

Model Selection and Estimation in the Gaussian Graphical Model¹

Ming Yuan and Yi Lin

(June 1, 2005)

Abstract

We propose a penalized likelihood method for estimating the concentration matrix in the Gaussian graphical model. The method leads to a sparse and shrinkage estimate of the concentration matrix that is positive definite, thus conducts model selection and estimation simultaneously in the Gaussian graphical model. The implementation of the method is nontrivial due to 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 method. The connection between our method and existing methods is illustrated. Simulations and real examples demonstrate the competitive performance of the new method.

Key words: penalized likelihood, maxdet algorithm, covariance selection, the Lasso.

¹Ming Yuan is Assistant Professor, School of Industrial and System Engineering, Georgia Institute of Technology, Atlanta, GA 30332 (E-mail: myuan@isye.gatech.edu). Yi Lin is Associate Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706 (E-mail: yilin@stat.wisc.edu). Lin's research was supported in part by National Science Foundation grant DMS-0134987.

1 Introduction

We consider the problem of estimating the concentration matrix of a p -dimensional normal distribution from an i.i.d. sample. Let $X = (X^{(1)}, \dots, X^{(p)})$ 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, \dots, 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 $c_{ij} = 0$ in the concentration matrix 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 and 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 vertexes corresponding to each of the p coordinates and the edges $E = (e_{ij})_{1 \leq i < j \leq p}$ describe the conditional independence relationship among $X^{(1)}, \dots, X^{(p)}$. 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, which is equivalent to that the element of the concentration matrix c_{ij} is zero. Thus parameter estimation and model selection in the Gaussian concentration graph model is 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 for model selection in Gaussian graphical models is the 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 (see, for example, Edwards, 2000). Another drawback of the common stepwise procedure is its computational complexity. At each single step the edge selection or deletion requires fitting a large number of candidate models (of the order p^2). This is computationally infeasible for even moderate p 's. To remedy these

problems, Drton and Perlman (2004, 2005) 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 and Bühlmann (2004) proposed a computationally attractive method to covariance selection that can be used for very large Gaussian graphs. They perform neighborhood 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.

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. The ℓ_1 penalty encourages sparsity and at the same time gives shrinkage estimates. Another important feature of our approach is that we explicitly take care of the natural constraint that the concentration matrix is positive definite. Therefore our method leads directly to a sparse and shrinkage estimate of the concentration matrix that is positive definite. This is in contrast with other methods that estimate the concentration matrix based on the model selected, since it is known that shrinkage is beneficial when multiple parameters are to be estimated.

The formulation of our method is introduced in Section 2, and the algorithm is given in Section 3. Solving the optimization problem in our formulation is nontrivial due to the nonlinear objective function and the positive definite constraint. Our algorithm takes advantage of the close connection between our formulation and the so-called maxdet-problem considered in Vandenberghe, Boyd, and Wu (1998). The maxdet-problem is a particular type of convex optimization problem with linear matrix inequalities that can be solved very efficiently. In our algorithm we make use of the efficient interior point algorithm developed by Vandenberghe, Boyd, and Wu (1998) for the maxdet-problem. In Section 3 we also introduce a “nonnegative garrote” method that is closely related to the penalized likelihood

method and propose a BIC type criterion for the selection of the tuning parameter in the methods.

There is a connection between the neighborhood selection method in Meinshausen and Bühlmann (2004) and our penalized likelihood approach, which we illustrate in Section 4. The neighborhood 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 neighborhood selection method is computationally faster due to its simpler form and that it does not consider the positive definite constraint. Our method is more efficient due to the incorporation of the positive definite constraint and the use of likelihood. Results from a simulation study comparing our penalized likelihood method with other methods are reported in Section 5. We can see that the penalized likelihood method gives very competitive performance. Some real examples are given in Section 6. We conclude with some discussions in Section 7.

2 Methodology

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

The log-likelihood for μ and $C = \Sigma^{-1}$ based on an i.i.d. sample X_1, \dots, X_n of X is

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

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

$$\bar{A} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})' (X_i - \bar{X}). \quad (2)$$

The commonly used sample covariance matrix is $S = n\bar{A}/(n-1)$. The concentration matrix C can be naturally estimated with \bar{A}^{-1} or S^{-1} . However, due to the large number of unknown

parameters to be estimated (the total number of parameters is $p(p+1)/2$), S is not a stable estimate of Σ for moderate or large p 's. In general, the matrix S^{-1} is positive definite when $n \geq p$, but does not lead to “sparse” graph structure since the matrix typically contains no zero entry.

In order to achieve a “sparse” graph structure and to give a better estimate of the concentration matrix, we adapt the idea of the Lasso in linear regression models to the covariance selection problem. We find the minimizer $(\hat{\mu}, \hat{C})$ of

$$-\ln |C| + \frac{1}{n} \sum_{i=1}^n (X_i - \mu)' C (X_i - \mu) \quad \text{subject to} \quad \sum_{i \neq j} |c_{ij}| \leq t, \quad (3)$$

over the set of positive definite matrices. Here $t \geq 0$ is a tuning parameter. When $t = \infty$, the solution to (3) is the maximum likelihood estimate \bar{A}^{-1} given that the inverse exists. On the other hand, if $t = 0$, then the constraint forces $X^{(1)}, \dots, X^{(p)}$ to be mutually independent. It is clear that $\hat{\mu} = \bar{X}$ regardless of t . Since the observations are centered, we have $\hat{\mu} = \mathbf{0}$. Therefore, \hat{C} is the positive definite matrix that minimizes

$$-\ln |C| + \frac{1}{n} \sum_{i=1}^n (X_i)' C X_i \quad \text{subject to} \quad \sum_{i \neq j} |c_{ij}| \leq t. \quad (4)$$

We can further rewrite (4) as

$$-\ln |C| + \text{trace}(C\bar{A}) \quad \text{subject to} \quad \sum_{i \neq j} |c_{ij}| \leq t. \quad (5)$$

An equivalent formulation to (5) is the penalized likelihood

$$-\ln |C| + \text{trace}(C\bar{A}) + \lambda \sum_{i \neq j} |c_{ij}|, \quad (6)$$

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

3 Computation

The nonlinearity of the objective function and the positive definite constraint make the optimization problem (5) nontrivial. We take advantage of the connection between (5) and the determinant maximization problem (maxdet problem) well studied in the convex

programming literature. The maxdet problem (Vandenberghe, Boyd, and Wu, 1998) is a special type of convex optimization problem that can be solved very efficiently with the interior point algorithm. It is of the form

$$\begin{aligned}
& \min_{x \in R^m} && b'x - \ln |G(x)| \\
& \text{subject to} && G(x) \text{ is positive definite} \\
& && F(x) \text{ is positive semi-definite}
\end{aligned} \tag{7}$$

where $b \in R^m$ and the functions $G : R^m \rightarrow R^{l \times l}$ and $F : R^m \rightarrow R^{l \times l}$ are affine:

$$\begin{aligned}
G(x) &= G_0 + x_1 G_1 + \dots + x_m G_m, \\
F(x) &= F_0 + x_1 F_1 + \dots + x_m F_m
\end{aligned}$$

where F_i and G_i are symmetric matrices. Vandenberghe, Boyd, and Wu (1998) developed an effective algorithm for this problem. To use their algorithm, it is also required that $F_i, i = 1, \dots, m$ and $G_i, i = 1, \dots, m$ are linearly independent respectively.

3.1 Algorithm

It is not hard to see that if the signs of c'_{ij} s are known, (5) can be cast as a maxdet-problem. More specifically, (5) can be expressed as

$$\begin{aligned}
& \min_C && 2 \sum_{i < j} a_{ij} c_{ij} + \sum_i a_{ii} c_{ii} - \ln \left| \sum_i c_{ii} I^{(i)} + \sum_{i < j} c_{ij} I^{(ij)} \right| \\
& \text{subject to} && \sum_i c_{ii} I^{(i)} + \sum_{i < j} c_{ij} I^{(ij)} \text{ is positive definite} \\
& && t - 2 \sum_{i < j} c_{ij} s_{ij} \geq 0, \quad s_{ij} c_{ij} \geq 0
\end{aligned} \tag{8}$$

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

- (1) Initialize $\hat{C}_{old} = \bar{A}^{-1}$ and $s = \text{sign}(\bar{A}^{-1})$
- (2) Solve (8) over positive definite matrices and denote the solution by \hat{C}_{new} .
- (3) Check if $\hat{C}_{new} = \hat{C}_{old}$.
 - (i) If true, return $\hat{C} = \hat{C}_{new}$.
 - (ii) If not, set $\hat{C}_{old} = \hat{C}_{new}$ and $s_{ij} = -s_{ij}$ for any pair (i, j) such that $\hat{c}_{ij} = 0$ and go back to step 2.

In our experience the algorithm usually converges within several iterations. Clearly, other initial values for s can also be used. We have the following

Lemma 1 *The above algorithm always converges and converges to the solution to (5).*

Proof. We first show that the algorithm will terminate in finite iterations. Notice that the objective function in (8) is strictly convex, and that at each iteration, \hat{C}_{old} lies in the feasible region of Step 2. Now if the algorithm does not terminate, that is, at each step $\hat{C}_{new} \neq \hat{C}_{old}$, 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 with the fact that there are only a finite total of $2^{p(p-1)/2}$ possible choices of the sign matrix S . Therefore the algorithm has to terminate.

Now we show that the algorithm converges to the solution to (5). Denote the solution at the convergence of the algorithm by \hat{C} . By the algorithm we see there exists two sign matrices \hat{S} and \tilde{S} , with $\hat{s}_{ij}\hat{c}_{ij} \geq 0$, $\tilde{s}_{ij}\hat{c}_{ij} \geq 0$, 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} . Denote $l(C) = -\ln|C| + \text{trace}(C\bar{A})$. Then by the Karush-Kuhn-Tucker conditions (Boyd and Vandenberghe, 2003), there exist $\lambda_1 > 0$ and $\lambda_2 > 0$, such that

$$\left. \frac{\partial l}{\partial c_{ij}} \right|_{C=\hat{C}} \hat{s}_{ij} = -\lambda_1 \quad \forall \hat{c}_{ij} \neq 0 \quad (9)$$

$$\left. \frac{\partial l}{\partial c_{ij}} \right|_{C=\hat{C}} \hat{s}_{ij} \geq -\lambda_1 \quad \forall \hat{c}_{ij} = 0. \quad (10)$$

and

$$\left. \frac{\partial l}{\partial c_{ij}} \right|_{C=\hat{C}} \tilde{s}_{ij} = -\lambda_2 \quad \forall \hat{c}_{ij} \neq 0 \quad (11)$$

$$\left. \frac{\partial l}{\partial c_{ij}} \right|_{C=\hat{C}} \tilde{s}_{ij} \geq -\lambda_2 \quad \forall \hat{c}_{ij} = 0. \quad (12)$$

Together with the fact that for any $\hat{c}_{ij} \neq 0$, $\hat{s}_{ij} = \tilde{s}_{ij}$, (9) and (11) imply $\lambda_1 = \lambda_2$. Combining with (10) and (12), we conclude that \hat{C} is also the solution to (5) again by the Karush-Kuhn-Tucker conditions. ■

3.2 Extension

Most of the time consumed by the proposed algorithm is spent on estimating $\text{sign}(C)$. In the case where a relatively reliable estimate of C is available, a simpler method can be utilized. Let \hat{C} be such a preliminary estimate. A shrinkage estimate can be defined through $c_{ij} = d_{ij}\hat{c}_{ij}$, where the symmetric matrix $D = (d_{ij})$ is the minimizer of

$$-\ln|C| + \text{trace}(C\bar{A}) \quad \text{subject to} \quad \sum_{i \neq j} d_{ij} \leq t, d_{ij} \geq 0, \quad (13)$$

and C is positive definite. For a relatively large sample size, \bar{A}^{-1} is an obvious choice for the preliminary estimate. This estimation procedure is in spirit similar to the nonnegative garrote estimate proposed by Breiman (1995) for linear regression. (13) is a maxdet-problem and can be solved directly using the algorithm of Vandenberghe, Boyd, and Wu (1998).

3.3 Tuning

So far we have concentrated on the calculation of the minimizer of (5) 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 multi-fold cross-validation score. A computationally more efficient alternative is the BIC for model selection and estimation. Similar to the Lasso in linear regression, in which the degrees of freedom can be defined as the number of nonzero coefficients in the solution (Zou, Hastie, and Tibshirani, 2004), the degrees of freedom in our problem can be naturally defined as the number of

nonzero entries in the solution. This also corresponds to the number of edges in the estimated graph. More specifically, for any tuning parameter t , we define the corresponding BIC score as

$$\text{BIC}(t) = -\ln |\widehat{C}(t)| + \text{trace}(\widehat{C}(t)\bar{A}) + \frac{\ln n}{n} \sum_{i \leq j} \widehat{e}_{ij}(t), \quad (14)$$

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

4 Quadratic Approximation

Assuming that \bar{A} is nonsingular, a second order approximation to the objective function of (5) around \bar{A}^{-1} can be written as (Boyd and Vandenberghe, 2003):

$$\text{trace} \left((C - \bar{A}^{-1})\bar{A}(C - \bar{A}^{-1})\bar{A} \right). \quad (15)$$

Therefore the solution to (6) can be approximated by the solution to

$$\text{trace} \left((C - \bar{A}^{-1})\bar{A}(C - \bar{A}^{-1})\bar{A} \right) + \lambda |C|_{\ell_1} \quad (16)$$

This second order approximation is closely connected to the approach proposed by Meinshausen and Bühlmann (2004), hereafter referred to as MB. In MB's approach, for each $i = 1, \dots, p$, we seek the minimizer $\widehat{\theta}_{i,-i} = (\widehat{\theta}_{i1}, \dots, \widehat{\theta}_{i(i-1)}, \widehat{\theta}_{i(i+1)}, \dots, \widehat{\theta}_{ip}) \in R^{p-1}$ to

$$\frac{1}{n} \left\| X^{(i)} - X^{[-i]} \theta_{i,-i} \right\|^2 + \lambda \sum_{j \neq i} |\theta_{ij}|, \quad (17)$$

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

Note that θ_{ii} , $i = 1, \dots, p$, are not determined. For notational purpose, we write $\theta_{ii} = 1$ for $i = 1, \dots, p$. Recall that we scale each component of X so that all the diagonal elements of the sample concentration matrix are one. The following lemma reveals a close connection between MB's approach and the second order approximation (16).

Lemma 2 Matrix $\Theta = (\theta_{ij})$ defined by (17) is the unconstrained solution to

$$\min_C \text{trace} \left((C - \bar{A}^{-1})' \bar{A} (C - \bar{A}^{-1}) \right) + \lambda |C|_{\ell_1}, \quad (18)$$

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

Proof. Denote $B = \bar{A}^{-1}$. Then $b_{ii} = 1$ according to our scaling and $b_{i,-i}$ is the least square estimate by regressing $X^{(i)}$ on the other elements (Lauritzen, 1996; Meinshausen and Bühlmann, 2004). Using this fact, (18) can be written as

$$\frac{1}{n} \sum_{i=1}^p \left\| X^{[-i]} b_{i,-i} - X^{[-i]} \theta_{i,-i} \right\|^2 + \lambda \sum_{i \neq j} |\theta_{ij}| \quad (19)$$

To minimize this function, we have $\theta_{ii} = 1$ and $\theta_{i,-i}$ is the minimizer of (17). ■

From Lemma 2, we see that MB's approach seeks a sparse C close to \bar{A}^{-1} . However, it does not incorporate the symmetry and positive definite constraint in the estimation of the concentration matrix, therefore an additional step is needed to estimate either the covariance matrix or the concentration matrix. Also, the loss function utilized by Meinshausen and Bühlmann is different from the quadratic approximation to the log-likelihood, therefore the approach is expected to be less efficient than our penalized likelihood method or the corresponding quadratic approximation (16).

5 Simulation

We consider eight different models in our simulation.

- (1) Heterogeneous: $\Sigma = \text{diag}(1, 2, \dots, n)$
- (2) AR(1): $c_{ii} = 1$ and $c_{i,i-1} = c_{i-1,i} = 0.5$
- (3) AR(2): $c_{ii} = 1$, $c_{i,i-1} = c_{i-1,i} = 0.5$ and $c_{i,i-2} = c_{i-2,i} = 0.25$
- (4) AR(3): $c_{ii} = 1$, $c_{i,i-1} = c_{i-1,i} = 0.4$ and $c_{i,i-2} = c_{i-2,i} = c_{i,i-3} = c_{i-3,i} = 0.2$
- (5) AR(4): $c_{ii} = 1$, $c_{i,i-1} = c_{i-1,i} = 0.4$, $c_{i,i-2} = c_{i-2,i} = c_{i,i-3} = c_{i-3,i} = 0.2$ and $c_{i,i-4} = c_{i-4,i} = 0.1$

(6) Full model: $c_{ij} = 2$ if $i = j$ and 1 otherwise

(7) Star: every node is connect with the first node. $c_{ii} = 1$, $c_{1,i} = c_{i,1} = 0.2$ and 0 otherwise

(8) Circle: $c_{ii} = 1$, $c_{i,i-1} = c_{i-1,i} = 0.5$ and $c_{1n} = c_{n1} = 0.4$

For each model, we simulated samples with size 25 and dimension $p = 5$, or size 50 and dimension 10. We compare our methods with MB's approach and the method proposed by Drton and Perlman (2004) (hereafter referred to as DP) in terms of the Kullback-Leibler loss (KL)

$$-\ln|\hat{C}| + \text{trace}(\hat{C}\Sigma) - (-\ln|\Sigma^{-1}| + p), \quad (20)$$

the number of false positives (FP) and the number of false negatives (FN). MB's approach was implemented using the LARS package from R and the method by DP has also been implemented in the SIN package of R. The method of DP 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 MB's method and DP's method 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 due to its connection to the idea of the Lasso in linear regression. Similarly, the extension described in Section 3.2 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. MB's method and both versions of DP's method tend to have more false negatives, which may partly explain their relatively poor performance. It is also suggested, though, by the simulation that the proposed penalized likelihood approach combined with BIC may have relatively more false positives. Because of this, the solution path of (5) may be more informative in determining the graph structure. For example, Figures 1 provides the graphs visited by (5) when the tuning parameter t decreases from $+\infty$ to 0 for a typical dataset generated from the AR(1) model.

p	Model	Lasso			Garrote			MB			SIN (0.05)			SIN (0.25)		
		KL	FP	FN	KL	FP	FN	KL	FP	FN	KL	FP	FN	KL	FP	FN
5	1	0.27 (0.02)	0.20 (0.06)	0.00 (0.00)	0.31 (0.02)	0.42 (0.08)	0.00 (0.00)	0.45 (0.04)	0.91 (0.08)	0.00 (0.00)	0.26 (0.02)	0.05 (0.03)	0.00 (0.00)	0.32 (0.03)	0.26 (0.07)	0.00 (0.00)
	2	0.70 (0.05)	3.31 (0.12)	0.07 (0.03)	0.67 (0.05)	1.20 (0.12)	0.14 (0.04)	0.63 (0.05)	0.68 (0.08)	0.16 (0.04)	1.88 (0.06)	0.03 (0.02)	2.47 (0.10)	1.40 (0.06)	0.15 (0.05)	1.59 (0.09)
	3	0.89 (0.05)	1.29 (0.10)	2.24 (0.24)	0.87 (0.04)	0.60 (0.08)	2.58 (0.18)	0.98 (0.04)	0.47 (0.06)	3.68 (0.09)	1.16 (0.04)	0.01 (0.01)	5.42 (0.12)	1.06 (0.05)	0.07 (0.04)	4.16 (0.15)
	4	0.79 (0.03)	0.22 (0.04)	5.60 (0.29)	0.80 (0.04)	0.16 (0.04)	5.86 (0.23)	0.83 (0.04)	0.16 (0.04)	6.23 (0.11)	0.93 (0.03)	0.00 (0.00)	8.14 (0.11)	0.90 (0.03)	0.01 (0.01)	6.97 (0.15)
	5	0.78 (0.04)	0.00 (0.00)	7.06 (0.30)	0.76 (0.03)	0.00 (0.00)	6.98 (0.23)	0.80 (0.04)	0.00 (0.00)	7.26 (0.12)	0.88 (0.03)	0.00 (0.00)	9.13 (0.11)	0.86 (0.03)	0.00 (0.00)	7.94 (0.16)
	6	1.09 (0.04)	0.00 (0.00)	4.53 (0.44)	1.11 (0.04)	0.00 (0.00)	4.58 (0.41)	1.30 (0.04)	0.00 (0.00)	7.05 (0.11)	1.28 (0.04)	0.00 (0.00)	6.18 (0.24)	1.18 (0.05)	0.00 (0.00)	3.77 (0.25)
	7	0.45 (0.02)	0.31 (0.08)	3.47 (0.10)	0.51 (0.03)	0.46 (0.08)	3.02 (0.12)	0.61 (0.03)	0.55 (0.07)	2.75 (0.10)	0.43 (0.02)	0.00 (0.00)	3.92 (0.03)	0.50 (0.03)	0.13 (0.05)	3.61 (0.06)
	8	0.73 (0.05)	2.55 (0.13)	0.11 (0.03)	0.77 (0.05)	1.28 (0.12)	0.26 (0.06)	0.80 (0.05)	0.17 (0.05)	0.37 (0.06)	1.89 (0.05)	0.03 (0.02)	3.29 (0.10)	1.48 (0.05)	0.11 (0.04)	2.12 (0.09)
10	1	0.22 (0.01)	0.26 (0.09)	0.00 (0.00)	0.26 (0.01)	0.75 (0.14)	0.00 (0.00)	0.63 (0.03)	3.48 (0.17)	0.00 (0.00)	0.23 (0.01)	0.07 (0.03)	0.00 (0.00)	0.26 (0.02)	0.23 (0.05)	0.00 (0.00)
	2	1.42 (0.04)	31.76 (0.26)	0.00 (0.00)	0.60 (0.02)	4.83 (0.27)	0.00 (0.00)	0.58 (0.02)	2.25 (0.15)	0.00 (0.00)	4.01 (0.14)	0.06 (0.02)	3.75 (0.14)	2.39 (0.12)	0.25 (0.06)	2.05 (0.12)
	3	1.22 (0.04)	10.87 (0.59)	3.30 (0.34)	1.03 (0.03)	5.86 (0.34)	3.05 (0.21)	1.52 (0.03)	2.92 (0.15)	7.07 (0.14)	1.90 (0.03)	0.07 (0.03)	11.26 (0.20)	1.60 (0.03)	0.21 (0.05)	8.91 (0.20)
	4	1.22 (0.04)	3.37 (0.32)	14.10 (0.52)	1.07 (0.03)	2.14 (0.19)	12.64 (0.39)	1.28 (0.03)	1.95 (0.14)	14.33 (0.20)	1.68 (0.02)	0.05 (0.02)	20.80 (0.20)	1.49 (0.03)	0.15 (0.04)	18.34 (0.27)
	5	1.21 (0.03)	1.08 (0.18)	23.34 (0.56)	1.06 (0.03)	0.98 (0.12)	20.58 (0.47)	1.23 (0.03)	1.02 (0.09)	21.19 (0.20)	1.42 (0.02)	0.02 (0.01)	26.66 (0.22)	1.30 (0.02)	0.08 (0.03)	24.13 (0.31)
	6	1.66 (0.01)	0.00 (0.00)	44.60 (0.16)	1.66 (0.01)	0.00 (0.00)	44.30 (0.18)	2.08 (0.02)	0.00 (0.00)	38.05 (0.20)	2.10 (0.04)	0.00 (0.00)	17.29 (1.01)	1.95 (0.05)	0.00 (0.00)	9.94 (0.79)
	7	0.71 (0.01)	0.73 (0.15)	7.61 (0.21)	0.69 (0.02)	2.14 (0.24)	5.82 (0.25)	0.97 (0.03)	3.37 (0.16)	4.77 (0.16)	0.78 (0.01)	0.07 (0.03)	8.79 (0.05)	0.81 (0.02)	0.23 (0.05)	8.29 (0.09)
	8	0.89 (0.04)	19.24 (0.63)	0.02 (0.02)	0.65 (0.03)	5.81 (0.30)	0.03 (0.02)	0.93 (0.02)	3.58 (0.19)	0.00 (0.00)	6.83 (0.23)	0.06 (0.02)	4.50 (0.16)	4.03 (0.18)	0.25 (0.06)	2.55 (0.13)

Table 1: Results for the eight models considered in the simulation.

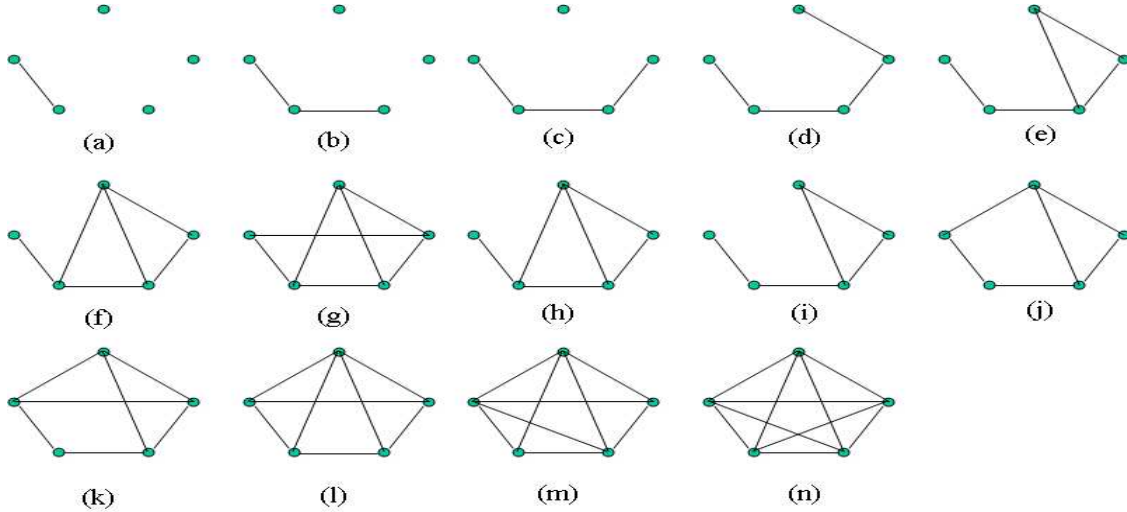


Figure 1: AR(1): MB selects (e); both SIN (0.05) and SIN (0.25) select (b); Lasso with BIC selects (h); Garrote with BIC select (d). The truth is (d)

6 Real World Examples

We first consider three real-world examples. The first is the Cork borings dataset. The data are presented in Whittaker (1990) 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. Another dataset we considered is Fret's heads dataset. The data contain 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, the head breadth of the second son. The data are also presented in Whittaker (1990). The last is the Mathematics marks dataset. Mardia et al. (1979) present 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).

Figures 2-4 depict the solution paths of (5) for each of the three datasets.

To compare the accuracy of different methods, five fold cross-validation was applied on the datasets. Table 2 documents the average KL distances for each method. The KL distance we used is

$$-\ln |\hat{C}| + \text{trace}(\hat{C}\hat{\Sigma}), \quad (21)$$

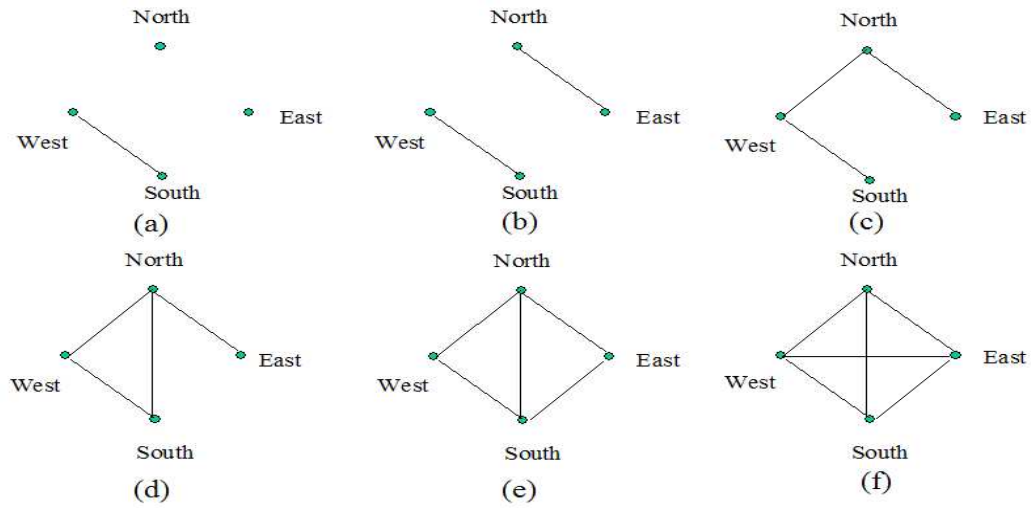


Figure 2: Cork borings dataset: MB selects (d); both SIN (0.05) and SIN (0.25) select (b). Both Lasso and Garrote with BIC select (e).

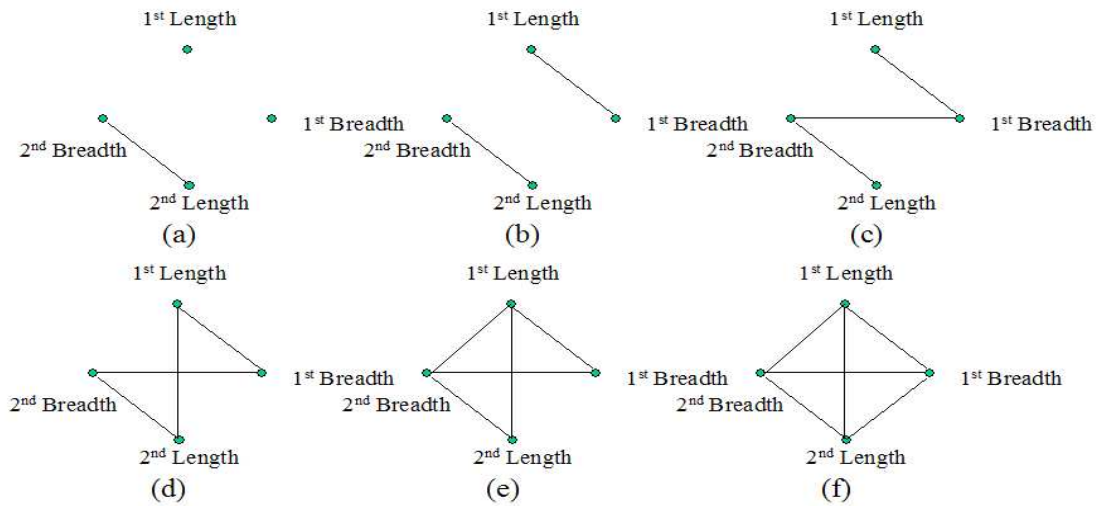


Figure 3: Fret's heads dataset: MB selects (f); SIN (0.05) selects (a); SIN (0.25) selects (b); both Lasso and Garrote with BIC select (f).

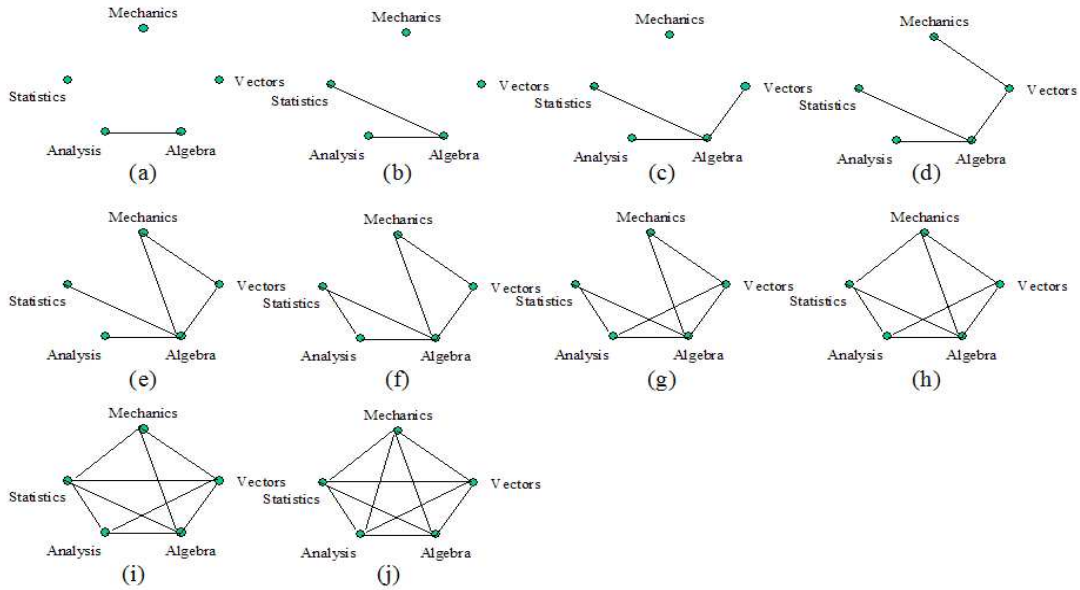


Figure 4: Math marks dataset: MB selects (f); SIN (0.05) selects (b) with an additional edge between Mechanics and Vectors; SIN (0.25) selects (d) with an additional edge between Statistics and Analysis; Lasso with BIC selects (i); and Garrote with BIC selects (j).

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

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
Math marks	29.52	29.53	29.83	30.66	30.26	29.84

Table 2: Averaged KL loss estimated by 5-fold cross-validation.

Next we considered a larger scale problem. The open prices of 35 stocks were collected for Years 2003 and 2004. Different methods were applied to estimate the covariance matrix using the data from Year 2003. The KL loss of the estimates are then evaluated using the data from Year 2004. The following table reports the improved KL loss over the sample covariance matrix.

As shown in Tables 2 and 3, the proposed penalized likelihood methods enjoy very com-

Dataset	Lasso	Garrote	MB	SIN (0.05)	SIN (0.25)
Stock Market	0.05	0.16	-0.58	-5.89	-4.81

Table 3: Improvement of predictive KL loss over sample covariance matrix.

petitive performance.

7 Conclusion

We have introduced a penalized likelihood method for the problem of covariance selection and estimation. This naturally leads to a method for model building in the Gaussian graphical model. The method has connection to some of the existing methods, but is expected to be more efficient. The implementation of our method takes advantage of recent advances in convex optimization.

References

- [1] Boyd, S. and Vandenberghe, L. (2003), *Convex Optimization*, Cambridge University Press, Cambridge.
- [2] Breiman, L. (1995), Better subset regression using the nonnegative garrote, *Technometrics*, **37**, 373-384.
- [3] Cox, D. R. and Wermuth, N. (1996), *Multivariate Dependencies: Models, Analysis and Interpretation*, Chapman & Hall, London.
- [4] Dempster, A. P. (1972), Covariance selection, *Biometrika*, **32**, 95108.
- [5] Drton, M. and Perlman, M. (2004), Model selection for gaussian concentration graphs, *Biometrika* **91(3)**, 591-602.
- [6] Drton, M. and Perlman, M. (2005), A SINful approach to gaussian graphical model selection, *Technical Report*, Department of Statistics, University of Washington.

- [7] Edwards, D. M. (2000), *Introduction to Graphical Modelling*, Springer, New York.
- [8] Efron, B., Johnstone, I., Hastie, T. and Tibshirani, R. (2004), Least angle regression, *Ann. Statist.*, **32** 407-499.
- [9] Lauritzen, S. L. (1996), *Graphical Models*, Clarendon Press, Oxford.
- [10] Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979) *Multivariate Analysis*, Academic Press, London.
- [11] Meinshausen, N. and Bühlmann, P. (2004), Consistent neighbourhood selection for high-dimensional graphs with the Lasso, *Technical Report*, ETH Zürich.
- [12] Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *J. Royal. Statist. Soc. B.*, **58**, 267-288.
- [13] Vandenberghe, L., Boyd, S. and Wu, S.-P., (1998), Determinant maximization with linear matrix inequality constraints, *SIAM Journal on Matrix Analysis and Applications*, **19(2)**, 499-533.
- [14] Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics*, John Wiley and Sons, Chichester.
- [15] Wu, S.-P., Vandenberghe, L. and Boyd, S. (1996), Software for determinant maximization problems – user’s guild, available at <http://www.stanford.edu/~boyd/maxdet>.
- [16] Zou, H., Hastie, T. and Tibshirani, R. (2004) On the “Degrees of Freedom” of the Lasso, available at stat.stanford.edu/~hastie/pub.htm.