Biometrika (2007), **94**, 1, *pp.* 19–35 doi:10.1093/biomet/asm018

Printed in Great Britain

Model selection and estimation in the Gaussian graphical model

BY MING YUAN

School of Industrial and Systems Engineering, Georgia Institute of Technology, 755 Ferst Drive NW, Atlanta, Georgia 30332, U.S.A.
myuan@isye.gatech.edu

AND YI LIN

Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706, U.S.A.
yilin@stat.wisc.edu

SUMMARY

We propose penalized likelihood methods for estimating the concentration matrix in the Gaussian graphical model. The methods lead to a sparse and shrinkage estimator of the concentration matrix that is positive definite, and thus conduct model selection and estimation simultaneously. The implementation of the methods is nontrivial because of the positive definite constraint on the concentration matrix, but we show that the computation can be done effectively by taking advantage of the efficient maxdet algorithm developed in convex optimization. We propose a BIC-type criterion for the selection of the tuning parameter in the penalized likelihood methods. The connection between our methods and existing methods is illustrated. Simulations and real examples demonstrate the competitive performance of the new methods.

Some key words: Covariance selection; Lasso; Maxdet algorithm; Nonnegative garrote; Penalized likelihood.

1. INTRODUCTION

Let $X = (X^{(1)}, \ldots, X^{(p)})$ be a p-dimensional random vector following a multivariate normal distribution $\mathcal{N}_p(\mu, \Sigma)$ with unknown mean μ and nonsingular covariance matrix Σ. Given a random sample $X_1, \ldots, X_n$ of X, we wish to estimate the concentration matrix $C = \Sigma^{-1}$. Of particular interest is the identification of zero entries in the concentration matrix $C = (c_{ij})$, since a zero entry $c_{ij} = 0$ indicates the conditional independence between the two random variables $X^{(i)}$ and $X^{(j)}$ given all other variables. This is the covariance selection problem (Dempster, 1972) or the model-selection problem in the Gaussian concentration graph model (Cox & Wermuth, 1996).

A Gaussian concentration graph model for the Gaussian random vector X is represented by an undirected graph $G = (V, E)$, where V contains p vertices corresponding to the p coordinates and the edges $E = (e_{ij})_{1 \leqslant i < j \leqslant p}$ describe the conditional independence relationships among $X^{(1)}, \ldots, X^{(p)}$. The edge between $X^{(i)}$ and $X^{(j)}$ is absent if and only if $X^{(i)}$ and $X^{(j)}$ are independent conditional on the other variables, and corresponds to $c_{ij} = 0$. Thus parameter estimation and model selection in the Gaussian concentration graph model are equivalent to estimating parameters and identifying zeros in the concentration matrix;

see Whittaker (1990), Lauritzen (1996) and Edwards (2000) for statistical properties of Gaussian concentration graph models and commonly used model selection and parameter estimation methods in such models.

The standard approach to model selection in Gaussian graphical models is greedy stepwise forward-selection or backward-deletion, and parameter estimation is based on the selected model. In each step the edge selection or deletion is typically done through hypothesis testing at some level α. It has long been recognized that this procedure does not correctly take account of the multiple comparisons involved (Edwards, 2000). Another drawback of the common stepwise procedure is its computational complexity. To remedy these problems, Drton & Perlman (2004) proposed a method that produces conservative simultaneous $1 - \alpha$ confidence intervals, and uses these confidence intervals to do model selection in a single step. The method is based on asymptotic considerations. Meinshausen & Bühlmann (2006) proposed a computationally attractive method for covariance selection that can be used for very large Gaussian graphs. They perform neighbourhood selection for each node in the graph and combine the results to learn the structure of a Gaussian concentration graph model. They showed that their method is consistent for sparse high-dimensional graphs. In all of the above mentioned methods, model selection and parameter estimation are done separately. The parameters in the concentration matrix are typically estimated based on the model selected. As demonstrated by Breiman (1996), the discrete nature of such procedures often leads to instability of the estimator: small changes in the data may result in very different estimates. Other recent advances include a Duke University discussion paper by A. Dobra and M. West, who presented a novel Bayesian framework for building Gaussian graphical models and illustrated their approach in a large scale gene expression study, and Li & Gui (2006), who adopted gradient-directed regularization, which is described in a technical report by J. Friedman and B. Popescu, available at http://www-stat.stanford.edu/~jhf, to construct sparse Gaussian graphical models.

In this paper, we propose a penalized-likelihood method that does model selection and parameter estimation simultaneously in the Gaussian concentration graph model. We employ an ℓ_1 penalty on the off-diagonal elements of the concentration matrix. This is similar to the idea of the lasso in linear regression (Tibshirani, 1996). The ℓ_1 penalty encourages sparsity and at the same time gives shrinkage estimates. In addition, we explicitly ensure that the estimator of the concentration matrix is positive definite. We also introduce a 'nonnegative garrote' type method that is closely related to the aforementioned approach.

There is a connection between the neighbourhood-selection method in Meinshausen & Bühlmann (2006) and our penalized-likelihood approach, which we illustrate in § 5. The neighbourhood-selection method can be cast as a penalized M-estimation without incorporating the positive definiteness or symmetry constraint. The loss function in the penalized M-estimation is a particular quadratic form. The neighbourhood selection method is computationally faster because of its simpler form and because it does not consider the positive definite constraint. Our method is more efficient because of the incorporation of the positive definite constraint and the use of likelihood.

Throughout the paper we assume that the observations are suitably centred and scaled. The sample mean is centred to be zero. One may scale to have the diagonal elements of the sample covariance matrix equal to one or to have the diagonal elements of the sample concentration matrix equal to one. In our experience these two scalings give very similar performance, and in this paper we assume the latter since it seems to be more natural for estimating the concentration matrix.

2. METHODOLOGY

2·1. *Lasso-type estimator*

The loglikelihood for μ and $C = \Sigma^{-1}$ based on a random sample $X_1, \ldots, X_n$ of X is

$$\frac{n}{2}\log|C| - \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)' C (X_i - \mu)$$

up to a constant not depending on μ and C. The maximum likelihood estimator of (μ, Σ) is $(\bar{X}, \bar{A})$, where

$$\bar{A} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})'.$$

The commonly used sample covariance matrix is $S = n\bar{A}/(n-1)$. The concentration matrix C can be naturally estimated by $\bar{A}^{-1}$ or S^{-1}. However, because of the large number, $p(p+1)/2$, of unknown parameters to be estimated, S is not a stable estimator of Σ for moderate or large p. In general, the matrix S^{-1} is positive definite when $n \geqslant p$, but does not lead to 'sparse' graph structure since the matrix typically contains no zero entry.

To achieve 'sparse' graph structure and to give a better estimator of the concentration matrix, we adapt the lasso idea and seek the minimizer $(\hat{\mu}, \hat{C})$ of

$$-\log|C| + \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)' C (X_i - \mu) \qquad \text{subject to} \qquad \sum_{i \neq j}|c_{ij}| \leqslant t, \tag{1}$$

over the set of positive definite matrices. Here $t \geqslant 0$ is a tuning parameter. When $t = \infty$, the solution to (1) is the maximum likelihood estimator $\bar{A}^{-1}$ provided that the inverse exists. On the other hand, if $t = 0$, then the constraint forces C to be diagonal, which implies that $X^{(1)}, \ldots, X^{(p)}$ are mutually independent. It is clear that $\hat{\mu} = \bar{X}$ regardless of t. Since the observations are centred, we have $\hat{\mu} = 0$. Therefore, $\hat{C}$ is the positive definite matrix that minimizes

$$-\log|C| + \frac{1}{n}\sum_{i=1}^{n} X_i' C X_i \qquad \text{subject to} \qquad \sum_{i \neq j}|c_{ij}| \leqslant t. \tag{2}$$

We can further rewrite (2) as

$$-\log|C| + \operatorname{tr}(C\bar{A}) \qquad \text{subject to} \qquad \sum_{i \neq j}|c_{ij}| \leqslant t. \tag{3}$$

Since both the objective function and feasible region of (3) are convex, we can equivalently use the Lagrangian form

$$-\log|C| + \operatorname{tr}(C\bar{A}) + \lambda\sum_{i \neq j}|c_{ij}|, \tag{4}$$

with $\lambda \geqslant 0$ being the tuning parameter.

2·2. *Nonnegative garrote-type estimator*

If a relatively reliable estimator $\tilde{C}$ of C is available, a shrinkage estimator can be defined through $c_{ij} = d_{ij}\tilde{c}_{ij}$, where the symmetric matrix $D = (d_{ij})$ is the minimizer of

$$-\log|C| + \operatorname{tr}(C\bar{A}) \qquad \text{subject to} \qquad \sum_{i \neq j} d_{ij} \leqslant t, d_{ij} \geqslant 0, \tag{5}$$

and with C positive definite. Equivalently, this can be written as

$$-\log|C| + \operatorname{tr}(C\bar{A}) + \lambda \sum_{i \neq j} \frac{c_{ij}}{\tilde{c}_{ij}}, \tag{6}$$

subject to $c_{ij}/\tilde{c}_{ij} \geqslant 0$ and with C positive definite. For a relatively large sample size, $\bar{A}^{-1}$ is an obvious choice for the preliminary estimator. This procedure is similar in spirit to the nonnegative garrote estimator proposed by Breiman (1995) for linear regression.

2·3. *Illustration*

Consider the special case in which $p = 2$. Denote the maximum likelihood estimator of the concentration matrix by

$$\hat{C}_0 = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix},$$

where the diagonal elements are 1 because of the scaling. Therefore,

$$\bar{A} = \hat{C}_0^{-1} = \frac{1}{1-r^2} \begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix}.$$

Substitution in (4) gives

$$-\log\left(c_{11}c_{22} - c_{12}^2\right) + \frac{c_{11} + c_{22}}{1-r^2} - \frac{2rc_{12}}{1-r^2} + 2\lambda|c_{12}|,$$

where we used the fact that C is symmetric.

LEMMA 1. *In the case of the bivariate normal, the proposed penalized likelihood estimator given by the solution to* (4) *is*

$$\hat{c}_{12} = \left(\frac{(1-r^2)\{|r| - \lambda(1-r^2)\}}{1 - \{|r| - \lambda(1-r^2)\}^2} \right)_+ \operatorname{sign}(r),$$

where $(x)_+ = \max(x, 0)$ *and*

$$\hat{c}_{11} = \hat{c}_{22} = \frac{1}{2}\left[(1-r^2) + \sqrt{\{(1-r^2)^2 + 4\hat{c}_{12}^2\}}\right]. \tag{7}$$

Similarly, the garrote type estimator can also be obtained in an explicit form in this case.

LEMMA 2. *With* $\tilde{C} = \bar{A}^{-1}$, *the minimizer of* (6) *is*

$$\hat{c}_{12} = \left(\frac{(1-r^2)\{r^2 - \lambda(1-r^2)\}}{|r| - \{r^2 - \lambda(1-r^2)\}^2/|r|} \right)_+ \operatorname{sign}(r)$$

and $\hat{c}_{11} = \hat{c}_{22}$ *are given by* (7).

The estimators are illustrated Fig. 1. If the true value of c_{12} is zero, r will tend to be small in magnitude. With an appropriately chosen λ, both estimates of c_{12} can be shrunk to zero, so that model selection for the graphical model can be achieved. Note that the proposed estimators are continuous functions of r and consequently of the data. Such continuity, not shared by the existing methods that perform maximum likelihood estimation on a selected graph structure, ensures the stability of our estimators. The garrote-type estimator penalizes large r's less heavily than small r's. As will be demonstrated in the next section,

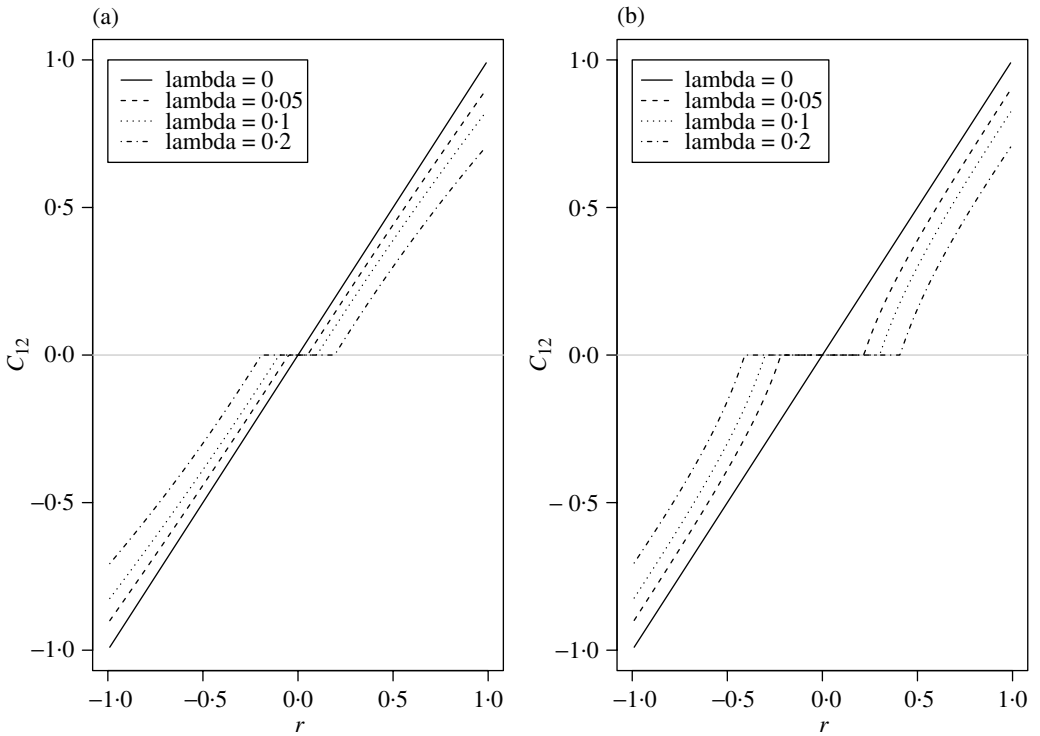

Fig. 1. (a) Lasso-type estimator, (b) garrote type estimator for the case of $p = 2$.

this can be advantageous for model-fitting. However, the disadvantage of the garrote-type estimator is that it can only be applied when a good initial estimator is available.

3. Asymptotic theory

In this section, we derive asymptotic properties of the proposed estimators that are analogous to those for the lasso (Knight & Fu, 2000). For simplicity, we assume that p is held fixed as the sample size $n \to \infty$. Although it might be more realistic to consider the case when $p \to \infty$ as $n \to \infty$, the following results nevertheless provide an adequate illustration of the mechanism of the proposed estimators.

Theorem 1. *If* $\sqrt{n}\lambda \to \lambda_0 \geqslant 0$ *as* $n \to \infty$, *the lasso-type estimator is such that*

$$\sqrt{n}(\hat{C} - C) \to \arg\min_{U=U'}(V),$$

in distribution, where

$$V(U) = \mathrm{tr}\,(U\Sigma U\Sigma) + \mathrm{tr}(UW) + \lambda_0 \sum_{i \neq j} \left\{u_{ij}\mathrm{sign}(c_{ij})I(c_{ij} \neq 0) + |u_{ij}|I(c_{ij} = 0)\right\},$$

in which W *is a random symmetric* $p \times p$ *matrix such that* $\mathrm{vec}(W) \sim \mathcal{N}(0, \Lambda)$, *and* Λ *is such that*

$$\mathrm{cov}(w_{ij}, w_{i'j'}) = \mathrm{cov}(X^{(i)}X^{(j)}, X^{(i')}X^{(j')}).$$

As an illustration, consider an example where $p = 3$ and

$$C = \begin{pmatrix} 1 & \frac{1}{3} & 0 \\ \frac{1}{3} & 1 & \frac{2}{3} \\ 0 & \frac{2}{3} & 1 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 1{\cdot}25 & -0{\cdot}75 & 0{\cdot}5 \\ -0{\cdot}75 & 2{\cdot}25 & -1{\cdot}5 \\ 0{\cdot}5 & -1{\cdot}5 & 2 \end{pmatrix}.$$

Note that, for $i \neq j \neq k \neq l$,

$$\begin{aligned} E\left\{(X^{(i)})^4\right\} &= 3\Sigma_{ii}^2 \\ E\left\{(X^{(i)})^3 X^{(j)}\right\} &= 3\Sigma_{ii}\Sigma_{ij} \\ E\left\{(X^{(i)})^2 (X^{(j)})^2\right\} &= \Sigma_{ii}\Sigma_{jj} + 2\Sigma_{ij}^2 \\ E\left\{(X^{(i)})^2 X^{(j)} X^{(k)}\right\} &= \Sigma_{ii}\Sigma_{jk} + 2\Sigma_{ij}\Sigma_{ik} \\ E\left(X^{(i)} X^{(j)} X^{(k)} X^{(l)}\right) &= \Sigma_{ij}\Sigma_{kl} + \Sigma_{ik}\Sigma_{jl} + \Sigma_{il}\Sigma_{jk}. \end{aligned}$$

After some tedious algebraic manipulation, we obtain that

$$\begin{pmatrix} W_{11} \\ W_{12} \\ W_{13} \\ W_{22} \\ W_{23} \\ W_{33} \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} 3{\cdot}125 & -1{\cdot}875 & 1{\cdot}25 & 1{\cdot}125 & -0{\cdot}75 & 0{\cdot}5 \\ -1{\cdot}875 & 3{\cdot}375 & -2{\cdot}25 & -3{\cdot}375 & 2{\cdot}25 & -1{\cdot}5 \\ 1{\cdot}25 & -2{\cdot}25 & 2{\cdot}75 & 2{\cdot}25 & -2{\cdot}25 & 2 \\ 1{\cdot}125 & -3{\cdot}375 & 2{\cdot}25 & 10{\cdot}125 & -6{\cdot}75 & 4{\cdot}5 \\ -0{\cdot}75 & 2{\cdot}25 & -2{\cdot}25 & -6{\cdot}75 & 6{\cdot}75 & -6 \\ 0{\cdot}5 & -1{\cdot}5 & 2 & 4{\cdot}5 & -6 & 8 \end{pmatrix}\right).$$

We simulated 1000 observations from the distribution of $\arg\min V$. Figure 2 gives the scatterplot of the off-diagonal elements for $\lambda_0 = 0$, 0·5 and 1. When $\lambda_0 = 0$, our estimator is asymptotically equivalent to the maximum likelihood estimator, and the asymptotic distribution for the elements of C is multivariate normal; see Fig. 2(a). If $\lambda_0 > 0$, the proposed estimator will have a positive probability of estimating c_{13} by its true value 0, and this probability increases as λ_0 increases. From Theorem 1 $\mathrm{pr}(\hat{c}_{13} = 0)$ tends to 0·30 if $\lambda_0 = 0{\cdot}5$ and to 0·45 when $\lambda_0 = 1$.

Similarly to Theorem 1, we can derive the asymptotic properties of the nonnegative garrote-type estimator.

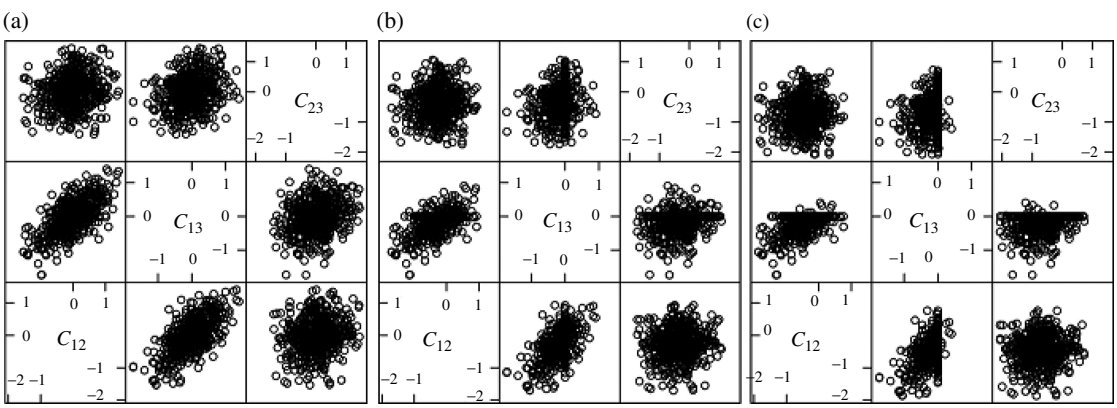

Fig. 2. Example with $p = 3$. Asymptotic distribution of our estimator as estimated by 1000 simulations, for different values of λ_0, (a) $\lambda_0 = 0$, (b) $\lambda_0 = 0{\cdot}5$, (c) $\lambda_0 = 1$.

THEOREM 2. *Denote by* $\hat{C}$ *the minimizer of* (6) *with initial estimator* $\tilde{C} = \bar{A}^{-1}$. *If* $n\lambda \to \infty$ *and* $\sqrt{n}\lambda \to 0$ *as* $n \to \infty$, *then* $\mathrm{pr}(\hat{c}_{ij} = 0) \to 1$ *if* $c_{ij} = 0$, *and other elements of* $\hat{C}$ *have the same limiting distribution as the maximum likelihood estimator on the true graph structure.*

Theorem 2 indicates that the garrote-type estimator enjoys the so-called oracle property: it selects the right graph with probability tending to one and at the same time gives a root-n consistent estimator of the concentration matrix.

4. COMPUTATION

4·1. *The maxdet problem*

The nonlinearity of the objective function and the positive-definiteness constraint make the optimization problem (3) nontrivial. We take advantage of the connection between (3) and the determinant-maximization problem, the maxdet problem (Vandenberghe et al., 1998), which can be solved very efficiently with the interior point algorithm.

The maxdet problem is of the form

$$\min_{x \in R^m} b'x - \log|G(x)|,$$

where $b \in R^m$, $G(x)$ is positive definite, $F(x)$ is positive semidefinite, and the functions $G : R^m \to R^{l \times l}$ and $F : R^m \to R^{l \times l}$ are affine:

$$G(x) = G_0 + x_1 G_1 + \cdots + x_m G_m,$$
$$F(x) = F_0 + x_1 F_1 + \cdots + x_m F_m,$$

where F_i and G_i are symmetric matrices. To use the algorithm of Vandenberghe et al. (1998), it is also required that $F_i, i = 1, \ldots, m$, be linearly independent and the same be true of $G_i, i = 1, \ldots, m$. It is not hard to see that the garrote-type estimator (5) solves a maxdet problem.

4·2. *Algorithm for lasso-type estimator*

If the signs of the c_{ij}'s are known, (3) can be expressed as the following maxdet problem:

$$\min_C \; 2\sum_{i<j} a_{ij}c_{ij} + \sum_i a_{ii}c_{ii} - \log\left|\sum_i c_{ii}I^{(i)} + \sum_{i<j} c_{ij}I^{(ij)}\right|,$$

subject to $\sum_i c_{ii}I^{(i)} + \sum_{i<j} c_{ij}I^{(ij)}$ being positive definite,

$$t - 2\sum_{i<j} c_{ij}s_{ij} \geqslant 0, \quad s_{ij}c_{ij} \geqslant 0, \tag{8}$$

where $C = (c_{ij})$, $S = (s_{ij})$, $\bar{A} = (a_{ij})$, $I^{(i)}$ is an $n \times n$ matrix with the (i, i)th entry being 1 and all other entries being 0, $I^{(ij)}$ is an $n \times n$ matrix with the (i, j)th and the (j, i)th entries being 1 and all other entries being 0, and s_{ij} is the sign of c_{ij}. Since the signs of the c_{ij}'s are not known in advance, we propose to update the s_{ij}'s and c_{ij}'s iteratively using the following steps.

Step 1. Let $\hat{C}_{\mathrm{old}} = \bar{A}^{-1}$ and $s_{ij} = \mathrm{sign}\{(\hat{C}_{\mathrm{old}})_{ij}\}$ for all $i \neq j$.

Step 2. Let $\hat{C}_{\mathrm{new}}$ solve (8) over the set of positive definite matrices.

Step 3. If $\hat{C}_{\text{new}} = \hat{C}_{\text{old}}$, then stop and let $\hat{C} = \hat{C}_{\text{new}}$. Otherwise, set $\hat{C}_{\text{old}} = \hat{C}_{\text{new}}$ and $s_{ij} = -s_{ij}$ for all pairs (i, j) such that $\hat{c}_{ij} = 0$ and go back to Step 2.

In our experience the algorithm usually converges within a small number of iterations. Clearly, other initial values for s can also be used.

LEMMA 3. *The above algorithm always converges and converges to the solution to* (3).

4·3. *Tuning*

So far we have concentrated on the calculation of the minimizer of (3) for any fixed tuning parameter t. In practice, we need to choose a tuning parameter so as to minimize a score measuring the goodness-of-fit. A commonly used such score is the multifold crossvalidation score, but a computationally more efficient alternative is the BIC for model selection and estimation. To evaluate the BIC for the current setting, one must first obtain an estimate of the degrees of freedom, which is defined as the number of unknown parameters in the case of the maximum likelihood estimate.

Let $\hat{A} = \hat{C}^{-1}$ and $\Omega = \{(i, j) : \hat{c}_{ij} \neq 0\}$. From the Karush-Kuhn-Tucker conditions, it is not hard to see that the lasso-type estimator satisfies $\hat{a}_{ij} = \bar{a}_{ij} + \lambda \text{sign}(\hat{c}_{ij})$ for all pairs $(i, j) \in \Omega$. The remaining $\text{card}(\Omega^c)/2$ unique entries of $\hat{A}$ can be obtained by solving $\text{card}(\Omega^c)/2$ equations, $\hat{C}_{ij} = 0$, where $i < j$ and $\text{card}(\cdot)$ is the cardinality of a set. Therefore, $\hat{C}$ relies on the observations only through $\bar{a}_{ij}$, $(i, j) \in \Omega$. Note that the number of parameters in $\{\bar{a}_{ij} : (i, j) \in \Omega\}$ is $\sum_{i \leqslant j} \hat{e}_{ij}$. Since the $\bar{a}_{ij}$'s are maximum likelihood estimates, we can define, for a given tuning parameter t,

$$\text{BIC}(t) = -\log|\hat{C}(t)| + \text{tr}\{\hat{C}(t)\bar{A}\} + \frac{\log n}{n} \sum_{i \leqslant j} \hat{e}_{ij}(t),$$

where $\hat{e}_{ij} = 0$ if $\hat{c}_{ij} = 0$, and $\hat{e}_{ij} = 1$ otherwise.

5. QUADRATIC APPROXIMATION

Provided that $\bar{A}$ is nonsingular, a second-order approximation to the objective function of (3) around $\bar{A}^{-1}$ can be written as (Boyd & Vandenberghe, 2003)

$$\text{tr}\left\{(C - \bar{A}^{-1})\bar{A}(C - \bar{A}^{-1})\bar{A}\right\}.$$

Therefore, the solution to (4) can be approximated by the solution to

$$\text{tr}\left\{(C - \bar{A}^{-1})\bar{A}(C - \bar{A}^{-1})\bar{A}\right\} + \lambda|C|_{\ell_1}. \tag{9}$$

This second-order approximation is closely connected to the approach proposed by Meinshausen & Bühlmann (2006). In their approach, for each $i = 1, \ldots, p$, we seek the minimizer $\hat{\theta}_{i,-i} = (\hat{\theta}_{i1}, \ldots, \hat{\theta}_{i(i-1)}, \hat{\theta}_{i(i+1)}, \ldots, \hat{\theta}_{ip}) \in R^{p-1}$ of

$$\frac{1}{n}\left|\left|X^{(i)} - X^{[-i]}\theta_{i,-i}\right|\right|^2 + \lambda \sum_{j \neq i} |\theta_{ij}|, \tag{10}$$

where $X^{[-i]}$ is the $n \times (p-1)$ matrix resulting from the deletion of the ith column from the data matrix X. A vertex j is taken to be a neighbour of vertex i if and only if $\hat{\theta}_{ij} \neq 0$. The two vertices are connected by an edge in the graphical model if either vertex is the neighbour of the other one.

Note that $\theta_{ii}, i = 1, \ldots, p$, are not determined. For notational purposes, we write $\theta_{ii} = 1$ for $i = 1, \ldots, p$. Recall that we scale each component of X so that all the diagonal elements of the sample concentration matrix are unity. The following lemma reveals a close connection between the approach of Meinshausen & Bühlmann (2006) and the second-order approximation (9).

LEMMA 4. *The matrix* $\Theta = (\theta_{ij})$ *defined by* (10) *is the unconstrained solution to*

$$\min_C \operatorname{tr}\left\{(C - \bar{A}^{-1})' \bar{A} (C - \bar{A}^{-1})\right\} + \lambda |C|_{\ell_1}, \tag{11}$$

over all $p \times p$ *matrices with diagonal elements fixed at* 1.

Lemma 4 shows that the approach of Meinshausen & Bühlmann (2006) seeks a sparse C close to $\bar{A}^{-1}$. However, it does not incorporate the symmetry and positive-definiteness constraint in the estimation of the concentration matrix, and therefore an additional step is needed to estimate either the covariance matrix or the concentration matrix. Also, the loss function used by Meinshausen & Bühlmann is different from the quadratic approximation to the loglikelihood, and therefore the approach is expected to be less efficient than our penalized likelihood method or the corresponding quadratic approximation (9).

6. SIMULATION

We consider eight different models in our simulation.

Model 1. Heterogeneous model with $\Sigma = \operatorname{diag}(1, 2, \ldots, n)$.

Model 2. An AR(1) model with $c_{ii} = 1$ and $c_{i,i-1} = c_{i-1,i} = 0{\cdot}5$.

Model 3. An AR(2) model with $c_{ii} = 1$, $c_{i,i-1} = c_{i-1,i} = 0{\cdot}5$ and $c_{i,i-2} = c_{i-2,i} = 0{\cdot}25$.

Model 4. An AR(3) model with $c_{ii} = 1$, $c_{i,i-1} = c_{i-1,i} = 0{\cdot}4$ and $c_{i,i-2} = c_{i-2,i} = c_{i,i-3} = c_{i-3,i} = 0{\cdot}2$.

Model 5. An AR(4) model with $c_{ii} = 1$, $c_{i,i-1} = c_{i-1,i} = 0{\cdot}4$, $c_{i,i-2} = c_{i-2,i} = c_{i,i-3} = c_{i-3,i} = 0{\cdot}2$ and $c_{i,i-4} = c_{i-4,i} = 0{\cdot}1$.

Model 6. Full model with $c_{ij} = 2$ if $i = j$ and $c_{ij} = 1$ otherwise.

Model 7. Star model with every node connected to the first node, with $c_{ii} = 1$, $c_{1,i} = c_{i,1} = 0{\cdot}2$ and $c_{ij} = 0$ otherwise.

Model 8. Circle model with $c_{ii} = 1$, $c_{i,i-1} = c_{i-1,i} = 0{\cdot}5$ and $c_{1n} = c_{n1} = 0{\cdot}4$.

For each model, we simulated samples with size 25 and dimension $p = 5$, or size 50 and dimension 10. We compare our methods with the approach of Meinshausen & Bühlmann (2006) and the method proposed by Drton & Perlman (2004) in terms of the Kullback–Leibler loss,

$$\text{KL} = -\log|\hat{C}| + \operatorname{tr}(\hat{C}\Sigma) - (-\log|\Sigma^{-1}| + p),$$

the number of false positives (FP; incorrectly identified edges) and the number of false negatives (FN; incorrectly missed edges). The approach of Meinshausen & Bühlmann (2006) was implemented using the LARS package from R and the method of Drton & Perlman (2004) has also been implemented in the SIN package of R. Their method gives

each edge of the full graph a p-value and two different cut-off values, 5% and 25%, were suggested in their original paper. Both of these methods focus on model selection and do not consider the problem of estimating the covariance matrix or the concentration matrix. For comparison, we estimate the concentration matrix by the maximum likelihood estimate after the graph structure is selected using their methods. Table 1 documents the means and standard errors, in parentheses, from 100 runs for each combination. Our penalized likelihood method is referred to as Lasso in the table because of its connection to the idea of the lasso in linear regression. Similarly, the extension described in § 2·2 is referred to as Garrote in the table.

As shown in Table 1, the proposed penalized likelihood methods enjoy better performance than the other methods. The method of Meinshausen & Bühlmann (2006) and both versions of the Drton–Perlman method tend to have larger FN, which may partly explain their relatively poor performance. However, the results suggest that the proposed penalized likelihood approach combined with BIC may have relatively larger FP. The solution path of (3) may therefore be more informative in determining the graph structure.

All four methods under comparison require the selection of tuning parameters that control the trade-off between sensitivity and specificity. To appreciate better the merits of different methods independently of the tuning parameter selection, we plotted the receiver operating characteristic curves for different models and methods, each averaged over the 100 simulated datasets. The AR(4) and full models are not included because the specificity in these cases is not well defined when $p = 5$. From the plot, not shown here because of space restrictions, the proposed methods outperform the other approaches for all models when $p = 5$ and $n = 25$. In the cases when $p = 10$ and $n = 50$, the performance of all methods improves but Lasso and Garrote still enjoy competitive performance when compared with the other approaches.

7. REAL WORLD EXAMPLES

We first consider three real-world examples. The cork borings data are presented in Whittaker (1990, p. 267) and were originally used by Rao (1948). The $p = 4$ measurements are the weights of cork borings on $n = 28$ trees in the four directions, north, east, south and west. Fret's heads dataset contains head measurements on the first and the second adult son in a sample of 25 families. The 4 variables are the head length of the first son, the head breadth of the first son, the head length of the second son and the head breadth of the second son. The data are also presented in Whittaker (1990, p. 255). The Mathematics marks dataset (Mardia et al., 1979, p. 3) contains the marks of $n = 88$ students in the $p = 5$ examinations in mechanics, vectors, algebra, analysis and statistics. The data also appear in Whittaker (1990, p. 1).

Figures 3–5 depict the solution paths of (3) for each of the three datasets.

To compare the accuracy of different methods, fivefold crossvalidation was applied on the datasets. Table 2 documents the average values of KL distances for each method, where now

$$\text{KL} = -\log|\hat{C}| + \text{tr}(\hat{C}\hat{\Sigma}),$$

in which $\hat{C}$ is the concentration matrix estimated on the training set and $\hat{\Sigma}$ is the sample covariance matrix evaluated on the test set.

Next we considered a larger problem. The opening prices of 35 stocks were collected for the years 2003 and 2004. Different methods were applied to estimate the covariance matrix

Table 1. *Results for the eight simulated models. Averages and standard errors from* 100 *runs*

		Lasso			Garrote			MB			SIN (0·05)			SIN (0·25)		
p	Model	KL	FP	FN	KL	FP	FN	KL	FP	FN	KL	FP	FN	KL	FP	FN
	1	0·27	0·20	0·00	0·31	0·42	0·00	0·45	0·91	0·00	0·26	0·05	0·00	0·32	0·26	0·00
		(0·02)	(0·06)	(0·00)	(0·02)	(0·08)	(0·00)	(0·04)	(0·08)	(0·00)	(0·02)	(0·03)	(0·00)	(0·03)	(0·07)	(0·00)
	2	0·70	3·31	0·07	0·67	1·20	0·14	0·63	0·68	0·16	1·88	0·03	2·47	1·40	0·15	1·59
		(0·05)	(0·12)	(0·03)	(0·05)	(0·12)	(0·04)	(0·05)	(0·08)	(0·04)	(0·06)	(0·02)	(0·10)	(0·06)	(0·05)	(0·09)
	3	0·89	1·29	2·24	0·87	0·60	2·58	0·98	0·47	3·68	1·16	0·01	5·42	1·06	0·07	4·16
		(0·05)	(0·10)	(0·24)	(0·04)	(0·08)	(0·18)	(0·04)	(0·06)	(0·09)	(0·04)	(0·01)	(0·12)	(0·05)	(0·04)	(0·15)
	4	0·79	0·22	5·60	0·80	0·16	5·86	0·83	0·16	6·23	0·93	0·00	8·14	0·90	0·01	6·97
		(0·03)	(0·04)	(0·29)	(0·04)	(0·04)	(0·23)	(0·04)	(0·04)	(0·11)	(0·03)	(0·00)	(0·11)	(0·03)	(0·01)	(0·15)
5	5	0·78	0·00	7·06	0·76	0·00	6·98	0·80	0·00	7·26	0·88	0·00	9·13	0·86	0·00	7·94
		(0·04)	(0·00)	(0·30)	(0·03)	(0·00)	(0·23)	(0·04)	(0·00)	(0·12)	(0·03)	(0·00)	(0·11)	(0·03)	(0·00)	(0·16)
	6	1·09	0·00	4·53	1·11	0·00	4·58	1·30	0·00	7·05	1·28	0·00	6·18	1·18	0·00	3·77
		(0·04)	(0·00)	(0·44)	(0·04)	(0·00)	(0·41)	(0·04)	(0·00)	(0·11)	(0·04)	(0·00)	(0·24)	(0·05)	(0·00)	(0·25)
	7	0·45	0·31	3·47	0·51	0·46	3·02	0·61	0·55	2·75	0·43	0·00	3·92	0·50	0·13	3·61
		(0·02)	(0·08)	(0·10)	(0·03)	(0·08)	(0·12)	(0·03)	(0·07)	(0·10)	(0·02)	(0·00)	(0·03)	(0·03)	(0·05)	(0·06)
	8	0·73	2·55	0·11	0·77	1·28	0·26	0·80	0·17	0·37	1·89	0·03	3·29	1·48	0·11	2·12
		(0·05)	(0·13)	(0·03)	(0·05)	(0·12)	(0·06)	(0·05)	(0·05)	(0·06)	(0·05)	(0·02)	(0·10)	(0·05)	(0·04)	(0·09)
	1	0·22	0·26	0·00	0·26	0·75	0·00	0·63	3·48	0·00	0·23	0·07	0·00	0·26	0·23	0·00
		(0·01)	(0·09)	(0·00)	(0·01)	(0·14)	(0·00)	(0·03)	(0·17)	(0·00)	(0·01)	(0·03)	(0·00)	(0·02)	(0·05)	(0·00)
	2	1·42	31·76	0·00	0·60	4·83	0·00	0·58	2·25	0·00	4·01	0·06	3·75	2·39	0·25	2·05
		(0·04)	(0·26)	(0·00)	(0·02)	(0·27)	(0·00)	(0·02)	(0·15)	(0·00)	(0·14)	(0·02)	(0·14)	(0·12)	(0·06)	(0·12)
	3	1·22	10·87	3·30	1·03	5·86	3·05	1·52	2·92	7·07	1·90	0·07	11·26	1·60	0·21	8·91
		(0·04)	(0·59)	(0·34)	(0·03)	(0·34)	(0·21)	(0·03)	(0·15)	(0·14)	(0·03)	(0·03)	(0·20)	(0·03)	(0·05)	(0·20)
	4	1·22	3·37	14·10	1·07	2·14	12·64	1·28	1·95	14·33	1·68	0·05	20·80	1·49	0·15	18·34
		(0·04)	(0·32)	(0·52)	(0·03)	(0·19)	(0·39)	(0·03)	(0·14)	(0·20)	(0·02)	(0·02)	(0·20)	(0·03)	(0·04)	(0·27)
10	5	1·21	1·08	23·34	1·06	0·98	20·58	1·23	1·02	21·19	1·42	0·02	26·66	1·30	0·08	24·13
		(0·03)	(0·18)	(0·56)	(0·03)	(0·12)	(0·47)	(0·03)	(0·09)	(0·20)	(0·02)	(0·01)	(0·22)	(0·02)	(0·03)	(0·31)
	6	1·66	0·00	44·60	1·66	0·00	44·30	2·08	0·00	38·05	2·10	0·00	17·29	1·95	0·00	9·94
		(0·01)	(0·00)	(0·16)	(0·01)	(0·00)	(0·18)	(0·02)	(0·00)	(0·20)	(0·04)	(0·00)	(1·01)	(0·05)	(0·00)	(0·79)
	7	0·71	0·73	7·61	0·69	2·14	5·82	0·97	3·37	4·77	0·78	0·07	8·79	0·81	0·23	8·29
		(0·01)	(0·15)	(0·21)	(0·02)	(0·24)	(0·25)	(0·03)	(0·16)	(0·16)	(0·01)	(0·03)	(0·05)	(0·02)	(0·05)	(0·09)
	8	0·89	19·24	0·02	0·65	5·81	0·03	0·93	3·58	0·00	6·83	0·06	4·50	4·03	0·25	2·55
		(0·04)	(0·63)	(0·02)	(0·03)	(0·30)	(0·02)	(0·02)	(0·19)	(0·00)	(0·23)	(0·02)	(0·16)	(0·18)	(0·06)	(0·13)

MB, method of Meinshausen & Bühlmann (2006); SIN (0·05), method of Drton & Perlman (2004) based on a cut-off of 0·05; SIN (0·25), method of Drton & Perlman (2004) based on a cut-off of 0·25; KL, Kullback-Leibler loss defined in (6); FP, number of incorrectly identified edges; FN, number of incorrectly missed edges.

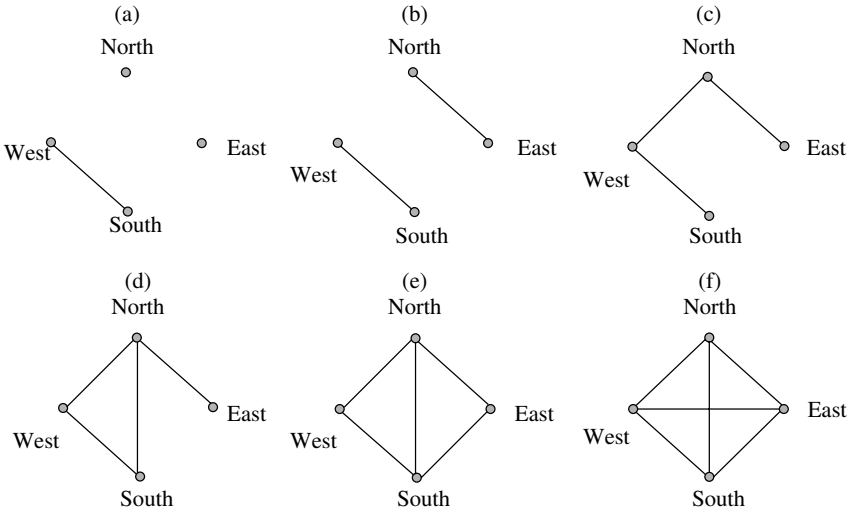

Fig. 3. Cork borings dataset. The Meinshausen–Bühlmann method selects (d); both methods based on SIN, with cut-off 0·05 and 0·25, select (b); both Lasso and Garrote with BIC select (e).

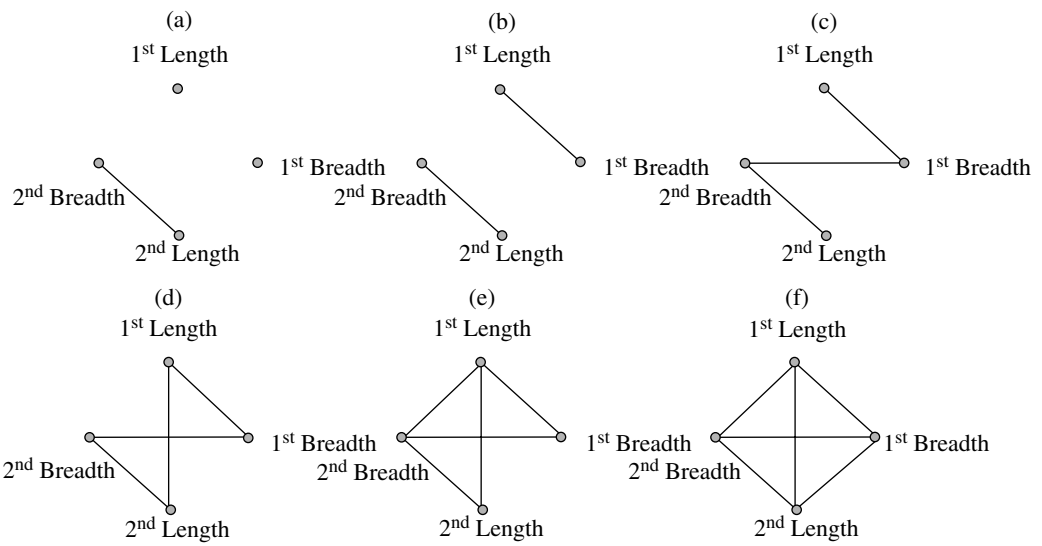

Fig. 4. Fret's heads dataset. The Meinshausen–Bühlmann method selects (f); SIN with cut-off 0·05 selects (a); SIN with cut-off 0·25 selects (b); both Lasso and Garrote with BIC select (f).

using the data from 2003. The KL loss of the estimates are then evaluated using the data from 2004, and Table 3 reports the improved KL loss over the sample covariance matrix.

As shown in Tables 2 and 3, the proposed penalized likelihood methods enjoy very competitive performance.

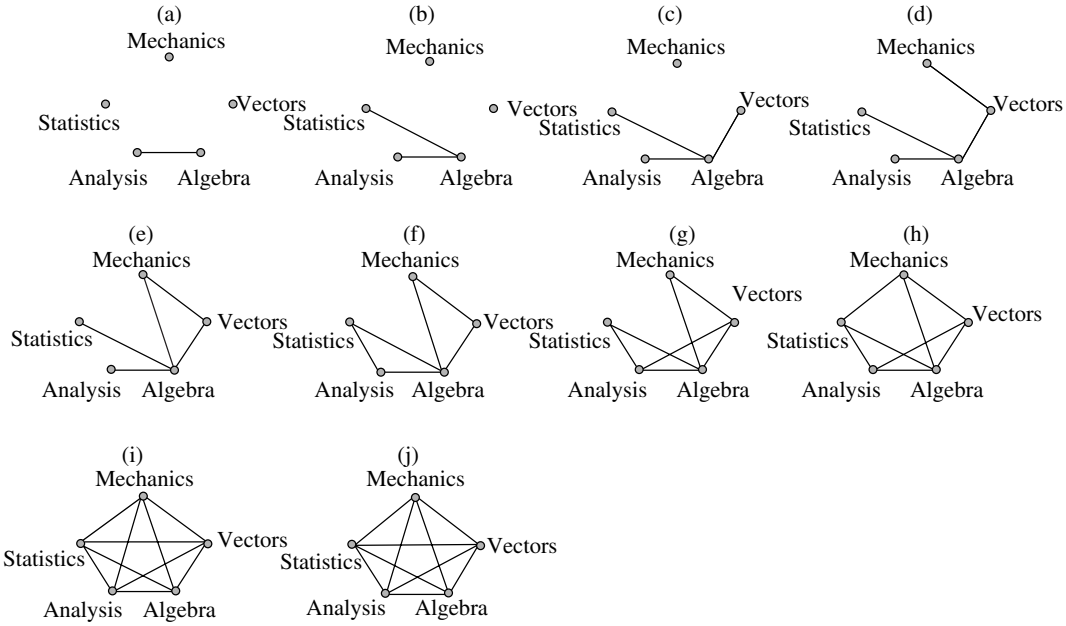

Fig. 5. Mathematics marks dataset. The Meinshausen–Bühlmann method selects (f); SIN with cut-off 0·05 selects (b) with an additional edge between Mechanics and Vectors; SIN with cut-off 0·25 selects (f); Lasso with BIC selects (j); and Garrote with BIC selects (f).

Table 2. *Small-scale examples. Averaged* KL *loss estimated by fivefold crossvalidation*

Dataset	Lasso	Garrote	MB	SIN (0·05)	SIN (0·25)	Sample
Cork borings	21·65	22·28	22·46	25·21	24·45	22·68
Fret's heads	18·68	18·33	20·15	21·10	21·22	20·00
Maths marks	29·52	29·53	29·83	30·66	30·26	29·84

MB, method of Meinshausen & Bühlmann (2006); SIN (0·05), method of Drton & Perlman (2004) based on a cut-off of 0·05; SIN (0·25), method of Drton & Perlman (2004) based on a cut-off of 0·25.

Table 3. *Stock market example. Improvement of predictive* KL *loss over sample covariance matrix*

Dataset	Lasso	Garrote	MB	SIN (0·05)	SIN (0·25)
Stock Market	0·05	0·16	−0·58	−5·89	−4·81

Acknowledgement

The authors wish to thank the editor and two anonymous referees for comments that greatly improved the paper. This research was supported in part by grants from the U.S. National Science Foundation.

APPENDIX

Proofs

Proof of Lemma 1. The first-order condition leads to

$$-\frac{c_{22}}{c_{11}c_{22}-c_{12}^2}+\frac{1}{1-r^2}=0, \tag{A1}$$

$$-\frac{c_{11}}{c_{11}c_{22}-c_{12}^2}+\frac{1}{1-r^2}=0, \tag{A2}$$

$$\frac{2c_{12}}{c_{11}c_{22}-c_{12}^2}-\frac{2r}{1-r^2}+2\lambda\text{sign}(c_{12})=0, \tag{A3}$$

where $\text{sign}(c_{12})=1$ if $c_{12}>0$, $\text{sign}(c_{12})=-1$ if $c_{12}<-1$, and $\text{sign}(c_{12})$ is anywhere between -1 and 1 if $c_{12}=0$. Equation (7) can be easily obtained from (A1) and (A2). Together with (A3), we conclude that

$$\frac{2c_{12}}{\frac{1}{2}(1-r^2)\left[(1-r^2)+\sqrt{\{(1-r^2)^2+4c_{12}^2\}}\right]}+2\lambda\text{sign}(c_{12})=\frac{2r}{1-r^2}. \tag{A4}$$

The sign of the left-hand side of the above equation, $\text{sign}(c_{12})$, should therefore be equal to the sign of the right-hand side, $\text{sign}(r)$. It follows that (A4) implies that

$$\frac{|c_{12}|}{\frac{1}{2}\left[(1-r^2)+\sqrt{\{(1-r^2)^2+4c_{12}^2\}}\right]}=\frac{\sqrt{\{(1-r^2)^2+4c_{12}^2\}}-(1-r^2)}{2|c_{12}|}=|r|-\lambda(1-r^2). \tag{A5}$$

The proof can be completed by the solution of (A5).

□

Proof of Theorem 1. Define $V_n(U)$ as

$$\begin{aligned}V_n(U) &= -\log\left|C+\frac{U}{\sqrt{n}}\right|+\text{tr}\left\{\left(C+\frac{U}{\sqrt{n}}\right)\bar{A}\right\}+\lambda\sum_{i\neq j}\left|c_{ij}+\frac{u_{ij}}{\sqrt{n}}\right|\\ &\quad+\log|C|-\text{tr}(C\bar{A})+\lambda\sum_{i\neq j}|c_{ij}|.\end{aligned}$$

Note that

$$\log\left|C+\frac{U}{\sqrt{n}}\right|-\log|C|=\log\left|I+\frac{\Sigma^{1/2}U\Sigma^{1/2}}{\sqrt{n}}\right|=\sum_{i=1}^{p}\log\left\{1+\sigma_i(\Sigma^{1/2}U\Sigma^{1/2})/\sqrt{n}\right\},$$

where $\sigma_i(\cdot)$ denotes the ith-largest eigenvalue of a matrix. Since

$$\log\left\{1+\sigma_i(\Sigma^{1/2}U\Sigma^{1/2})/\sqrt{n}\right\}=\frac{\sigma_i(\Sigma^{1/2}U\Sigma^{1/2})}{\sqrt{n}}-\frac{\sigma_i^2(\Sigma^{1/2}U\Sigma^{1/2})}{n}+o\left(\frac{1}{n}\right),$$

we conclude that

$$\begin{aligned}\log\left|C+\frac{U}{\sqrt{n}}\right|-\log|C| &= \sum_i\frac{\sigma_i(\Sigma^{1/2}U\Sigma^{1/2})}{\sqrt{n}}-\frac{\text{tr}(\Sigma^{1/2}U\Sigma U\Sigma^{1/2})}{n}+o\left(\frac{1}{n}\right)\\ &= \frac{\text{tr}(\Sigma^{1/2}U\Sigma^{1/2})}{\sqrt{n}}-\frac{\text{tr}(\Sigma^{1/2}U\Sigma U\Sigma^{1/2})}{n}+o\left(\frac{1}{n}\right)\\ &= \frac{\text{tr}\,(U\Sigma)}{\sqrt{n}}-\frac{\text{tr}\,(U\Sigma U\Sigma)}{n}+o\left(\frac{1}{n}\right).\end{aligned}$$

On the other hand,

$$\operatorname{tr}\left\{\left(C+\frac{U}{\sqrt{n}}\right)\bar{A}\right\}-\operatorname{tr}(C\bar{A})=\operatorname{tr}\left(\frac{U\bar{A}}{\sqrt{n}}\right)$$
$$=\frac{\operatorname{tr}(U\Sigma)}{\sqrt{n}}+\frac{\operatorname{tr}\left\{U(\bar{A}-\Sigma)\right\}}{\sqrt{n}}.$$

Together with the fact that

$$\lambda\sum_{i\neq j}\left(\left|c_{ij}+\frac{u_{ij}}{\sqrt{n}}\right|-|c_{ij}|\right)=\frac{\lambda}{\sqrt{n}}\sum_{i\neq j}\left\{u_{ij}\operatorname{sign}(c_{ij})I(c_{ij}\neq 0)+|u_{ij}|I(c_{ij}=0)\right\},$$

$nV_n(U)$ can be re-written as

$$\operatorname{tr}(U\Sigma U\Sigma)+\operatorname{tr}(UW_n)+\sqrt{n}\lambda\sum_{i\neq j}\left\{u_{ij}\operatorname{sign}(c_{ij})I(c_{ij}\neq 0)+|u_{ij}|I(c_{ij}=0)\right\}+o(1),$$

where $W_n=\sqrt{n}(\bar{A}-\Sigma)\to\mathcal{N}(0,\Lambda)$. Therefore, $nV_n(U)\to V(U)$, in distribution. Since both $V(U)$ and $nV_n(U)$ are convex and $V(U)$ has a unique minimum, it follows that,

$$\arg\min nV_n(U)=\sqrt{n}(\hat{C}-C)\to\arg\min V(U).\qquad\square$$

Proof of Theorem 2. The proof proceeds in the same fashion as that of Theorem 1. Define $V_n(U)$ as

$$V_n(U)=-\log\left|C+\frac{U}{\sqrt{n}}\right|+\operatorname{tr}\left\{\left(C+\frac{U}{\sqrt{n}}\right)\bar{A}\right\}+\lambda\sum_{i\neq j}\frac{c_{ij}+u_{ij}/\sqrt{n}}{\tilde{c}_{ij}}$$
$$+\log|C|-\operatorname{tr}(C\bar{A})+\lambda\sum_{i\neq j}\frac{c_{ij}}{\tilde{c}_{ij}}.$$

As before,

$$nV_n(U)=\operatorname{tr}(U\Sigma U\Sigma)+\operatorname{tr}(UW_n)+\sqrt{n}\lambda\sum_{i\neq j}\frac{u_{ij}}{\tilde{c}_{ij}}+o(1).$$

Note that $\tilde{c}_{ij}=O_p(n^{-1/2})$ if $c_{ij}=0$, $\tilde{c}_{ij}\to c_{ij}$ in probability and $\sqrt{n}\lambda\to 0$. Therefore, the above expression can be rewritten as

$$nV_n(U)=\operatorname{tr}(U\Sigma U\Sigma)+\operatorname{tr}(UW_n)+n\lambda\sum_{c_{ij}=0}\frac{u_{ij}}{\tilde{c}_{ij}\sqrt{n}}+o(1).$$

Since $n\lambda\to\infty$, we conclude that the minimizer of $nV_n(U)$ satisfies $u_{ij}=0$ if $c_{ij}=0$ with probability tending to one. The proof is now completed if we note that the maximum likelihood estimator $\hat{C}^{\text{true}}$ for the true graph $(V, E=(c_{ij}\neq 0))$ is such that

$$\sqrt{n}(\hat{C}^{\text{true}}-C)\to\arg\min\left\{\operatorname{tr}(U\Sigma U\Sigma)+\operatorname{tr}(UW)\right\},$$

in distribution, where the minimum is taken over all symmetric matrices U such that $u_{ij}=0$ if $c_{ij}=0$. $\square$

Proof of Lemma 2. Simple matrix calculus shows that the matrix of second derivatives of the objective function in (8) is positive definite and therefore the objective function is strictly convex. Since the feasible region is compact, $\hat{C}_{\text{new}}$ is always well defined. We now show that the algorithm will terminate in a finite number of iterations. Note that, at each iteration, $\hat{C}_{\text{old}}$ lies in the feasible region of Step 2. If the algorithm does not terminate, that is, at each step $\hat{C}_{\text{new}}\neq\hat{C}_{\text{old}}$, then the minimum attained at Step 2 is strictly smaller than that from the previous iteration. The minima attained in the iterations form a strictly decreasing sequence, which in turn implies that the sign matrix in (8) must

be different for all iterations. However, this contradicts the fact that there are only a finite number, $2^{p(p-1)/2}$, of possible choices for the sign matrix S. Therefore the algorithm has to terminate.

Now we show that the algorithm converges to the solution to (3). Denote the solution at convergence of the algorithm by $\hat{C}$. By the algorithm we see there exist two sign matrices $\hat{S}$ and $\tilde{S}$, with $\hat{s}_{ij}\hat{c}_{ij} \geqslant 0$, $\tilde{s}_{ij}\hat{c}_{ij} \geqslant 0$, and $\hat{s}_{ij} = -\tilde{s}_{ij}$ for any $\hat{c}_{ij} = 0$, such that $\hat{C}$ solves (8) with both $\hat{S}$ and $\tilde{S}$. Let $l(C) = -\log|C| + \mathrm{tr}(C\bar{A})$. Then, by the Karush-Kuhn-Tucker conditions (Boyd & Vandenberghe, 2003), there exist $\lambda_1 > 0$ and $\lambda_2 > 0$ such that

$$\left.\frac{\partial l}{\partial c_{ij}}\right|_{C=\hat{C}} \hat{s}_{ij} = -\lambda_1 \qquad \text{for all } \hat{c}_{ij} \neq 0 \tag{A6}$$

$$\left.\frac{\partial l}{\partial c_{ij}}\right|_{C=\hat{C}} \hat{s}_{ij} \geqslant -\lambda_1 \qquad \text{for all } \hat{c}_{ij} = 0, \tag{A7}$$

and

$$\left.\frac{\partial l}{\partial c_{ij}}\right|_{C=\hat{C}} \tilde{s}_{ij} = -\lambda_2 \qquad \text{for all } \hat{c}_{ij} \neq 0 \tag{A8}$$

$$\left.\frac{\partial l}{\partial c_{ij}}\right|_{C=\hat{C}} \tilde{s}_{ij} \geqslant -\lambda_2 \qquad \text{for all } \hat{c}_{ij} = 0. \tag{A9}$$

Together with the fact that $\hat{s}_{ij} = \tilde{s}_{ij}$ for any $\hat{c}_{ij} \neq 0$, (A6) and (A8) imply that $\lambda_1 = \lambda_2 \equiv \lambda$. Combining this with (A7) and (A9), we conclude that

$$\left.\frac{\partial l}{\partial c_{ij}}\right|_{C=\hat{C}} = -\lambda\hat{s}_{ij} \qquad \text{for all } \hat{c}_{ij} \neq 0$$

$$-\lambda \leqslant \left.\frac{\partial l}{\partial c_{ij}}\right|_{C=\hat{C}} \leqslant \lambda \qquad \text{for all } \hat{c}_{ij} = 0,$$

which implies that $\hat{C}$ is also the solution to (4), and equivalently (3), again by the Karush-Kuhn-Tucker conditions. □

Proof of Lemma 3. Let $B = \bar{A}^{-1}$. Then $b_{ii} = 1$ according to our scaling and $b_{i,-i}$ is the least-squares estimator corresponding to regressing $X^{(i)}$ on the other elements; see Lauritzen (1996) and Meinshausen & Bühlmann (2006). Using this fact, we may write (11) as

$$\frac{1}{n}\sum_{i=1}^{p} \left|\left|X^{[-i]}b_{i,-i} - X^{[-i]}\theta_{i,-i}\right|\right|^2 + \lambda \sum_{i \neq j} |\theta_{ij}|.$$

To minimize this function, we have $\theta_{ii} = 1$ and $\theta_{i,-i}$ as the minimizer of (10). □

REFERENCES

BOYD, S. & VANDENBERGHE, L. (2003). *Convex Optimization*. Cambridge: Cambridge University Press.

BREIMAN, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373–84.

BREIMAN, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.* **24**, 2350–83.

COX, D. R. & WERMUTH, N. (1996). *Multivariate Dependencies: Models, Analysis and Interpretation*. London: Chapman and Hall.

DEMPSTER, A. P. (1972). Covariance selection. *Biometrika* **32**, 95–108.

DRTON, M. & PERLMAN, M. (2004). Model selection for Gaussian concentration graphs. *Biometrika* **91**, 591–602.

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

KNIGHT, K. & FU, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28**, 1356–78.

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

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

MARDIA, K. V., KENT, J. T. & BIBBY, J. M. (1979). *Multivariate Analysis*. London: Academic Press.

MEINSHAUSEN, N. & BÜHLMANN, P. (2006). High-dimensional graphs with the Lasso. *Ann. Statist.* **34**, 1436–62.
RAO, C. (1948). Tests of significance in multivariate analysis. *Biometrika*, **35**, 58–79.
TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* **58**, 267–88.
VANDENBERGHE, L., BOYD, S. & WU, S.-P. (1998). Determinant maximization with linear matrix inequality constraints. *SIAM J. Matrix Anal. Appl.* **19**, 499–533.
WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester: John Wiley and Sons.

[*Received January 2006. Revised August 2006*]