULTS

In this section simulation studies are carried out to evaluate the performance of the proposed procedure and to compare it with other methods for precision matrix estimation and support recovery. Four models presented in Table 1 are considered. For each model, we generate a $p \times q$ coefficient matrix Γ and a $p \times p$ precision matrix Ω with $\text{pr}(\Gamma_{ij} \neq 0)$ and $\text{pr}(\Omega_{ij} \neq 0 \mid i \neq j)$ shown in Table 1. If $\Gamma_{ij} \neq 0$ or $\Omega_{ij} \neq 0$ $(i \neq j)$, we generate Γ_{ij} or Ω_{ij} $(i \neq j)$ from $\text{Unif}([0{\cdot}5, 1] \cup [1, 0{\cdot}5])$. The diagonal of Ω is set to be a common value so that the condition number of Ω is equal to p. This is to make sure that Ω is positive definite and invertible. Let $\Sigma = \Omega^{-1}$. We generate $n \times q$ design matrix X and an $n \times p$ random error matrix so that X_{ij} and Z_{ij} independently follow $N(0, 1)$ distributions. The $n \times p$ outcome matrix is set to be $Y = X\Gamma + Z\Sigma^{1/2}$.

Model 1 has small values of p and q and is considered to mimic the applications on finding small-scale gene regulatory pathways or constructing networks in social sciences. Models 2–4 have moderate or large p and q, simulating the settings in most genomic applications.

The performance of our proposed method is compared with several other methods, including those of Cai et al. (2011) and Friedman et al. (2008) that ignore the covariate effects and that of Yin & Li (2011). For all these estimators, the tuning parameters are chosen using five-fold crossvalidation by maximizing the crossvalidated log-likelihood function,

$$\log \det(\Omega) - \text{tr}(S_{yy}\Omega),$$

where $S_{yy} = n^{-1} \sum_{k=1}^{n} (y_k - \bar{y})(y_k - \bar{y})^{\mathrm{T}}$ for the methods of Cai et al. (2011) and Friedman et al. (2008), and $S_{yy} = n^{-1} \sum_{k=1}^{n} (y_k - \hat{\Gamma} x_k)(y_k - \hat{\Gamma} x_k)^{\mathrm{T}}$ for our method and

Table 2. *Simulation results: estimation errors of four different methods for the precision matrix as measured by different matrix norms based on* 50 *replications. Numbers in parentheses are the simulation standard errors*

	(p, q, n)	Method	Spectral norm	Frobenius norm	Matrix ℓ_1 norm
Model 1	(60, 30, 100)	CAPME	4·4 (0·2)	15·8 (0·2)	9·6 (0·4)
		CLIME	4·7 (0·1)	16·2 (0·1)	11·2 (0·4)
		cGGM	3·1 (0·2)	13·4 (0·1)	7·7 (0·5)
		GLASSO	5·6 (0·1)	16·9 (0·0)	12·1 (0·2)
Model 2	(200, 200, 200)	CAPME	10·5 (0·3)	30·2 (0·1)	24·8 (0·7)
		CLIME	13·1 (0·0)	34·0 (0·0)	29·4 (0·1)
		cGGM	11·4 (0·2)	33·0 (0·2)	26·4 (0·6)
		GLASSO	6·9 (0·2)	41·0 (0·0)	13·9 (0·0)
Model 3	(200, 200, 100)	CAPME	8·2 (0·5)	48·5 (1·7)	26·6 (1·8)
		CLIME	8·8 (0·1)	48·8 (0·1)	19·4 (0·2)
		cGGM	11·0 (5·2)	54·8 (3·2)	26·4 (0·6)
		GLASSO	9·6 (0·0)	50·0 (0·0)	20·1 (0·0)
Model 4	(800, 300, 200)	CAPME	14·2 (0·1)	69·5 (0·1)	31·6 (0·4)
		CLIME	10·8 (0·6)	111·5 (2·5)	37·8 (0·9)
		cGGM	14·4 (0·3)	69·1 (0·3)	37·3 (5·6)
		GLASSO	15·4 (0·0)	82·4 (0·0)	34·2 (0·1)

CAPME, ℓ_1 constrained minimization adjusted for covariates; CLIME, the method of Cai et al. (2011); cGGM, the method of Yin & Li (2011); GLASSO, the method of Friedman et al. (2008).

Table 3. *Simulation results: variable selection performances as measured by overall error rate, sensitivity, specificity and the Matthews correlation coefficient, for four different procedures, based on* 50 *replications. Numbers in parentheses are the simulation standard errors. All the values are multiplied by* 100

Model	(p, q, n)	Method	MISR	SPE	SEN	MCC
Model 1	(60, 30, 100)	CAPME	17 (0)	89 (1)	58 (3)	45 (3)
		CLIME	29 (0)	77 (1)	37 (2)	12 (2)
		cGGM	17 (0)	87 (1)	61 (2)	44 (2)
		GLASSO	30 (0)	75 (1)	42 (3)	13 (2)
Model 2	(200, 200, 200)	CAPME	6 (0)	97 (0)	36 (2)	35 (1)
		CLIME	9 (0)	95 (0)	7 (1)	2 (1)
		cGGM	9 (0)	94 (0)	38 (2)	24 (0)
		GLASSO	20 (0)	83 (0)	19 (1)	1 (1)
Model 3	(200, 200, 100)	CAPME	16 (0)	87 (0)	19 (1)	4 (1)
		CLIME	16 (0)	95 (0)	5 (1)	1 (1)
		cGGM	37 (0)	65 (1)	4 (2)	1 (1)
		GLASSO	12 (0)	93 (0)	8 (1)	1 (1)
Model 4	(800, 300, 200)	CAPME	3 (0)	1 (0)	8 (0)	1 (1)
		CLIME	12 (0)	90 (0)	12 (0)	1 (0)
		cGGM	3 (0)	99 (0)	3 (1)	4 (1)
		GLASSO	32 (0)	69(0)	33 (0)	1 (0)

CAPME, ℓ_1 constrained minimization adjusted for covariates; CLIME, the method of Cai et al. (2011); cGGM, the method of Yin & Li (2011); GLASSO, the method of Friedman et al. (2008); MISR, misspecification rate; SEN, sensitivity; SPE, specificity; MCC, Matthews correlation coefficient.

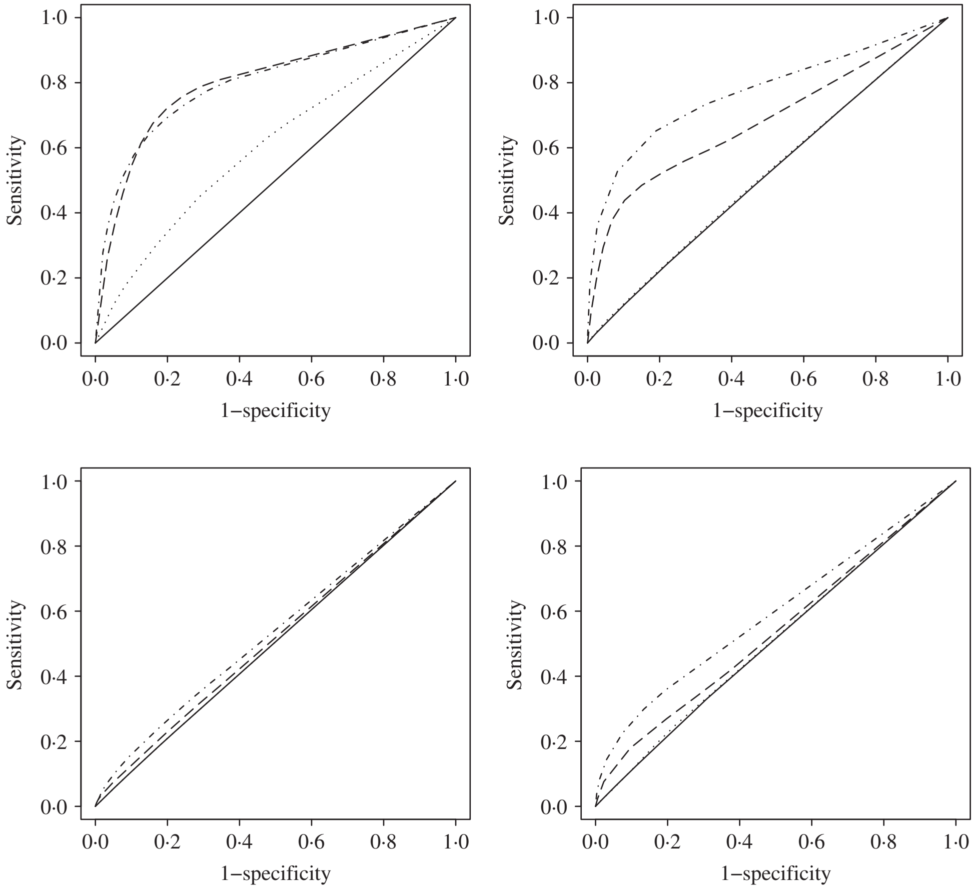

Fig. 1. The average receiver operating characteristic curves obtained by varying the tuning parameter τ_n. The upper left panel is for Model 1, the upper right panel is for Model 2, the bottom left panel is for Model 3 and the bottom right panel is for Model 4. The solid, dotted, dashed and dashed-dotted curves represent the methods of Friedman et al. (2008), Cai et al. (2011), Yin & Li (2011) and our method, respectively. The solid and the dotted curves overlap in the bottom plots.

that of Yin & Li (2011), with $\hat{\Gamma}$ computed from the training dataset. The final estimates are obtained using the chosen tuning parameters on the full datasets. No extra thresholding is applied to the estimators.

Several different measures are used to compare the performance of these estimators. The estimation error $\hat{\Omega} - \Omega$ is evaluated in terms of the spectral norm, Frobenius norm and ℓ_1 norm. The graph structure recovery is evaluated by the misspecification rate, specificity, sensitivity and Matthews correlation coefficient, which are defined as:

$$\textsc{misr}(\Omega_0, \hat{\Omega}) = \frac{\textsc{fn} + \textsc{fp}}{p(p-1)}, \quad \textsc{spe} = \frac{\textsc{tn}}{\textsc{tn} + \textsc{fp}}, \quad \textsc{sen} = \frac{\textsc{tp}}{\textsc{tp} + \textsc{fn}},$$

$$\textsc{mcc} = \frac{\textsc{tp} \times \textsc{tn} - \textsc{fp} \times \textsc{fn}}{\{(\textsc{tp} + \textsc{fp})(\textsc{tp} + \textsc{fn})(\textsc{tn} + \textsc{fp})(\textsc{tn} + \textsc{fn})\}^{1/2}}.$$

Here, TP, TN, FP, FN are the numbers of true positives, true negatives, false positives and false negatives, respectively, where true positives are the nonzero entries of the nondiagonal elements of Ω. The performances over 50 replications are reported in Tables 2 and 3.

For the estimation error, see Table 2, when $\log(pq)/n$ is small or moderate as in Models 1 and 2, the performance of our method is comparable to that of the method in Yin & Li (2011). As $\log(pq)/n$ increases, the proposed estimator has the smallest estimation errors. In terms of graph structure recovery, see Table 3, adjusting for covariates yields better performance in general as shown by the proposed method and the method of Yin & Li (2011). Our procedure performs better than the other methods for Models 2–4, and it has a performance comparable to the method of Yin & Li (2011) for Model 1.

The results presented in Tables 2 and 3 depend on the tuning parameters, which are selected by five-fold crossvalidation for all the estimators. To further compare the performance on graph structure recovery, we obtain the receiver operating characteristic curve for each simulated dataset by varying the turning parameter τ_n. The tuning parameter for the regression coefficients, λ_n, for our method and that of Yin & Li (2011) is fixed at the value selected by the crossvalidation. Figure 1 shows the receiver operating characteristic curves averaged over 50 replications. Our method has a comparable performance with that of Yin & Li (2011) for Model 1 and has better performance in the other models. Figure 1 also demonstrates that without adjusting for the covariate effects, existing precision matrix estimation methods perform poorly in terms of support recovery. The value of $\log(pq)/n$ is the key factor that determines the performance of these methods. When it is large as in Model 3, all the methods perform rather poorly. In Model 4, the dimension of the parameters $p^2 + pq$ is eleven times that of Model 3 and the sample size is only twice as large. However, since Model 4 has a smaller $\log(pq)/n$ ratio, all methods have better performance than for Model 3.

6. Analysis of yeast data

We illustrate our method using the yeast genetical genomics data set generated by Brem & Kruglyak (2005). The dataset contains 112 yeast segregants grown from a cross involving BY4716 and wild isolate RM11-1a. The RNA was isolated and cDNA was hybridized to microarrays with 6216 yeast genes assayed on each array. Each of the 112 segregants were individually genotyped at 2956 marker positions. Due to the small sample size and limited perturbation to the biological system, it is not possible to construct a gene network for all 6216 genes. We instead focused our analysis on two sets of genes that are biologically relevant: the first set of 54 genes that belong to the yeast mitogen-activated protein kinase signalling pathway provided by the Kyoto Encyclopedia of Genes and Genomes database (Kanehisa et al., 2010), another set of 1207 genes of the protein-protein interaction network obtained from a previously compiled set by Steffen et al. (2002) combined with protein physical interactions deposited in the Munich Information center for Protein Sequences (Mewes et al., 2002).

The first set of genes includes 54 genes that belong to the yeast mitogen-activated protein kinase signalling pathway. Figure 2 displays the illustrative pathway structure, showing how yeast genes respond to different cellular signals. Some gene nodes such as Ste20, Ste11 and Ste7 appear in multiple locations on this pathway. This directed signalling pathway explains how yeast cells respond to different cellular signals.

To apply our method, we first select the genetic markers based on simple screening. There are 188 markers that are marginally associated with at least two of the 54 genes with a p-value less than or equal to 0·01. A total of 702 such associations are observed, suggesting there is a large pool of possible confounders. We apply our method to this set of 54 genes and 188 markers and use five-fold crossvalidation to choose the tuning parameters as $\lambda = 0{\cdot}15$ and $\tau = 0{\cdot}24$. The covariate-adjusted estimation results in selecting 51 links among the 54 genes. In addition, the method identifies 597 nonzero entries for the coefficient matrix, indicating that many gene

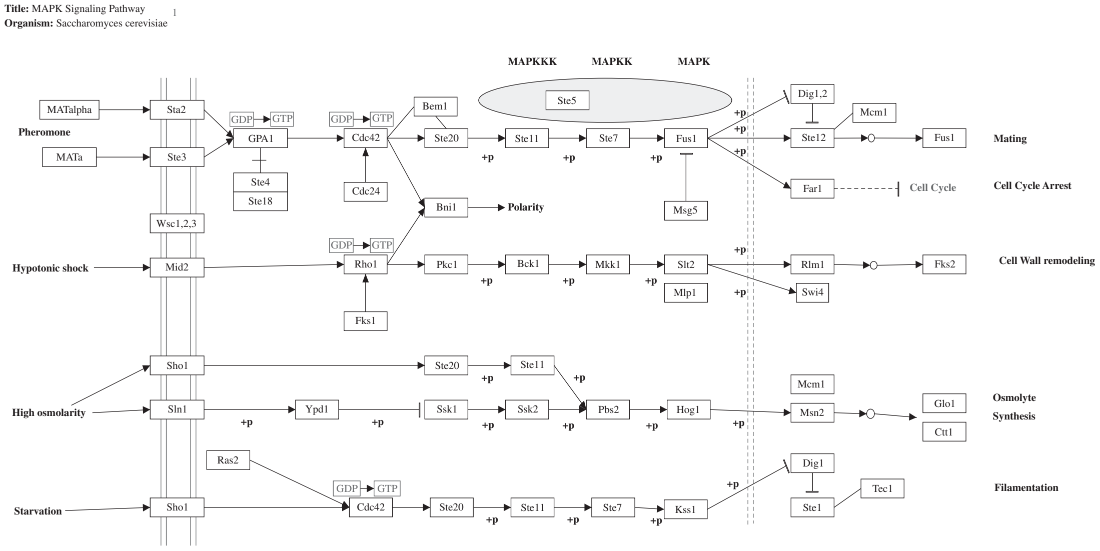

Fig. 2. The yeast mitogen-activated protein kinase signalling pathway, illustrating the signalling paths in responses to different signals. Some genes appear in multiple paths. The figure is downloaded from http://www.wikipathways.org/index.php/Pathway:WP510 (Kelder et al., 2012).

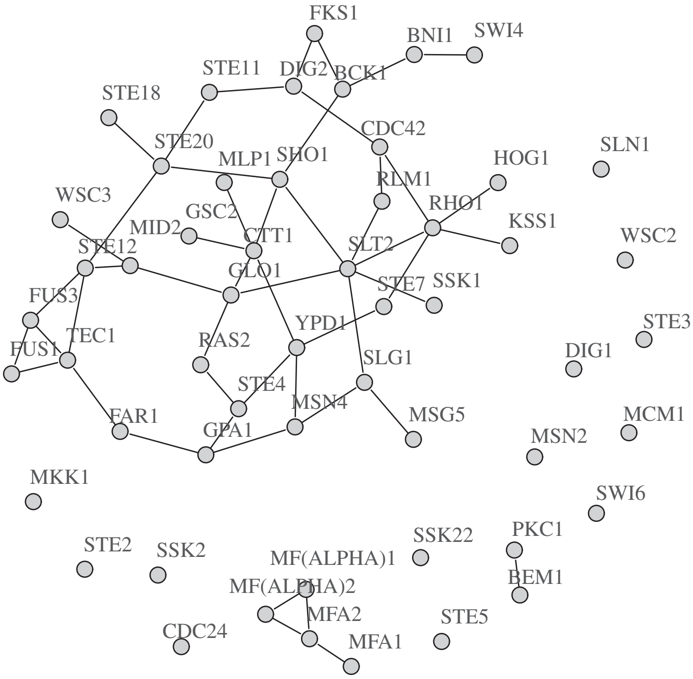

Fig. 3. Covariate-adjusted conditional independence graph constructed based on the estimated covariate-adjusted precision matrix for the 54 genes on the yeast mitogen-activated protein kinase signalling pathway.

expression levels are affected by genetic variants. There are 528 pairs of genes sharing at least one common genetic variant. Figure 3 shows the graph constructed by our method based on the estimated precision matrix. While we do not expect that the estimated conditional Gaussian graph can fully recover the true mitogen-activated protein kinase signalling pathway, we observe that the estimated undirected graph indeed has biological meanings. For example, Fus1, Fus3, Ste12, Ste20, Ste18, Ste11, Dig2 and Cdc42 are linked together, suggesting a strong interaction mechanism between these genes. These genes are all involved in the yeast pheromone and mating process. In contrast, genes Sho1, Ste20, Ste11, Ctt1, Glo1, Ypd1 and Msn4 are linked since they all participate in osmolyte synthesis. Finally, genes Swi4, Bni1, Bck1 and Fks1 are linked due to their interaction in the cell wall remodelling process.

For comparison we also obtain the Gaussian graph estimated by constrained ℓ_1 penalization of Cai et al. (2011) and the estimation of Friedman et al. (2008) without adjusting for the genetic effects on gene expressions. We use five-fold crossvalidation to choose the tuning parameter for both methods, resulting in $\lambda = 0{\cdot}20$ and $\lambda = 0{\cdot}15$, respectively. The method of Cai et al. (2011) identifies 146 links and the method of Friedman et al. (2008) identified 543 links. Both graphs include too many links and are hard to interpret biologically.

For the second dataset, we analyse genes that belong to the yeast protein-protein interaction network (Steffen et al., 2002). We select 1207 genes with variance greater than 0·05. Five-fold crossvalidation chooses the tuning parameters as $\lambda = 0{\cdot}15$ and $\tau = 0{\cdot}20$, leading to an estimated covariate-adjusted Gaussian graph with 3588 links out of 727 821 possible links. In contrast, the method of Friedman et al. (2008) identifies 25 117 links with an optimal tuning parameter $\lambda = 0{\cdot}23$, and the method of Cai et al. (2011) identifies 5983 links with the selected tuning parameter $\lambda = 0{\cdot}18$. Again, it seems that the covariate-adjusted Gaussian graphical model gives a sparser graph than the standard Gaussian graphical model when the genetic effects on gene expressions are ignored.

7. Extension

The two-stage procedure introduced in this paper can be extended to yield an iterative procedure. For fixed tuning parameters λ_n and τ_n, given the current estimate of Ω_0, say $\hat{\Omega}_0$, one can estimate Γ_0 by solving the optimization problem

$$\hat{\Gamma} \in \mathop{\arg\min}_{\Gamma \in R^{p\times q}} \{|\Gamma|_1 : |S_{x\hat{\Omega}_0 y} - \Gamma S_{x\hat{\Omega}_0 x}|_\infty \leqslant \lambda_n\},$$

where $S_{x\hat{\Omega}_0 y} = n^{-1}\sum_{k=1}^n (y_k - \bar{y})\hat{\Omega}_0^{-1}(x_k - \bar{x})^{\mathrm{T}}$ and $S_{x\hat{\Omega}_0 x} = n^{-1}\sum_{k=1}^n (x_k - \bar{x})\hat{\Omega}_0^{-1}(x_k - \bar{x})^{\mathrm{T}}$. One can then iteratively update Γ_0 and Ω_0 until convergence. This however increases the computational time dramatically.

Acknowledgement

This research was supported by the National Institutes of Health, the National Science Foundation, the Program for Professor of Special Appointment at Shanghai Institutions of Higher Learning and the Foundation of National Excellent Doctoral Dissertation of China. Xie thanks Dr. Axel Kholmeyer's help in organizing the cluster for computation.

Appendix

Proofs of the theorems

The first lemma is an exponential inequality from Cai & Liu (2011) on the partial sums of independent random variables.

Lemma A1. *Let $\xi_1, \ldots, \xi_n$ be independent random variables with mean zero. Suppose that there exists some $t > 0$ and $\bar{B}_n$ such that $\sum_{k=1}^n E\{\xi_k^2 e^{t|\xi_k|}\} \leqslant \bar{B}_n^2$. Then uniformly for $0 < x \leqslant \bar{B}_n$,*

$$\mathrm{pr}\left(\sum_{k=1}^n \xi_k \geqslant C_t \bar{B}_n x\right) \leqslant \exp(-x^2),$$

where $C_t = t + t^{-1}$.

Proof of Theorems 1 *and* 2. Without loss of generality, we assume that $E(x)=0$. Recall that $E(z)=0$. We show that with probability greater than $1-O\{(pq)^{-1}\}$,

$$|S_{xy}-\Gamma_0 S_{xx}|_\infty \leqslant \lambda_n. \tag{A1}$$

To prove (A1), it suffices to show that

$$\left|\frac{1}{n}\sum_{k=1}^{n}(z_k-\bar{z})(x_k-\bar{x})^{\mathrm{T}}\right|_\infty \leqslant \lambda_n. \tag{A2}$$

Taking $\xi_k = z_{ki}x_{kj}$ in Lemma A1 and noting that $\max_{i,j} E\exp(t|z_{ki}x_{kj}|)\leqslant K$ for all $|t|\leqslant \min(\eta, \eta/K)$, we have

$$\max_{i,j}\mathrm{pr}\left(n^{-1}\left|\sum_{k=1}^{n} z_{ki}x_{kj}\right|\geqslant \lambda_n/2\right)\leqslant 2(pq)^{-2}.$$

By Lemma A1, we have

$$\max_{j}\mathrm{pr}[|\bar{x}_j|\geqslant C\{\log(pq)/n\}^{1/2}]\leqslant 2(pq)^{-2},\quad \max_{i}\mathrm{pr}[|\bar{z}_i|\geqslant C\{\log(pq)/n\}^{1/2}]\leqslant 2(pq)^{-2}$$

for some constant $C>0$. This implies (A2). Let $\hat{\Gamma}=(\hat{\gamma}_{ij})=(\hat{\gamma}_1,\ldots,\hat{\gamma}_p)^{\mathrm{T}}$ be the solution of (3). Then by (A1), we have $|(\hat{\Gamma}-\Gamma_0)S_{xx}|\leqslant 2\lambda_n$. Moreover, by the equivalence between (3) and (4), we have $\sum_{j=1}^{q}|\hat{\gamma}_{ij}|\leqslant \sum_{j=1}^{q}|\gamma_{ij}^0|$ for all $i=1,\ldots,p$. Set $\|\Gamma_0\|_{L_\infty}=\max_{1\leqslant i\leqslant p}\sum_{j=1}^{q}|\gamma_{ij}^0|$. We have $\|\hat{\Gamma}\|_{L_\infty}\leqslant \|\Gamma_0\|_{L_\infty}$. Also by Lemma A1, we have

$$\mathrm{pr}[|\Sigma_x - S_{xx}|_\infty \geqslant C\{\log(pq)/n\}^{1/2}]\leqslant 2(pq)^{-1}$$

for some constant $C>0$. Then, with probability greater than $1-O\{(pq)^{-1}\}$, we have

$$\begin{aligned}|(\hat{\Gamma}-\Gamma_0)\Sigma_x|_\infty &\leqslant |(\hat{\Gamma}-\Gamma_0)S_{xx}|_\infty + |(\hat{\Gamma}-\Gamma_0)(\Sigma_x - S_{xx})|_\infty\\ &\leqslant 2\lambda_n + C\|\hat{\Gamma}-\Gamma_0\|_{L_\infty}\{\log(pq)/n\}^{1/2}.\end{aligned}$$

It follows that

$$\begin{aligned}|\hat{\Gamma}-\Gamma_0|_\infty &\leqslant |(\hat{\Gamma}-\Gamma_0)\Sigma_x|_\infty\|\Sigma_x^{-1}\|_{L_1}\\ &\leqslant 2\|\Sigma_x^{-1}|_{L_1}\lambda_n + C|\Sigma_x^{-1}\|_{L_1}\|\hat{\Gamma}-\Gamma_0\|_{L_\infty}\{\log(pq)/n\}^{1/2}.\end{aligned}\tag{A3}$$

Let $t_n=|\hat{\Gamma}-\Gamma_0|_\infty$. Define $h_j=(h_{j1},\ldots,h_{jq})^{\mathrm{T}}=\hat{\gamma}_j-\gamma_j^0$, $h_j^1=(\hat{\gamma}_{ji}I\{|\hat{\gamma}_{ji}|\geqslant 2t_n\}:1\leqslant i\leqslant q)^{\mathrm{T}}-\gamma_j^0$ and $h_j^2=h_j-h_j^1$. Then $|h_j^2|_1-|h_j^1|_1+|\gamma_j^0|_1\leqslant |h_j^2|_1+|h_j^1+\gamma_j^0|_1=|h_j+\gamma_j^0|_1\leqslant|\gamma_j^0|_1$. So we have $|h_j|_1\leqslant 2|h_j^1|_1$. It suffices to estimate $|h_j^1|_1$. We have

$$\begin{aligned}|h_j^1|_1 &= \sum_{i=1}^{q}|\hat{\gamma}_{ji}I\{|\hat{\gamma}_{ji}|\geqslant 2t_n\}-\gamma_{ji}^0|\\ &= \sum_{i=1}^{q}|\hat{\gamma}_{ji}-\gamma_{ji}^0|I\{|\hat{\gamma}_{ji}|\geqslant 2t_n\}+\sum_{i=1}^{q}|\gamma_{ji}^0|I\{|\hat{\gamma}_{ji}|<2t_n\}\\ &\leqslant \sum_{i=1}^{q}t_nI\{|\gamma_{ji}^0|\geqslant t_n\}+\sum_{i=1}^{q}|\gamma_{ji}^0|I\{|\gamma_{ji}^0|<3t_n\}\\ &\leqslant t_n^{1-\delta_1}\sum_{i=1}^{q}|\gamma_{ji}^0|^{\delta_1}+(3t_n)^{1-\delta_1}\sum_{i=1}^{q}|\gamma_{ji}^0|^{\delta_1}.\end{aligned}$$

Therefore,

$$\|\hat{\Gamma}-\Gamma_0\|_{L_\infty}\leqslant Cs_1(q)N_q^{1-\delta_1}\lambda_n^{1-\delta_1}+C\|\hat{\Gamma}-\Gamma_0\|_{L_\infty}^{1-\delta_1}s_1(q)N_q^{1-\delta_1}\lambda_n^{1-\delta_1}.$$

If $\|\hat{\Gamma}-\Gamma_0\|_{L_\infty} \leqslant 1$, then we have $\|\hat{\Gamma}-\Gamma_0\|_{L_\infty} \leqslant Cs_1(q)N_q^{1-\delta_1}\lambda_n^{1-\delta_1}$. If $\|\hat{\Gamma}-\Gamma_0\|_{L_\infty} > 1$, then by (7), we have for large n,

$$\|\hat{\Gamma}-\Gamma_0\|_{L_\infty} \leqslant Cs_1(q)N_q^{1-\delta_1}\lambda_n^{1-\delta_1} + \frac{1}{2}\|\hat{\Gamma}-\Gamma_0\|_{L_\infty}.$$

Thus (8) holds with probability greater than $1 - O\{(pq)^{-1}\}$. By (8) and (7), we have $\|\hat{\Gamma}-\Gamma_0\|_{L_\infty} \leqslant 1$ with probability greater than $1 - O\{(pq)^{-1}\}$. This, together with (A3), implies (10). Finally, (9) follows from (8), (10) and the inequality $p^{-1}\|\hat{\Gamma}-\Gamma_0\|_F^2 \leqslant |\hat{\Gamma}-\Gamma_0|_\infty \|\hat{\Gamma}-\Gamma_0\|_{L_\infty}$. □

Proof of Theorems 4 and 5. Recall that $E(z) = 0$. Set

$$\hat{\Sigma}_z = \frac{1}{n}\sum_{k=1}^{n} z_k z_k^{\mathrm{T}}.$$

We suppose that

$$|(S_{yy} - \hat{\Sigma}_z)\Omega_0|_\infty \leqslant \tau_n \tag{A4}$$

and

$$|(\hat{\Sigma}_z - \Sigma_0)\Omega_0|_\infty \leqslant \tau_n. \tag{A5}$$

Then we have $|I_{p\times p} - S_{yy}\Omega_0|_\infty = |(S_{yy}-\Sigma_0)\Omega_0|_\infty \leqslant 2\tau_n$. It follows that $|\hat{\Omega}_1 - \Omega_0|_\infty \leqslant |(I_{p\times p} - \Omega_0 S_{yy})\hat{\Omega}_1|_\infty + |\Omega_0(I_{p\times p} - S_{xx}\hat{\Omega}_1)|_\infty \leqslant 2\|\Omega_0\|_{L_1}\tau_n$. This proves Theorem 5. Following the arguments as the proof of Theorem 1, we can get Theorem 4.

It remains to prove (A4) and (A5). Write $\Delta_n = \hat{\Gamma} - \Gamma_0$. Then we have

$$S_{xx} = \frac{1}{n}\sum_{k=1}^{n}(z_k - \Delta_n x_k)(z_k - \Delta_n x_k)^{\mathrm{T}}.$$

We now prove that with probability greater than $1 - O\{(pq)^{-1}\}$,

$$\left|\frac{1}{n}\sum_{k=1}^{n} z_k x_k^{\mathrm{T}}\Delta_n^{\mathrm{T}}\right|_\infty \leqslant CM_p^{-1}\left\{\frac{\log(pq)}{n}\right\}^{1/2} \tag{A6}$$

and

$$\left|\frac{1}{n}\sum_{k=1}^{n} \Delta_n x_k x_k^{\mathrm{T}}\Delta_n^{\mathrm{T}}\right|_\infty \leqslant CM_p^{-1}\left\{\frac{\log(pq)}{n}\right\}^{1/2}. \tag{A7}$$

First, recall that

$$\max_{i,j} \mathrm{pr}\left(n^{-1}\left|\sum_{k=1}^{n} z_{ki}x_{kj}\right| \geqslant \lambda_n/2\right) \leqslant C(pq)^{-2}. \tag{A8}$$

Write $\Delta_n = (\delta_{ij})$, $x_k = (x_{k1},\ldots,x_{kq})^{\mathrm{T}}$ and $z_k = (z_{k1},\ldots,z_{kp})^{\mathrm{T}}$. To prove (A6), we need to show only that with probability greater than $1 - O\{(pq)^{-1}\}$,

$$\max_{i,l}\left|\frac{1}{n}\sum_{k=1}^{n}(z_{ki}x_{k1}\delta_{l1} + \cdots + z_{ki}x_{kq}\delta_{lq})\right| \leqslant C\left\{\frac{\log(pq)}{n}\right\}^{1/2}.$$

By (8), (12) and (A8),

$$\begin{aligned}\max_{i,l}\left|\frac{1}{n}\sum_{k=1}^{n}\sum_{j=1}^{q} z_{ki}x_{kj}\delta_{lj}\right| &\leqslant \|\hat{\Gamma}-\Gamma_0\|_{l_\infty}\max_{i,j}\left|\frac{1}{n}\sum_{k=1}^{n} z_{ki}x_{kj}\right| \\ &\leqslant CM_p^{-1}\max_{i,j}\left|\frac{1}{n}\sum_{k=1}^{n} z_{ki}x_{kj}\right| \\ &\leqslant CM_p^{-1}\{\log(pq)/n\}^{1/2}.\end{aligned}$$

Thus (A6) holds. It remains to show (A7), which is equivalent to show that with probability greater than $1-O\{(pq)^{-1}\}$,

$$\max_{i,l}\left|\frac{1}{n}\sum_{k=1}^{n}\sum_{j=1}^{q}\delta_{ij}x_{kj}\sum_{j=1}^{q}\delta_{lj}x_{kj}\right| \leqslant CM_p^{-1}\left\{\frac{\log(pq)}{n}\right\}^{1/2}. \tag{A9}$$

By Lemma A1, we can get

$$\max_j \mathrm{pr}\left(\frac{1}{n}\sum_{k=1}^{n}x_{kj}^2 \geqslant C\right) = O\{(pq)^{-2}\}$$

for some constant $C>0$. By (10), (8) and (12),

$$\max_i \frac{1}{n}\sum_{k=1}^{n}\left(\sum_{j=1}^{q}\delta_{ij}x_{kj}\right)^2 \leqslant \max_i \sum_{j=1}^{q}\delta_{ij}^2\frac{1}{n}\sum_{k=1}^{n}x_{kj}^2 \leqslant CM_p^{-1}\left\{\frac{\log(pq)}{n}\right\}^{1/2}$$

with probability greater than $1-O\{(pq)^{-1}\}$. This implies (A9).

We next prove (A5). Write

$$(\hat{\Sigma}_z-\Sigma_0)\Omega_0 = \frac{1}{n}\sum_{k=1}^{n}(z_k z_k^{\mathrm{T}}\Omega_0 - Ez_k z_k^{\mathrm{T}}\Omega_0).$$

Note that $\mathrm{var}(z_{ki}) = \sigma_{ii}^0$ and $\mathrm{var}\{(z_k^{\mathrm{T}}\Omega_0)_j\} = \omega_{jj}^0$. By assumption (A2), $\max_i \sigma_{ii}^0 \max_j \omega_{jj}^0 \leqslant C_0$. By Lemma A1, we have

$$\max_{i,j}\mathrm{pr}\left[\left|\frac{1}{n}\sum_{k=1}^{n}(z_{ki}(z_k^{\mathrm{T}}\Omega_0)_j - Ez_{ki}(z_k^{\mathrm{T}}\Omega_0)_j)\right| \geqslant C\left\{\frac{\log(pq)}{n}\right\}^{1/2}\right] \leqslant C(pq)^{-2}.$$

for some bounded constant C depending only on C_0, η and K. This yields (A5). □

References

Bickel, P. & Levina, L. (2008). Covariance regularization by thresholding. *Ann. Statist.* **6**, 2577–604.

Brem, R. & Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Nat. Acad. Sci.* **102**, 1572–7.

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

Cai, T., Liu, W. & Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Am. Statist. Assoc.* **106**, 594–607.

Candès, E. & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313–51.

Cheung, V. & Spielman, R. (2002). The genetics of variation in gene expression. *Nature Genet.* **32**, 522–5.

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

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* **38**, D335–60.

Kelder, T., van Iersel, M. P., Hanspers, K., Kutmon, M., Conklin, B. R., Evelo, C. & Pico, A. R. (2012). WikiPathways: building research communities on biological pathways. *Nucleic Acids Res*. 40, D1301–7.

Li, B., Chun, H. & Zhao, H. (2012). Sparse estimation of conditional graphical models with application to gene networks. *J. Am. Statist. Assoc.* **107**, 152–67.

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

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

Mewes, H., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Münsterkötter, M., Rudd, S. & Weil, B. (2002). MIPS: A database for genomes and protein sequences. *Nucleic Acids Res.* **30**, 31–4.

Obozinski, G., Wainwright, M. & Jordan, M. (2011). Support union recovery in high-dimensional multivariate regression. *Ann. Statist.* **39**, 1–47.

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

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D., Pollack, J. R. & Wang, P. (2009b). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Statist.* **41**, 53–77.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

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

Rothman, A., Levina, E. & Zhu, J. (2010). Sparse multivariate regression with covariate estimation. *J. Comp. Graph. Statist.* **19**, 947–62.

Segal, E., Friedman, N., Kaminski, N., Regev, A. & Koller, D. (2005). From signatures to models: Understanding cancer using microarrays. *Nature Genet.* **37**, S38–45.

Steffen, M., Petti, A., Aach, J., D'Haeseleer, P. & Church, G. (2002). Automated modelling of signal transduction networks. *BMC Bioinformatics* **3**, 34.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.

Turlach, B., Venables, W. & Wright, S. (2005). Simultaneous variable selection. *Technometrics* **47**, 349–63.

Yin, J. & Li, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetic genomics data. *Ann. Appl. Statist.* **5**, 2630–50.

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

[*Received June* 2011. *Revised August* 2012]